Medieval Manuscripts and Their Migrations: Using SPARQL to Investigate the Research Potential of an Aggregated Knowledge Graph

Toby Burrows, Laura Cleaver, Doug Emery, Eero Hyvönen, Mikko Koho, Lynn Ransom, Emma Thomson, Hanno Wijsman

Burrows, Toby, Laura Cleaver, Doug Emery, Eero Hyvönen, Mikko Koho, Lynn Ransom, Emma Thomson, and Hanno Wijsman. 2022. "Medieval Manuscripts and Their Migrations: Using SPARQL to Investigate the Research Potential of an Aggregated Knowledge Graph." *Digital Medievalist*, 15(1): 3, pp. 1–48. DOI: https://doi.org/10.16995/dm.8064

Open Library of Humanities

Medieval Manuscripts and Their Migrations: Using SPARQL to Investigate the Research Potential of an Aggregated Knowledge Graph

Toby Burrows, University of Oxford, UK, toby.burrows@oerc.ox.ac.uk
Laura Cleaver, University of London, UK, laura.cleaver@sas.ac.uk
Doug Emery, University of Pennsylvania Libraries, US, emery@pobox.upenn.edu
Eero Hyvönen, University of Helsinki, FI, eero.hyvonen@aalto.fi
Mikko Koho, University of Helsinki & Aalto University, FI, mikko.koho@aalto.fi
Lynn Ransom, University of Pennsylvania Libraries, US, lransom@upenn.edu
Emma Thomson, University of Pennsylvania Libraries, US, emmacaw@upenn.edu
Hanno Wijsman, Institut de recherche et d'histoire des textes (CNRS), FR, hannowijsman@gmail.com

Although the RDF query language SPARQL has a reputation for being opaque and difficult for traditional humanists to learn, it holds great potential for opening up vast amounts of Linked Open Data to researchers willing to take on its challenges. This is especially true in the field of premodern manuscripts studies as more and more datasets relating to the study of manuscript culture are made available online. This paper explores the results of a two-year long process of collaborative learning and knowledge transfer between the computer scientists and humanities researchers from the Mapping Manuscript Migrations (MMM) project to learn and apply SPARQL to the MMM dataset. The process developed into a wider investigation of the use of SPARQL to analyse the data, refine research questions, and assess the research potential of the MMM aggregated dataset and its Knowledge Graph. Through an examination of a series of six SPARQL query case studies, this paper will demonstrate how the process of learning and applying SPARQL to query the MMM dataset returned three important and unexpected results: 1) a better understanding of a complex and imperfect dataset in a Linked Open Data environment, 2) a better understanding of how manuscript description and associated data involving the people and institutions involved in the production, reception, and trade of premodern manuscripts needs to be presented to better facilitate computational research, and 3) an awareness of need to further develop data literacy skills among researchers in order to take full advantage of the wealth of unexplored data now available to them in the Semantic Web.

1 Introduction

§1 The primary goals of the Mapping Manuscript Migrations (MMM) project (for the project blog, see: http://blog.mappingmanuscriptmigrations.org/; technical descriptions and publications of the project are available at: https://mappingmanuscriptmigrations.org/en/), funded by the Digging into Data Challenge of the Trans-Atlantic Platform between 2017 and 2020, were to bring together data relating to the history and provenance of medieval and Renaissance manuscripts and to explore the research potential of the aggregated dataset. Based on the Linked Data publishing model (Heath and Bizer 2011) and the W3C Semantic Web standards and technologies (https://www.w3.org/standards/semanticweb), including Universal Resource Identifiers (URI), the RDF data model, ontologies (Staab and Studer 2009), and the SPARQL query language (SPARQL recommendation of W3C: https://www.w3.org/TR/sparql11-query/) for querying RDF data, the project resulted in establishing a Linked Open Data (LOD) service (SPARQL recommendation of W3C: https://www.w3.org/TR/sparql11-query/) and a public MMM portal (the MMM portal is available at: https://mappingmanuscriptmigrations.org). The data and the portal (Hyvönen et al. 2021) allow users to access and query across three distinct datasets, each focusing on premodern manuscript data but built to serve three different purposes: the University of Pennsylvania's *Schoenberg Database of Manuscripts*, the Institut de recherche et d'histoire des textes' *Bibale* database, and the Bodleian Library's online catalogue *Medieval Manuscripts in Oxford Libraries* (respectively, https://sdbm.library.upenn.edu; https://bibale.irht.cnrs.fr; and https://medieval.bodleian.ox.ac.uk). The MMM project also made the transformed datasets (for a full report on MMM data modelling and transformation from legacy databases, see Koho et al. 2021) available for direct searching and downloading on the Zenodo repository (https://zenodo.org/record/4019643).

§2 The work of modelling, combining, and presenting the MMM data was carried out by project team members from the e-Research Centre at Oxford University and the Semantic Computing Research Group at Aalto University, and was based on a series of twenty-four research questions determined at the outset by the project's manuscript researchers at the IRHT and the Schoenberg Institute for Manuscript Studies as well as by members of a focus group gathered in the early stages of the project. The questions were designed to serve as examples of the kinds of inquiries that researchers would want to make in order to identify the key data points they would want to access and query for the data modelling team. They were also used to analyze and test the data model and the viability of the aggregated data and were then used in the evaluation of the public MMM portal (Burrows et al. 2020). To these ends, the original research questions were fundamental to the shaping and successful implementation of the project.

§3 While the launch of the MMM LOD service and portal marked the formal end of the project, for the MMM project team it represented a path to a new frontier for research. The portal, based on the Sampo model (the Sampo model and series of semantic portals are described in: https://seco.cs.aalto.fi/applications/sampo/) and Sampo-UI framework (Ikkala et al. 2021) with its search, data exploration, and data analysis functionalities, is an interface that lies between the users and the underlying RDF data. The portal can be used without programming skills or knowledge about the SPARQL language. The user can choose from five perspectives—Manuscripts, Works, People, Places, and Events—that provide easy entrée into the dataset from different perspectives and facilitate searching and analyzing the data for users new to Linked Data. The perspectives are implemented using SPARQL queries to the underlying LOD service that mediate but also ultimately limit users' ability to query the data flexibly, extensively, and expansively. The perspectives are grounded in traditional research questions that were created outside of a computational context and are therefore not suited to take full advantage of the data model they helped to create. The really interesting data digging happens when the user confronts the RDF data directly via the SPARQL endpoint using custom made SPARQL queries for solving particular research questions. For this purpose, SPARQL editors, such as YASGUI (Rietveld and Hoekstra 2017) can be used, or alternatively programming environments, such as Google Colab (https://colab.research.google.com/notebooks/intro.ipynb) and Jupyter notebooks (https://jupyter.org) for Python scripting for visualizations and data analyses based on SPARQL queries.

§4 This paper explores this process as it was undertaken by members of the project team, the primary authors of the present article who participated in a two-year long process of collaborative learning and knowledge transfer between computer scientists and humanities researchers. The process developed into a wider investigation of the use of SPARQL to analyze the data, explore broader types of research questions, and assess the research potential of the MMM aggregated dataset and its Knowledge Graph. Through an examination of a series of six SPARQL query case studies, we will show that as we became more adept at querying, the better we understood that the scope of original research questions had fallen short of both the abilities and the potential of the MMM data to create new knowledge about the production and transmission of manuscripts across time and that a new approach to research questions would produce better and more transparent results. In addition to analyzing the queries themselves, we will also show what the case studies reveal about the structure and contents of the MMM data, and how lacunae in the data (especially around biographical details of persons) can be compensated for by drawing in information from other Linked Open Data resources like Wikidata.

2 The research questions

§5 Before turning to the SPARQL case studies, it is useful to provide further background to the development of the original research questions to provide context and highlight some of the key problems they presented when applied to the aggregated dataset. A research question is typically understood to be a question that a research project seeks to answer. Identifying a research question or set of questions is generally one of the first steps in developing the methods and techniques for scholarship, whether that scholarship is traditional or digital, because it provides a basis and a goal for starting work. The MMM research questions were based on the team's pre-existing knowledge of each dataset, but they also represented a set of expectations for what manuscript researchers might want to know about manuscripts in general (Table 1).

1.	How many manuscripts produced before 1600 in European countries survive?
2.	How many manuscripts were produced in Northern Italy and/or Lombardy?
3.	How many manuscripts were produced in the Low Countries?
4.	How many manuscripts were produced in London in the fifteenth century?
5.	How many manuscripts formerly owned by Sir Thomas Phillipps are in British Libraries?
6.	What is the average number of folios in a book of hours?
7.	How many surviving manuscripts that contain Spanish texts written in gothic rotunda were produced in Castile for an abbey or convent? How many were owned during the nineteenth century by English private collectors? Which of these are now owned by an institution in North America?
8.	What French collectors purchased manuscripts since the end of the Wars of Religion (after 1598)? Where are their manuscripts now?
9.	How many manuscripts containing texts by Ramon Llul were sold in the 19th century?
10.	Who collects manuscripts with texts by Ramon Llul?
11.	How many times do texts by Ramon Llul appear with texts by Albertus Magnus in the same manuscript?
12.	What was the most popular text by a medieval author in France in the seventeenth-century?
13.	Did Sir Thomas Phillipps own a thirteenth-century bible with historiated initials?
14.	How many illuminated manuscripts were in a specific collection?
15.	Who are the donors and owners of a collection?
16.	Research by subject, technique, language, artist, even the use of pigments in a collection?
17.	Details of a collection (subject, technique, place of production, etc.)? What are its gaps? What are its dominant features?
18.	Life of a collection, or of an illuminated book?

(Contd.)

19.	Which manuscripts have probably been lost?
20.	Which manuscript has been sold and can no longer be identified as part of a collection today?
21.	Which copies of a text are illuminated?
22.	What position does a copy of a text occupy in its transmission? Are there unique exemplars of works?
23.	What are the surviving versions of a work? Who made a French translation of an old text? When?
24.	What are the different surviving publications [copies] of a text (date, place of production, person(s) responsible, etc.)?

Table 1: Mapping Manuscript Migrations Original Research Questions.
This list is also referenced in Burrows et al. (2020). Questions 14 to 24 were borrowed from the *Biblissima* project's list of research questions available here: https://doc.biblissima.fr/ontologie-biblissima#m%C3%A9thodologie.

§6 The questions were designed to include different levels of complexity to test how well results could be retrieved. Simple questions such as 1–6 are based on elements easily identified across all data sets. For example, Questions 1 and 2 require results to be filtered by only one element: by date (before 1600) and by place (Northern Italy and Lombardy) respectively. The remaining questions introduce more complexity. For many of these, simply adding more elements elevated the level of complexity. For example, Question 7 "How many surviving manuscripts that contain Spanish texts written in gothic rotunda were produced in Castile for an abbey or convent?" requires five data elements: language, script type, place of production, former owner, and institution type.

§7 The questions provided a template of data elements for the data model development and helped to define the semantic relationships among the elements that would need to be encoded within the model. But were they good research questions in the sense defined above? Testing them against the RDF in the SPARQL endpoint revealed structural weaknesses in the questions. As the case studies will show, these included semantic ambiguity and misleading assumptions about certain data elements or what the combined datasets were capable of answering.

§8 A successful answer to a research question depends on how well the methods and techniques determined to answer that question are developed and applied to the research process. A successful answer will also depend on how well the research data is understood by those posing the question and how well the question can be mapped to the underlying data model. Querying the dataset using SPARQL exposed the difficulties arising from questions that had too much ambiguity to make computational querying

possible or that were based on flawed assumptions made by users about the abilities of the data to return the expected results. Gaining an awareness of these problems also helped the team refine the questions as their understanding of the available evidence and nature of the data increased.

3 SPARQL query language

§9 SPARQL is the query language designed for data that conform to the RDF model, and hence is a key component of Semantic Web and LOD services and platforms (DuCharme 2013). SPARQL queries follow the pattern of RDF triples, in that they are expressed in the "subject–predicate–object" pattern. Queries are usually run against a SPARQL endpoint exposed by a triple store. Multiple namespaces can be queried in the same query; so can multiple SPARQL endpoints. Some Linked Open Data triple stores containing humanities data offer a public SPARQL endpoint, such as the Getty Vocabularies endpoint and the Wikidata endpoint (http://vocab.getty.edu/sparql; https://query.wikidata.org/sparql).

§10 SPARQL has something of a reputation for being difficult to learn, however, and appears to have been little used by humanities researchers—or at least rarely promoted to them as an active tool for digital humanities projects (Schweizer and Geer 2021). There are few previous specific evaluations of SPARQL in a digital humanities setting. (One exception is: Ichinose et al. 2014. SPARQL is only mentioned briefly in: Meroño-Peñuela et al. 2015.) The best available resource for humanities researchers interested in learning SPARQL is the 2015 tutorial by Matthew Lincoln on the *Programming Historian* Website (Lincoln 2015). This site, however, has been officially "retired"; the examples depended on the British Museum's SPARQL endpoint to its Collections database which is no longer reliably available. Lincoln (2014), an earlier but much shorter introduction to SPARQL by Lincoln, uses Europeana as its basis.

§11 As noted above, the MMM team became interested in exploring different approaches to the aggregated data that went beyond the functionality of the public portal. Guided by the expertise of Semantic Web specialists from Aalto University, the project team conducted a weekly online SPARQL training workshop over the course of two years (May 2019–May 2021). During these sessions, the specialists were able to transfer knowledge to the humanists and in return the humanists provided insight into the research process for the Semantic Web specialists. The MMM project has also published its own introductory tutorial for using SPARQL queries with the MMM data (https://mapping-manuscript-migrations.github.io/sparql/sparql_tutorial.html).

4 The MMM data model and knowledge graph

§12 The MMM data model, which draws on the CIDOC-CRM (Doerr 2003; for the CRM standard online, see: http://www.cidoc-crm.org/) and FRBRoo (Riva, Doerr, and Žumer 2009) ontologies for its entity classes and properties but also adds some specific to MMM, has been discussed in detail elsewhere (Koho et al. 2021, 4–10). It was constructed mainly by inspecting and comparing the different data models used by the three data sources, with additional verification from the twenty-four MMM research queries. It is used to structure the MMM Knowledge Graph, which contains the following entities (as of January 2021):

- 222,605 manuscripts
- 435,428 works and expressions
- 56,685 actors (persons and organizations)
- 5,077 places
- 937,158 events

A significant number of resources in the MMM Knowledge Graph (primarily actors and places) are linked to external authorities. These links originate from the source datasets and from the work done in the MMM project to add shared identifiers to resources in the source datasets for reconciliation purposes. External linkages, in addition to resource level links to the original source datasets, include:

- 15,868 links to VIAF
- 4,617 links to Wikidata
- 4,311 links to data.bnf.fr
- 4,236 links to the Getty TGN
- 3,470 links to ISNI database
- 3,066 links to German national library catalogue
- 3,060 links to IdRef (Identifiers and Referentials)
- 2,572 links to The Library of Congress Linked Data Service
- 1,909 links to Bibliothèque nationale de France catalogue

The vocabularies for actors and places were automatically harmonized across the source data using these identifiers. Manuscripts were harmonized using shelf-marks or Phillipps numbers (assigned by the 19th-century collector Thomas Phillipps). The

names of works were harmonized by manual review of string matching on titles; this only covered titles in the same language, not translated titles in other languages.

§13 A temporal distribution of the events in the MMM data by decades is shown in Figure 1, with separate categories for (1) manuscript production events (`ecrm:E12_Production`), (2) manuscript observations (`ecrm:E10_Transfer_of_Custody` and `mmms:ManuscriptActivity`), and (3) all other events. Only events with an associated timespan are visualized, which accounts for 22.5% of all events. Some events span multiple decades, in which cases an event is counted for each decade. The data are skewed by manuscript survival, cataloguing practices, and most of all by what is catalogued and included in the databases. The SPARQL query used is as follows: https://api.triplydb.com/s/OYKNfOimm.

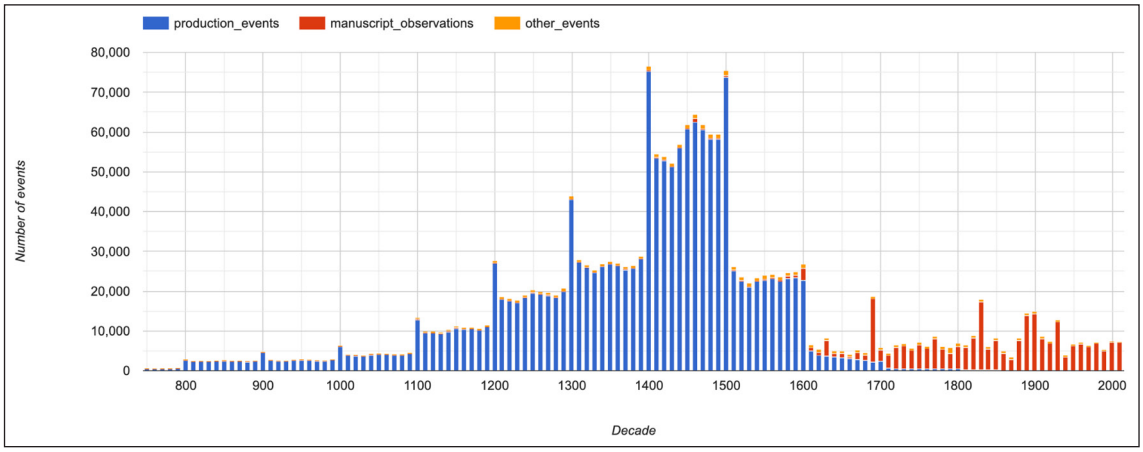


Figure 1: Distribution of events in MMM data, by decades.

§14 One of the important lessons learned from the SPARQL workshops was the necessity of understanding the underlying RDF data model and the semantic links between the data elements in order to perform functional queries. In his explanation of RDF, Joshua Tauberer notes: "What is meant by 'semantic' in Semantic Web is not that computers are going to understand the meaning of anything, but that the logical pieces of meaning can be mechanically manipulated by a machine to useful *human* ends" (Tauberer 2006). The humans using the machine, we learned, must therefore understand the logical structure in order to manipulate it for useful *computational* ends.

§15 When considering the MMM data model, it is important to keep in mind its relationship to the research questions. The data model is expressed in RDF, a method for describing data by defining relationships between data objects. The "subject–predicate–object" pattern produces triples that express the relationships. A triple is

the basic unit of an RDF knowledge graph. For many, the concept of triples is difficult to digest. Unlike most other data models that present data as lists of elements, such as a spreadsheet with well-defined columns or the tables in a relational database, the elements in RDF exist in something more comparable to a cloud of data, seemingly loosely connected by semantic statements. It is much harder to visualize and internalize the structure in one's mind, which may explain why understanding RDF and ways to query it are difficult for non-semantic web specialists.

§16 As the syntactical naming of the units comprising a triple suggests, triples work much like sentences. In a sentence, which can also be a question, subjects and objects are related by the action or state of being that links them. If one considers triples as a list of answers to questions (who did what, what is something, when was something done), then a query in RDF is simply a triple or series of triples statements expressed in the context of a search to identify desired data elements possessing certain relationships. A simple SPARQL query can be expressed as "Show me all things associated with this thing." Then, a further relationship can be added to refine results: "Then show me all the things associated with those things that share this value." Further triple statements can be added to the query indefinitely to execute a variety of search functions. The query, then, is only limited by three things: the researcher's ability to think of new questions to ask or new associations to make; how well the associations have been expressed in the data model in relation to the data; and how well the data has been structured so that the required data elements are accessible to the computer performing the search.

§17 As we noted above, the MMM RDF data model was derived in large part from the data elements identified in the research questions described in the previous section (manuscripts, texts, owners, places of production, dates, etc.) (**Figure 2**). These elements are the nodes represented in the model. The nodes are connected to each other by the properties derived from the MMM ontologies, which express all the possible relationships between the nodes, for example, "is composed of," "has former or current owner," "took place at," "has timespan," etc. In the RDF schema, the nodes are the subjects and objects connected to each other by the properties or predicates; the connections form the triples that can then be queried in a medium like SPARQL. To construct a query, one starts with a node, then follows the associations in any direction where there is a link. In such a flexible structure, the possibilities for what one can query and how are greatly expanded. For the MMM project team, achieving a high degree of familiarity with the data model enhanced the ability to query it and opened up new ways to approach the data well beyond the scope that the original research questions set out to achieve.

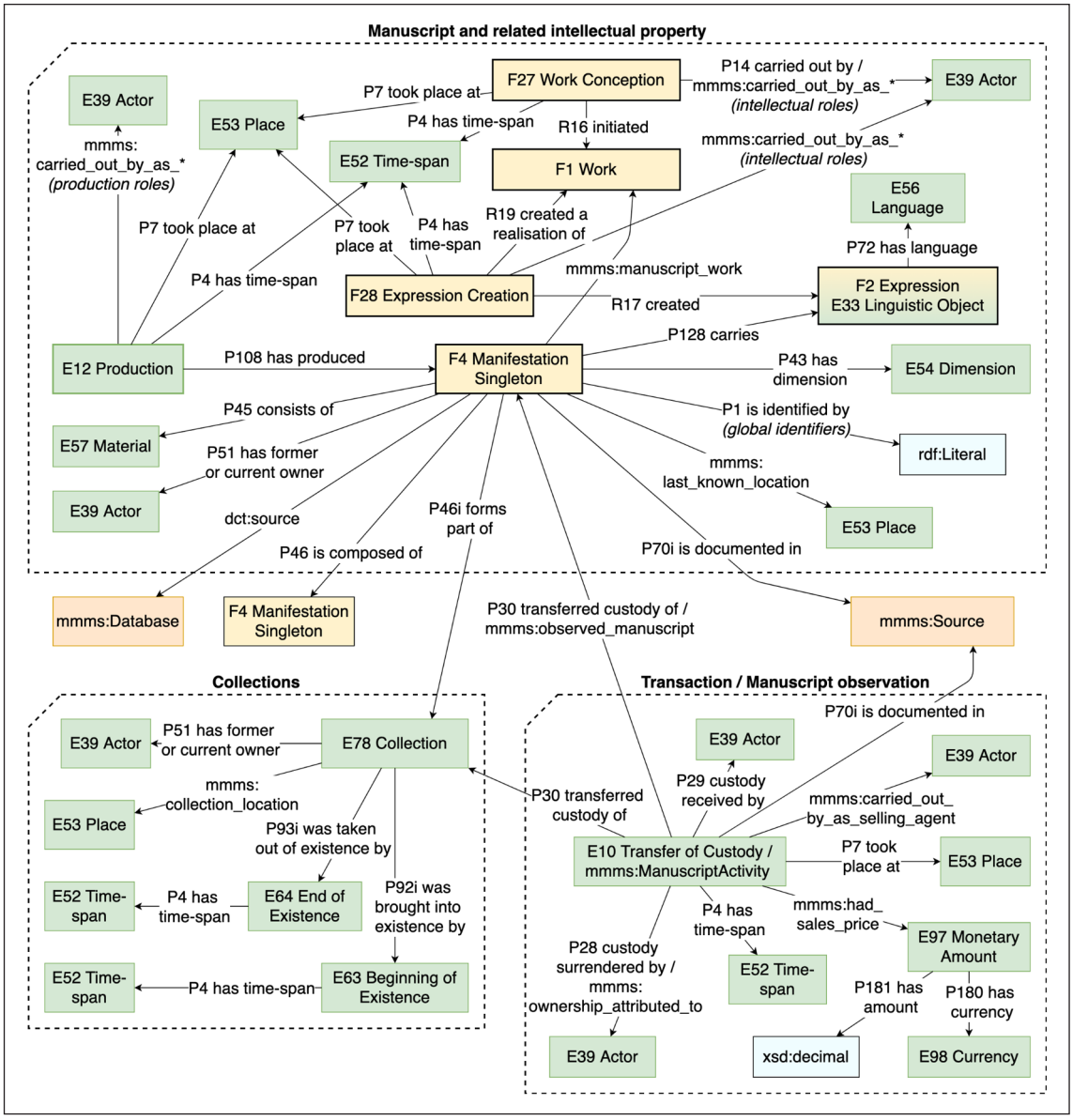


Figure 2: The MMM Data Model.

5 SPARQL queries as case studies

§18 Most of the original MMM research questions could, with some exceptions, be answered to greater or lesser extent through the Semantic Web portal interface to the MMM data, using a combination of filtering and searching. But the project team wanted to go beyond this interface—partly to tackle those questions that the interface could not answer, partly to explore new questions, and partly to explore the relationships and structures in the data more fully. SPARQL queries were used for this purpose, and

the remainder of this paper discusses a selection of these queries as case studies for investigating the research potential of the aggregated dataset and the MMM Knowledge Graph. They include three of the original MMM research questions, as well as further questions arising from another large European manuscript provenance project Cultivate MSS (https://www.ies.sas.ac.uk/research-projects-archives/cultivate-mss-project), an ERC-funded project led by Laura Cleaver at the Institute for English Studies at the University of London, and other questions intended to add data from sources outside the MMM Knowledge Graph. The queries based on these questions did not always return the expected results, but the lessons learned from them led to better questions and better results.

5.1 Query 1: How many manuscripts were produced in Lombardy or Northern Italy?

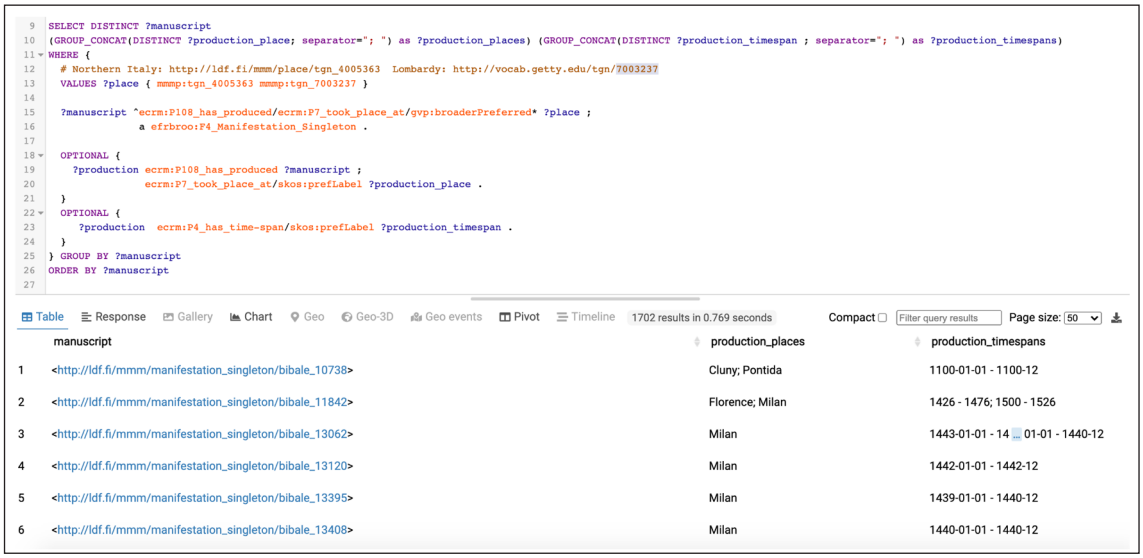
§19 The first query, based on Question 2 of the original research questions, is a simple one requiring only two data elements: show me all of the *manuscripts* associated with a certain *production place*. In this case, a researcher may want to find all illuminated manuscripts produced in or around Milan. Milan, a specific city located within the region of Lombardy, was a major and influential centre of illumination in northern Italy especially in the late Middle Ages, but only a fraction of manuscripts produced in this area during this time have been securely localized to the city in available sources. Shared stylistic features in script or decoration or textual references (e.g., calendars) of otherwise unlocalizable manuscripts can, however, point to affinities with this particularly "northern" style, leading cataloguers to tentatively assign "Northern Italy" as the place of production if the case for a secure tie to Milan or Lombardy is too tenuous to justify.

§20 The researcher will therefore want to cast a wide net to find all manuscripts with a possible connection to Milan. The query "show me all manuscripts produced in Lombardy and Northern Italy" will return a reasonable set of results to allow narrowing down the search for manuscripts produced in Milan. A search for manuscripts produced in Lombardy will helpfully limit results, but a wider search for manuscripts produced in Northern Italy could return more expansive results and more chances for finding manuscripts that have not yet been more accurately localized.

5.1.1 Query explanation https://api.triplydb.com/s/l6M4n5Eff

§21 The query (**Figure 3**) begins with a SELECT statement, which identifies the variable values to be returned by the query. The SELECT statement here (lines 9 to 10) includes variables that will return only *distinct*, or different, manuscript values and production place values. Also included are production timespans, though the timespan is not

essential to the original query. Multiple production place and production timespan values associated with the same manuscript value are concatenated to avoid showing the duplicated values within the same manuscript record.



```
SELECT DISTINCT ?manuscript
(GROUP_CONCAT(DISTINCT ?production_place; separator="; ") as ?production_places) (GROUP_CONCAT(DISTINCT ?production_timespan ; separator="; ") as ?production_timespans)
WHERE {
  # Northern Italy: http://ldf.fi/mmm/place/tgn_4005363  Lombardy: http://vocab.getty.edu/tgn/7003237
  VALUES ?place { mmmp:tgn_4005363 mmmp:tgn_7003237 }

  ?manuscript ^ecrm:P108_has_produced/ecrm:P7_took_place_at/gvp:broaderPreferred* ?place ;
              a efrbroo:F4_Manifestation_Singleton .

  OPTIONAL {
    ?production ecrm:P108_has_produced ?manuscript ;
                ecrm:P7_took_place_at/skos:prefLabel ?production_place .
  }
  OPTIONAL {
     ?production  ecrm:P4_has_time-span/skos:prefLabel ?production_timespan .
  }
} GROUP BY ?manuscript
ORDER BY ?manuscript
```

Table | Response | Gallery | Chart | Geo | Geo-3D | Geo events | Pivot | Timeline | 1702 results in 0.769 seconds | Compact | Filter query results | Page size: 50

	manuscript	production_places	production_timespans
1	<http://ldf.fi/mmm/manifestation_singleton/bibale_10738>	Cluny; Pontida	1100-01-01 - 1100-12
2	<http://ldf.fi/mmm/manifestation_singleton/bibale_11842>	Florence; Milan	1426 - 1476; 1500 - 1526
3	<http://ldf.fi/mmm/manifestation_singleton/bibale_13062>	Milan	1443-01-01 - 14 ... 01-01 - 1440-12
4	<http://ldf.fi/mmm/manifestation_singleton/bibale_13120>	Milan	1442-01-01 - 1442-12
5	<http://ldf.fi/mmm/manifestation_singleton/bibale_13395>	Milan	1439-01-01 - 1440-12
6	<http://ldf.fi/mmm/manifestation_singleton/bibale_13408>	Milan	1440-01-01 - 1440-12

Figure 3: SPARQL query for Query 1.

§22 The places are limited to those associated with the Getty's Thesaurus of Geographical Names (TGN) identifiers for Northern Italy (tgn_4005363) and Lombardy (tgn_7003237) (line 13). The predicates in line 15 (`ecrm:P108_has_produced/ecrm:P7_took_place_at/gvp:broaderPreferred*`) allow the capture of manuscripts associated with these places as well as all other places expressed within the hierarchy of the TGN terms. The asterisk symbol at the end of the predicate `gvp:broaderPreferred*` tells the query to include all results that equal the Getty Thesaurus of Geographic Names URIs for Northern Italy and Lombardy as well as any places that are nested within them. Lines 18 to 21 make optional the association between a value in production place and a manuscript, and lines 22 to 24 do the same for the production timespan.

5.1.2 Results

§23 The query returns 1,702 instances of manuscripts, or manifestation singletons, in the combined dataset that contain the TGN IDs for Northern Italy (tgn_4005363) and for Lombardy (tgn_7003272) as a production place value. The predicate `gvp:broaderpreferred*` in line 15 of the query also enables the capture of cities and sites within Lombardy without having to identify and enter all TGN IDs associated with Lombardy. For example, Results 3 to 6 show "manifestation singletons," which is how

the FRBRoo ontology defines a manuscript object, with Milan as the production place because Milan is contained within Lombardy in the TGN hierarchy (http://vocab.getty.edu/tgn/7003150).

§24 The results also show manifestation singletons with multiple production places. These results indicate more than one place attribution has been assigned to a particular manifestation singleton. There are two reasons for this result. The first has to do with the way that manuscripts are often described: a source description identifies two or more possible places of production either because a cataloguer is hedging bets, for example, a manuscript could be described as from "Austria or Northern Italy" (http://ldf.fi/mmm/manifestation_singleton/sdbm_24767), or because a manuscript contains two component parts that were produced in different places and later bound together. The second reason is due to the data modelling: two or more of the sources could give two different places, as in this example in which the Bodleian record gives one place of production and the SDBM gives another: http://ldf.fi/mmm/manifestation_singleton/bodley_manuscript_2010. In the case of the SDBM, a manuscript record may contain two or more entries that give different location data.

5.1.3 Lessons learned

§25 Some general conclusions can be drawn about the interpretation of the dataset based on these results. The results highlight inconsistencies inherent to manuscript description dependent upon human observation: differing opinions (Austria or Northern Italy?) or knowledge changes across time (it was considered to be made in Northern Italy, but recent studies now indicate that it may have been produced in Siena), and inconsistencies in data entry (production place was not provided in the source data). (The SDBM draws its data from catalogue sources that can vary widely in the amount of detail provided in manuscript description, from simple identification of author, title, and date to full codicological descriptions; it is common therefore for many details relating to the physical description of a manuscript not to be provided.) The query results therefore cannot be taken at face value and researchers must navigate through the manuscript links in the MMM record for further exploration and discovery.

§26 A review of the results for this query raises the question: are the SPARQL results better than the results from a similar query in the MMM portal or separate queries in the original source datasets? The MMM portal and all three source datasets represent places hierarchically based on LOD authorities, including TGN. Querying the original data sources would obviously lack the efficiency of the aggregated dataset, but a search in the MMM portal returns the same results as the SPARQL query (https://mappingmanuscriptmigrations.org/en/manuscripts/faceted-search/table?page=0),

and the visualization tool allows drilling down in the search results much more effectively. Thus, this particular SPARQL query offers only limited advantage over more direct searching in the source datasets and no advantage over the portal.

§27 While the query did not improve on the results provided by the MMM portal, the process of building the query gave shape to the data and insight into the limitations and character of the source data. This exercise, along with other early, relatively simple queries the group created, introduced the building blocks for SPARQL queries, like place and timespan techniques, that were returned to time and again.

5.2 Query 2: How many manuscripts survive that contain Spanish texts written in gothic rotunda script that were produced in Castile for an abbey or convent? How many of these were owned during the nineteenth century by English private collectors and are now owned by an institution in North America?

§28 The first case study shows a simple query that produces no further information beyond what can be gained from a filtered browse in the MMM portal. The second case study demonstrates how adding complexity to the question expands the potential of using SPARQL to query an RDF dataset. The question attempts to determine how many *manuscripts* exist today that were written in a certain *script type* and that were produced in a certain *institution type* existing in a specific *production place*. The question then proposes that those results be further limited to those manuscripts owned by a certain *collector type* from a specific *location* during a specific *timespan*. A final additional query limits the search again to those manuscripts with a specific *current location*.

§29 As one of the original research questions, this question was designed to be complex for complexity's sake, in order to demonstrate for the data modellers how a researcher might want to drill down with increasing specificity using a wide range of qualifiers. It is an intentionally challenging question that tests the limits of the source datasets. The question contains an element "script type" that was ultimately not included in the final data model because it was not adequately represented in the original data sources. The question also requires that the query be able to distinguish between types of institutions (religious, monastic) and types of collectors (private *versus* public) as well as distinguishing *current* locations among all locations identified in the data. Unlike the first case study, this question produced, not surprisingly, a fundamentally more complex query that tests not only the data model but also the user's ability to interpret the results of the query. The query was developed in two steps: first, to identify Castilian manuscripts with Spanish texts; then, to determine who produced them.

5.2.1 Query explanation

5.2.1.1 Step 1 (Figure 4): https://api.triplydb.com/s/GfPEtMgxX

```
SELECT ?manuscript ?manuscript_label ?expression ?expression_label ?production ?place_label WHERE {
  VALUES ?place {<http://ldf.fi/mmm/place/tgn_7002784> <http://ldf.fi/mmm/place/tgn_7002848> <http://ldf.fi/mmm/place/tgn_7007510> <http://ldf.fi/mmm/place/tgn_7010420> <http://ldf.fi/mmm/place/tgn_1061958> <http://ldf.fi/mmm/place/tgn_7007090> <http://ldf.fi/mmm/place/tgn_7002822>}
  ?manuscript a efrbroo:F4_Manifestation_Singleton ;
  ecrm:P128_carries ?expression ;
  skos:prefLabel ?manuscript_label .

  ?expression skos:prefLabel ?expression_label ;
 ecrm:P72_has_language <http://ldf.fi/mmm/language/sdbm_8> .

  ?production ecrm:P108_has_produced ?manuscript ;
              ecrm:P7_took_place_at ?place .

  ?place skos:prefLabel ?place_label .
}
```

Table | Response | Gallery | Chart | Geo | Geo-3D | Geo events | Pivot | Timeline — 286 results in 1.438 seconds

Compact | Filter query results | Page size: 50

	manuscript	manuscript_label	expression	expression_label	production	place_label
1	<http://ldf.fi/mmm/manifestation_singleton/sdbm_1669>	SDBM_MS_1669	<http://ldf.fi/mmm/expression/12028>	Fuero Real	<http://ldf.fi/mmm/event/sdbm_production_1669>	Castile
2	<http://ldf.fi/mmm/manifestation_singleton/sdbm_777>	SDBM_MS_777	<http://ldf.fi/mmm/expression/17046>	Epistolae	<http://ldf.fi/mmm/event/sdbm_production_777>	Castile

Figure 4: SPARQL query for Query 2: Step 1.

§30 This stage of the query finds manuscripts that were produced in Castile and contain texts written in the Spanish language. Following the `SELECT` statement, Line 10 uses a `VALUES` clause to assign the specific URIs for the `?place` variable, including the general region of Castile and more specific locations, both historical and modern, within the region (Valladolid, Toledo, Burgos, Madrid, Ciudad Real, Ávila, and Guadalajara). Lines 11 to 13 include statements defining the `?manuscript` variable as a "manifestation singleton" (`efrbroo:F4_Manifestation_Singleton`). This variable links to the `?expression variable` and returns the human-readabe label for the manuscript (`?manuscript_label`). The `?expression` variable is the text within a manuscript, following the FRBRoo conceptual model that defines the relationships between works, expressions, manifestations, and items in bibliographic records. Line 15 collects the label of the expression. Line 16 states that all expressions included in the results must be written in the Spanish language, encoded as the URI <http://ldf.fi/mmm/language/sdbm_8>.

§31 In MMM, manuscripts are linked to information about their place of production via a production event class. Lines 18–19 return information about these production events by stating that production events (represented by the `?production` variable) are linked to manuscripts via the `ecrm:P108_has_produced` property, and occurred at a

specific place (represented by the `?place` variable). This variable is the same variable defined in line 10; all manuscripts included in the results will thus have a production place that matches one of those values. Line 21 returns the human-readable label for the `?place` variable by using the `skos:prefLabel` predicate to return the label for the `?place` variable.

5.2.1.2 Step 2 (Figure 5): https://api.triplydb.com/s/C83qkzA_h

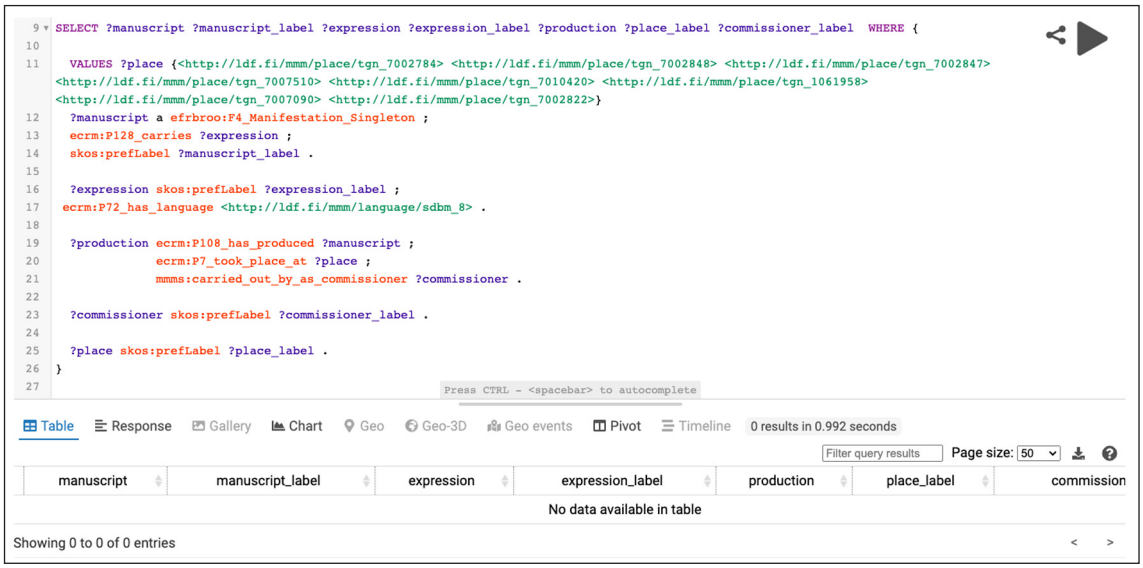


Figure 5: SPARQL query for Query 2: Step 2.

§32 In the next step, we modified the previous query to include "produced for" information. Commissioning data is part of the production event class, so we add the `?commissioner` variable at line 21 by linking it to the `?production` event variable via the predicate `mmms:carried_out_by_as_commissioner`. Line 23 collects the labels for the `?commissioner`. This modified query produces zero results, however, which tells us that no commissioning data is available for this group of manuscripts in the MMM dataset. We were thus not able to continue the query to find out more about later ownership.

5.2.2 Results

§33 The initial query produced a list of 286 texts, or "expressions," in manuscripts which were produced in Castile and written in the Spanish language. Amending the query to look for the person or organization who commissioned the production of these manuscripts produced no results. Further exploration of the data showed that the

MMM-specific property "`carried_out_by_as_commissioner`" is only relevant to the Bibale data. The SDBM identifies provenance agents but does not distinguish whether former owners could also be commissioners. The Bodleian records sometimes include ownership information, like the presence of a coat of arms, that suggests a manuscript was commissioned, but this encoding is not expressed as structured data that can be mapped to the MMM data model. A separate query to show which manuscripts in the whole dataset have a named commissioner (https://api.triplydb.com/s/1tQPkY-au) returned 234 records from Bibale. These results have no overlap with the manuscripts produced in Castile.

§34 As a result, this research question cannot be answered directly. An alternative would be to find the earliest owners of Castilian manuscripts as a proxy for potential commissioners. In the course of investigating this problem, the team produced an ancillary query to locate the distinct places of production and the owners of Spanish-language manuscripts produced in Castile: https://api.triplydb.com/s/M5lTr-KYy. The results can be inspected manually to find religious houses as the earliest owners, since the MMM data do not specify types of institutions. This approach avoids the dead-end of the commissioning relationship and can then be refined to look for 19th-century English owners and present-day American owners. This query in fact finds three Spanish-language manuscripts produced in Castile with religious houses as their (presumably first) owners: two from Madrid, (http://ldf.fi/mmm/manifestation_singleton/bibale_40694 and http://ldf.fi/mmm/manifestation_singleton/sdbm_23689), and one from Burgos (http://ldf.fi/mmm/manifestation_singleton/sdbm_5013). One of these was later owned by a 19th-century British collector (Thomas Phillipps), while another is now in a North American library (University of California Berkeley).

5.2.3 Lessons learned

§35 This research question, which combined place of origin, text language, script, type of commissioner, and then added the later location and ownership of manuscripts, was designed to test the limits of the dataset and is plainly artificial. A pattern emerged during the development of this query that we saw frequently. Often, the source datasets do not contain the specific information sought in the question, or do not encode it in a way that can be mapped to the specific MMM property. The Bodleian catalogue includes 78 cases where persons have the role statement "commissioner, dedicatee, or patron," including MS. Lat. class. d. 38, a Latin manuscript containing the arms of King Alfonso V of Aragon (https://medieval.bodleian.ox.ac.uk/catalog/manuscript_6383). In the MMM data, he is encoded as an owner of this manuscript, but is not linked to its production event. This suggests that some re-thinking of the transformation and

mapping of personal role statements from the Bodleian data in particular might be worth considering.

§36 Both "Castile" and "Spanish" are also problematic in this query. Historical regions like the Kingdom of Castile are not reflected in the TGN hierarchy of places, which is based on current administrative and jurisdictional boundaries, so the property `gvp:broaderpreferred` cannot be used. For this query, the `?place` variable had to be bound to a list of specific place URLs from the TGN that was roughly comprehensive. The lack of availability of geographical hierarchy information, and the fact that historical boundaries change over time, mean that there is no simple method for capturing places within historical regions. Records that represent Castilian manuscripts may simply list Spain as the place of production, but there is no way to determine more specific locations within Spain in the query.

§37 The term "Spanish" for language is also ambiguous. Spanish in its modern sense is a post-medieval phenomenon (Penny 2002); the MMM data sources are inconsistent in their encoding of medieval languages from the Iberian peninsula. The fullest and most accurate way of constructing this query would involve inspecting all these varieties of languages and places in the data sources, seeing the extent to which they are reflected in the MMM data, and ascertaining how best to specify them in the SPARQL query. Even a cursory look suggests a significant level of inconsistency in the source data. These considerations would still apply if the question was made much more specific along these lines: Which manuscripts containing texts in a vernacular language were produced in the Kingdom of Castile as it existed in 1217?

5.3 Query 3: What was the most popular text by a medieval author in France in the 17th century?

§38 This third query offers a further example of how an original research question can be difficult to translate into a satisfactory form that is appropriate for the MMM data model. It requires building a search around the data elements: *author*, *work*, and *place* and *date* associated with a specific *event*, in this case the acquisition of a manuscript with a certain text by a French collector in the 17th century, which is defined in the MMM data model as a "provenance event." All of these elements are included in the data model, but the challenge is to identify what data or combination of data determines popularity. What in the context of the MMM dataset does popularity mean? The following query explanation attempts to extract results based on this assumption. Because of the complexity of the query, the team broke the investigation down into a

series of four query steps in which each query builds upon the results of the previous one.

5.3.1 Query explanation

5.3.1.1 Step 1 (Figure 6): Provenance events occurring in France: https://api.triplydb.com/s/ZWE5m487i

§39 The first step of this query aims to identify all provenance events (dates optional) that occurred in France. Following the `SELECT` statement, Line 9 assigns a specific value to the `?event_type_uri` variable, `ecrm:E10_Transfer_of_Custody`, by using the `VALUES` clause. Thus, every event type returned in the results will be a provenance event involving the transfer of a manuscript from one owner to another, as opposed to more generic provenance events where a direct transfer of ownership is not necessarily known or confirmed by the data. Line 11 states that every location returned in the results (represented by the `?place_uri` variable) must be within the boundaries of France using the same predicate `gvp:broaderPreferred*` that was used in the first case study. Line 14 introduces the `?event_uri` variable, stating that every `?event_uri` must have occurred at the places assigned to (`ecrm:P7_took_place_at`) the `?place_uri` variable in Line 11. Lines 16–17 further define the types of information we return about events. In line 16, the symbol `a` is a shorthand for the `rdf:type` predicate to indicate that the `?event_uri` variable is an instance of the `?event_type_uri` class, which we defined in line 9 as a transfer of custody event. Line 18 is an optional clause that includes the date that an event took place, if that information is present in the data.

Figure 6: SPARQL query for Query 3: Step 1.

5.3.1.2 Step 2 (Figure 7): Manuscripts and their provenance events (dates optional) that occurred in France: https://api.triplydb.com/s/L1Pd3P9ZM

```
SELECT ?manuscript ?event_uri ?event_type_label ?date ?place_uri ?place_label WHERE {

  ?place_uri gvp:broaderPreferred* <http://ldf.fi/mmm/place/tgn_1000070> .
  ?place_uri gvp:placeTypePreferred ?place_type .
  ?place_uri owl:sameAs ?place_getty_uri .
  ?place_uri skos:prefLabel ?place_label .

  ?event_uri ecrm:P7_took_place_at ?place_uri .
  ?event_uri ecrm:P30_transferred_custody_of|mmms:observed_manuscript ?manuscript.
  ?manuscript a efrbroo:F4_Manifestation_Singleton.

  ?event_uri a ?event_type_uri .
  ?event_type_uri rdfs:label|skos:prefLabel ?event_type_label .
  OPTIONAL { ?event_uri ecrm:P4_has_time-span/skos:prefLabel ?date }
}
```

Table | Response | Gallery | Chart | Geo | Geo-3D | Geo events | Pivot | Timeline — 47805 results in 5.088 seconds

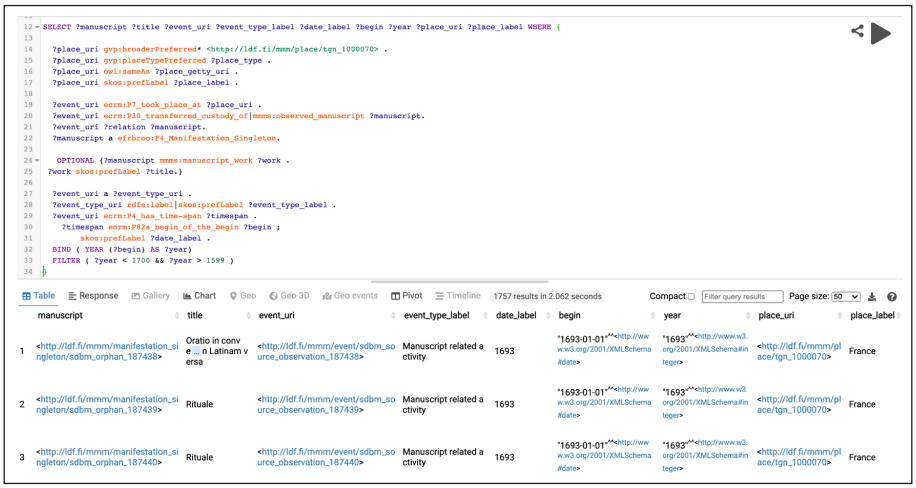
Filter query results — Page size: 50

	manuscript	event_uri	event_type_label	date	place_uri	place_label
1	<http://ldf.fi/mmm/manifestation_singleton/sdbm_orphan_111326>	<http://ldf.fi/mmm/event/sdbm_source_observation_111326>	Manuscript related activity	1735-07-18	<http://ldf.fi/mmm/place/tgn_1000070>	France
2	<http://ldf.fi/mmm/manifestation_singleton/sdbm_orphan_111327>	<http://ldf.fi/mmm/event/sdbm_source_observation_111327>	Manuscript related activity	1735-07-18	<http://ldf.fi/mmm/place/tgn_1000070>	France
3	<http://ldf.fi/mmm/manifestation_singleton/sdbm_orphan_111328>	<http://ldf.fi/mmm/event/sdbm_source_observation_111328>	Manuscript related activity	1735-07-18	<http://ldf.fi/mmm/place/tgn_1000070>	France

Figure 7: SPARQL query for Query 3: Step 2.

§40 Building on the above query, the second query's results include all the manuscripts associated with provenance events that occurred in France, with the dates on which they occurred if known. Line 19 states that the `?event_uri` variable is linked to the `?manuscript` variable via two potential provenance event predicates: either transfer of custody events or observed manuscript events, which are provenance events where a direct transfer of custody is not confirmed in the data.

5.3.1.3 Step 3 (Figure 8). Manuscripts with their titles (optional), that had a provenance event that occurred in France in the 17th century: https://api.triplydb.com/s/WVeDNDp7V



```
SELECT ?manuscript ?title ?event_uri ?event_type_label ?date_label ?begin ?year ?place_uri ?place_label WHERE {

  ?place_uri gvp:broaderPreferred* <http://ldf.fi/mmm/place/tgn_1000070> .
  ?place_uri gvp:placeTypePreferred ?place_type .
  ?place_uri owl:sameAs ?place_getty_uri .
  ?place_uri skos:prefLabel ?place_label .

  ?event_uri ecrm:P7_took_place_at ?place_uri .
  ?event_uri ecrm:P30_transferred_custody_of|mmms:observed_manuscript ?manuscript.
  ?event_uri ?relation ?manuscript.
  ?manuscript a efrbroo:F4_Manifestation_Singleton.

   OPTIONAL {?manuscript mmms:manuscript_work ?work .
?work skos:prefLabel ?title.}

  ?event_uri a ?event_type_uri .
  ?event_type_uri rdfs:label|skos:prefLabel ?event_type_label .
  ?event_uri ecrm:P4_has_time-span ?timespan .
    ?timespan ecrm:P82a_begin_of_the_begin ?begin ;
        skos:prefLabel ?date_label .
  BIND ( YEAR (?begin) AS ?year)
  FILTER ( ?year < 1700 && ?year > 1599 )
}
```

Table | Response | Gallery | Chart | Geo | Geo-3D | Geo events | Pivot | Timeline — 1757 results in 2.062 seconds — Compact — Filter query results — Page size: 50

	manuscript	title	event_uri	event_type_label	date_label	begin	year	place_uri	place_label
1	<http://ldf.fi/mmm/manifestation_singleton/sdbm_orphan_187438>	Oratio in conv e in Latinam v ersa	<http://ldf.fi/mmm/event/sdbm_source_observation_187438>	Manuscript related activity	1693	"1693-01-01"^^<http://www.w3.org/2001/XMLSchema#date>	"1693"^^<http://www.w3.org/2001/XMLSchema#integer>	<http://ldf.fi/mmm/place/tgn_1000070>	France
2	<http://ldf.fi/mmm/manifestation_singleton/sdbm_orphan_187439>	Rituale	<http://ldf.fi/mmm/event/sdbm_source_observation_187439>	Manuscript related activity	1693	"1693-01-01"^^<http://www.w3.org/2001/XMLSchema#date>	"1693"^^<http://www.w3.org/2001/XMLSchema#integer>	<http://ldf.fi/mmm/place/tgn_1000070>	France
3	<http://ldf.fi/mmm/manifestation_singleton/sdbm_orphan_187440>	Rituale	<http://ldf.fi/mmm/event/sdbm_source_observation_187440>	Manuscript related activity	1693	"1693-01-01"^^<http://www.w3.org/2001/XMLSchema#date>	"1693"^^<http://www.w3.org/2001/XMLSchema#integer>	<http://ldf.fi/mmm/place/tgn_1000070>	France

Figure 8: SPARQL query for Query 3: Step 3.

§41 This query expands and refines the results further by adding the titles of works within the manuscripts and limiting the timeframe of provenance events to those that occurred in the 17th century. Lines 24–25 feature an `OPTIONAL` clause to retrieve the works included in the manuscripts (if known) and the labels of those works, represented by the variable `?titles`.

§42 Lines 29–34 include statements related to the dates when a provenance event took place. For this research question, we are interested in events that occurred in the 17th century. The range of timespans included in the results need to have begun after the year 1599 but before the year 1700. To specify these parameters in SPARQL, we take the beginning of the timespan specified in each event (the `?begin` variable), use the `BIND` and `YEAR` functions to extract the year from each timespan and assign each year to a new variable, `?year`, and then `FILTER` the results to include only those years that are less than 1700 but greater than 1599.

5.3.1.4 Step 4 (Figure 9). Manuscript with texts by authors who lived between 450–1500, with provenance events that occurred in France in the 17th century: https://api.triplydb.com/s/_9cC7UFM-

Figure 9: SPARQL query for Query 3: Step 4.

§43 This query adds information about authors and their life dates to the results. The strategy for limiting authors by their life dates is similar to the route taken in the previous query to find 17th century provenance events. Since the research question is interested in works composed by medieval authors, our results need to be limited to authors who lived during the medieval period, which we defined as between 450–1550 CE. Works are linked to their authors via the `mmm:carried_out_by_as_possible_author` predicate, as seen in lines 33–34. Authors with known life dates will have their birth and/or death dates (which could each vary widely in specificity from a precise date to a range of time) stored separately in the database, so we need to filter on birth and death dates separately. The parameters for the authors' births are stated in line

35–39. We link from the author to their birth event (line 35), from that birth event to the timespan for that event, and then to the beginning of the timespan (`?author_birth_begin`). Just as in the previous query, we use the `BIND` and `FILTER` functions on lines 38–39 to extract the year from the timespan and then filter to include only years that are less than 1500 but greater than 450. We use the same process to limit authors' death dates to the medieval period in lines 41–45, by using predicates specific to author death events and limiting the dates to between 500 and 1550.

5.3.2 Results

§44 Query 3: Step 1 returned 1,765 results. Event dates ranged from "after 877" (e.g., <http://ldf.fi/mmm/event/bibale_transfer_association:3972>) to "2020-04-01 – 2020-04-03" (e.g., <http://ldf.fi/mmm/event/sdbm_source_observation_260557). About 20% of these records had no associated date. The second query returned 47,805 results. More than 96% of these were generic "manuscript-related events" rather than transfers of custody. About 50% of them had no associated date. The third query reduced the number of records drastically. Only 1757 records were identified as "transfer of custody" or "observed ownership" events that could be localized to 17th-century France.

§45 When authors' birth and death dates were added to find "medieval" authors in the fourth query, the number of results was further reduced to 1,262 records. This list contains all the possible combinations of authors, works, dates, and manuscripts. The query can be analyzed to reveal that the list contains 264 manuscripts, 757 distinct works, and 153 different authors. Because the titles of works have not been harmonized across versions in different languages—and also because of the way in which the SDBM records multiple works contained in a single manuscript—it is impossible to say with any certainty which work occurs most frequently. (In the SDBM, multiple works and multiple authors occurring in the same manuscript are listed separately and are not linked to each other. This means that the MMM mapping has to describe each author as the "possible author" of each work, even though they may only be the author of one of the works in question.) But the most frequently occurring authors are clear: Isidore of Seville (25 manuscripts), Bede (22), Bernard of Clairvaux (13), Anselm of Canterbury (12), and Boethius (11).

5.3.3 Lessons learned

§46 The complexity of this research question meant that we had to break it down into parts, write queries for each part, and then assemble these into a single SPARQL query. It also revealed that questions can be expressed in a form that is difficult to map to the

terms and relationships used in the data model and the aggregated dataset. To approach a set of results that could be used to answer the question "what was the most popular text?" meant tackling a series of definitional problems and making choices about how best to define them in the context of the MMM data.

§47 The phrase "most popular text" is ambiguous, for a start. It could mean the most-read text, the most-quoted text, the most-owned text, or the most-circulated text. Only the latter two have any relevance to the MMM data, since they can be expressed respectively in terms of "the text in those manuscripts with the most recorded owners" in 17th-century France, or "the text in those manuscripts with the most ownership events" in the same period. Does the question refer to manuscript owners associated with France, or manuscript provenance events which occurred in France? Does "medieval author" cover anonymous or pseudonymous works and expressions as well as those with known authors? If so, how do we identify anonymous "medieval" texts, since works and expressions do not have dates directly associated with them?

§48 Whatever choices were made in relation to these definitional difficulties, the important point was to ensure that those choices were documented and explained. It might also have been possible to consider reframing the question in a less prescriptive way: "Which manuscripts with medieval texts were owned by French collectors in the 17th century?" This could have been addressed by identifying owners living in France in the 17th century and looking at the manuscripts they owned and the associated works.

§49 As mentioned earlier, one factor affecting these results significantly is that titles of works have not been harmonized across translations in different languages. There is little in the way of authoritative Linked Open Data vocabularies and identifiers for medieval and Renaissance works, and the absence of consistent conventional titles for works in this period makes the process of reconciling them between their occurrences in different manuscripts extremely difficult (Sharpe 2003). Without this kind of reconciliation, we cannot easily construct a query that takes a work and looks for all manuscripts containing that work. We should either try to identify all the variant titles of a work and include them in the query, or focus on manuscripts and authors instead. The way in which the SDBM treats multiple works in a single manuscript (as described above) also has a significant effect on queries of this kind.

5.3.4 New research questions and wider explorations

§50 The three case studies considered so far attempted to apply research questions devised before the data model was designed and implemented. The results of each query were mixed. While the simplest query produced the expected results primarily because of its simplicity, it did not really test the ability of the dataset to return results. The more

complex questions with a significantly greater degree of ambiguity were much harder to translate into the elements and relationships expressed in the data model. They revealed, amongst other things, that some queries could be too specific or too complex in their combination of criteria to produce meaningful results. They also revealed that some relationships in the MMM data (e.g., between authors and works) were too ambiguous to produce reliable results. And they showed that questions involving pre-modern languages or pre-modern political and administrative jurisdictions needed careful mapping to modern authoritative vocabularies for places and languages. But they helped to teach some of the intricacies of SPARQL in the context of a relatively complex data model and a dataset that contains important ambiguities.

§51 Moving the SPARQL queries beyond the initial set of research questions became an important goal and the focus of more recent workshop sessions. For this second round of investigations, we looked particularly at ways of visualizing data in response to comparative quantitative and exploratory questions. We also examined ways of extending the reach of questions by using data sources outside MMM to add missing contextual information. The questions were mostly derived from active research projects into the history of manuscript collecting, a topic for which the MMM data should be particularly relevant.

5.4 Query 4: What are the ratios of height to width in medieval liturgical manuscripts?

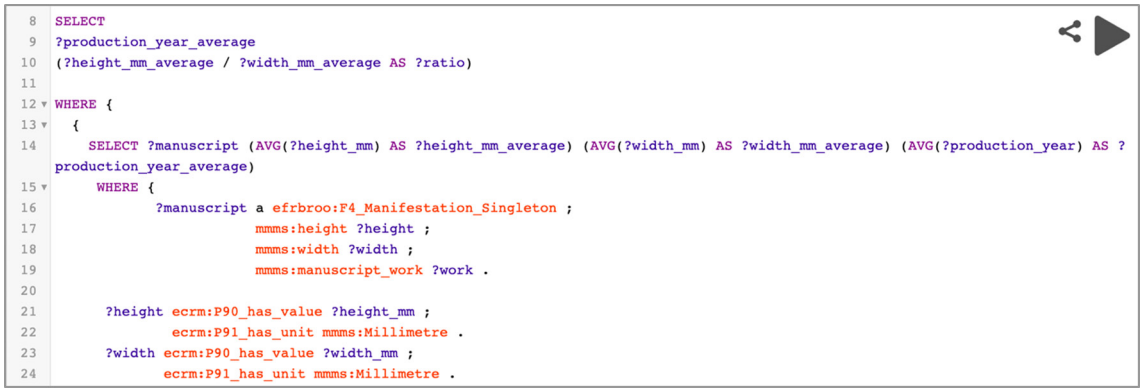
§52 The aggregated MMM data (like the source datasets) contain various quantifiable elements relating to the physical properties of individual manuscripts. These include height, width, folio count, and number of lines on a page, as well as the numbers of miniatures and decorated initials. A Twitter thread in December 2020 devoted to the importance of recording a manuscript's size and folio count (Smith 2020) included a visualization of height-to-width ratios in 3,413 manuscripts, using data from the SDBM (Davis 2020). This prompted an exploration of the same kind of data in MMM using a new set of SPARQL queries in an effort to confirm this visualization and to correct the problem of multiple entries referring to the same manuscript, potentially skewing the results. In the MMM dataset, duplicate manuscript entries from the SDBM data as well as from the Bodleian and Bibale data were reconciled into one record, thus reducing a certain amount of noise in the results.

§53 To construct the query, a formula to calculate ratios is applied to two elements, *height* and *width* restricted to a specific *manuscript type*, in this case, liturgical manuscripts. These manuscripts contain the prayers, readings, and hymns recited or sung during the Mass or as part of the Divine Office. They include missals and graduals for the Mass, and breviaries and antiphonaries for the Divine Office. Other less common

types of liturgical manuscripts include sacramentaries, sequentaries, pontificals, and ordinals. While manuscript dimensions are available in all three source datasets, none of the datasets include the element "*manuscript type*" in their respective data models. The solution was to query for records containing specific titles reflecting liturgical manuscripts.

5.4.1 Query explanation

5.4.1.1 Step 1 (Figure 10): Manuscript production year averages and ratios of height: https://api.triplydb.com/s/nfhgCrlyB



```
8  SELECT
9  ?production_year_average
10 (?height_mm_average / ?width_mm_average AS ?ratio)
11
12 WHERE {
13   {
14     SELECT ?manuscript (AVG(?height_mm) AS ?height_mm_average) (AVG(?width_mm) AS ?width_mm_average) (AVG(?production_year) AS ?
   production_year_average)
15      WHERE {
16            ?manuscript a efrbroo:F4_Manifestation_Singleton ;
17                        mmms:height ?height ;
18                        mmms:width ?width ;
19                        mmms:manuscript_work ?work .
20
21       ?height ecrm:P90_has_value ?height_mm ;
22               ecrm:P91_has_unit mmms:Millimetre .
23       ?width ecrm:P90_has_value ?width_mm ;
24              ecrm:P91_has_unit mmms:Millimetre .
```

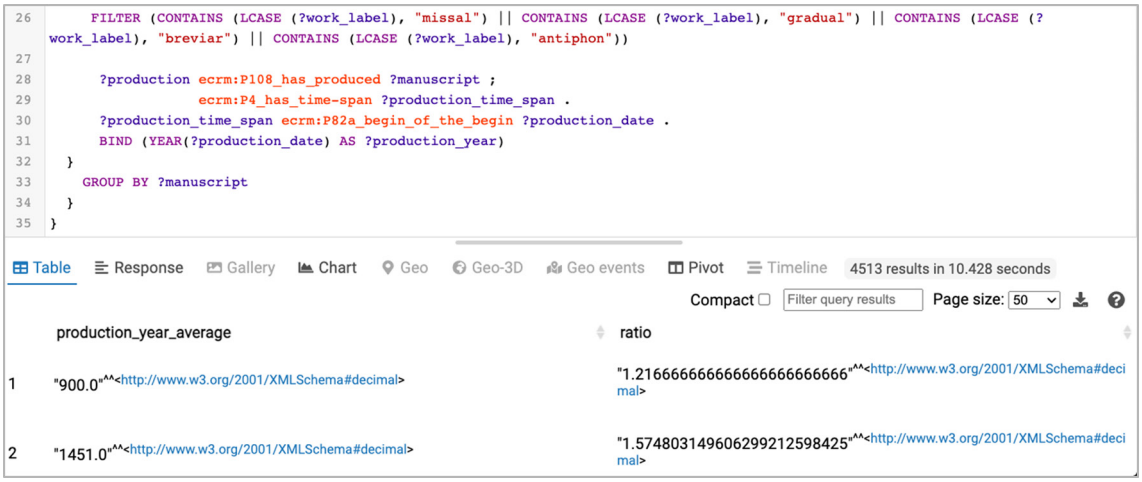
Figure 10: SPARQL query for Query 4: Step 1 (lines 1–24).

§54 This query begins with a very simple `SELECT` statement that includes only two variables: one representing a manuscript's production year average, and another representing the ratio of a manuscript's size. We defined this ratio as a manuscript's average height divided by its average width. In MMM, manuscripts often have multiple different values for their heights, widths, and production years because our data about them comes from many different sources created over time. This necessitates that we use the averages of these values.

§55 Calculating averages requires a subquery nested within the `WHERE` clause of our main query, beginning on lines 12–14. The `SELECT` statement in the subquery begins with the `?manuscript` variable, followed by three instances of the average aggregate function (AVG) that will calculate the averages of the `?height_mm`, `?width_mm`, and `?production_year` variables. The `WHERE` clause beginning on line 15 defines the desired triple (i.e., subject–predicate–object) patterns in these variables. Lines 16–19 pertain to the `?manuscript` variable, defining it as a manifestation singleton and returning height, width, and work data. Lines 21–24 refine the height and width

information by returning the value of those fields and defining that value as a unit of length in millimeters.

§56 To narrow the results of this query to only include liturgical manuscripts, we can filter the manuscripts based on their text titles, which are modeled as work labels in MMM (**Figure 11**). Line 26 includes a `FILTER` clause that refines the results to only include manuscripts that contain works that include the characters "missal," "gradual," "breviar," or "antiphon" in their titles. The character strings will be matched exactly, so the work labels are abbreviated in order to accommodate the various spellings found in MMM's data sources. Likewise, each `CONTAINS` function also includes the `LCASE` function to convert the values in the `?work_label` field to all lowercase letters so as to include both upper and lowercase title variations in our `FILTER` clause.



```
26      FILTER (CONTAINS (LCASE (?work_label), "missal") || CONTAINS (LCASE (?work_label), "gradual") || CONTAINS (LCASE (?work_label), "breviar") || CONTAINS (LCASE (?work_label), "antiphon"))
27
28      ?production ecrm:P108_has_produced ?manuscript ;
29                  ecrm:P4_has_time-span ?production_time_span .
30      ?production_time_span ecrm:P82a_begin_of_the_begin ?production_date .
31      BIND (YEAR(?production_date) AS ?production_year)
32    }
33    GROUP BY ?manuscript
34  }
35}
```

Table | Response | Gallery | Chart | Geo | Geo-3D | Geo events | Pivot | Timeline | 4513 results in 10.428 seconds

Compact | Filter query results | Page size: 50

	production_year_average	ratio
1	"900.0"^^<http://www.w3.org/2001/XMLSchema#decimal>	"1.216666666666666666666666"^^<http://www.w3.org/2001/XMLSchema#decimal>
2	"1451.0"^^<http://www.w3.org/2001/XMLSchema#decimal>	"1.574803149606299212598425"^^<http://www.w3.org/2001/XMLSchema#decimal>

Figure 11: SPARQL query for Query 4: Step 1 (lines 25–35).

§57 Lines 28–31 construct the `?production_year` variable. Production event information is its own class in MMM, which links to specific manuscripts via the `ecrm:P108_has_produced` predicate (line 28) and timespan data via the `ecrm:P4_has_time-span` predicate (line 29). MMM timespans follow the CIDOC-CRM model for modeling the range of a timespan. For this query, we've elected to use the "beginning of the begin" of the timespan data (the terminus post quem) as a manuscript's production date (line 30). Line 31 uses the `BIND` and `YEAR` functions to extract the year information out of the production timespan and assign it as the `?production_year` variable. This concludes the subquery. As a last step, we group our results according to the `?manuscript` variable on line 33 to ensure that our results table displays information for only one manuscript per row.

5.4.1.2 Step 2 (Figure 12): Revised query to filter results for manuscripts produced after 700: https://api.triplydb.com/s/-9C8qoZtb

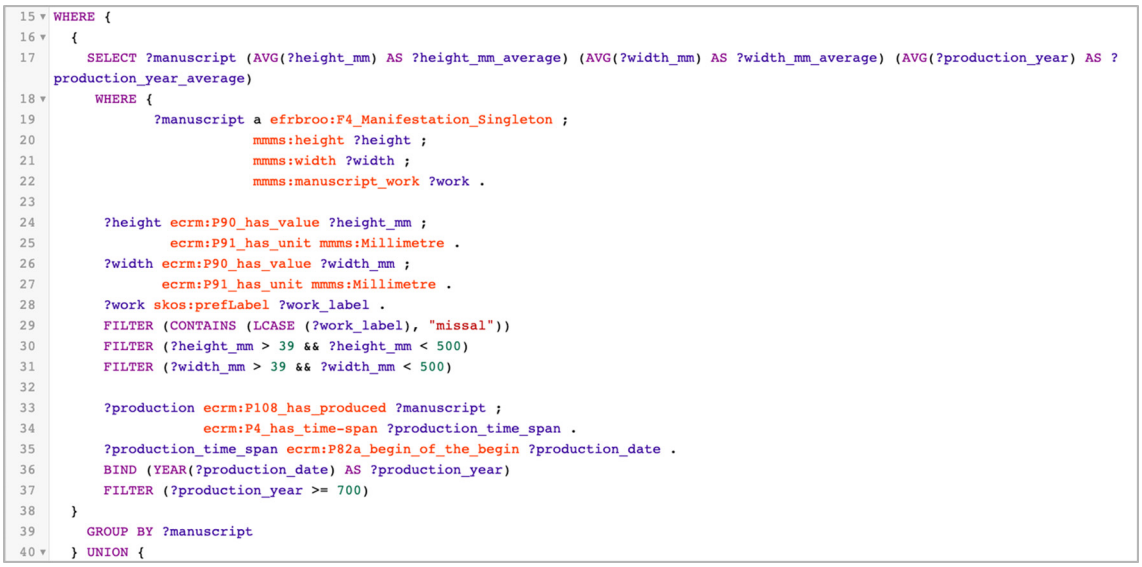
Figure 12: SPARQL query for Query 4: Step 2 (lines 21–37).

§58 This query is nearly identical to the previous query, except that it includes three extra `FILTER` functions to refine the results further. On lines 30–31, two `FILTER` functions state that all height and width measurements included in the calculations must be greater than 39 millimeters and less than 500 millimeters. Filtering the results in this way helps ensure that our results do not include typos or other data entry mistakes that sometimes appear in the measurement data. Line 37 filters the production year results to include only manuscripts produced on or after 700 CE. The choice to filter by this production year stems from a cosmetic need to produce a chart of the results that is easier to read. Since few manuscripts in the MMM dataset were produced before 700 CE, removing those manuscripts from the results creates a more efficient x-axis and greater legibility of the individual data points in the chart.

5.4.2 Alternative Query 4 (Figures 13a-c): Comparing ratios of different liturgical books: breviaries and missals: https://api.triplydb.com/s/qrzY6bd0e

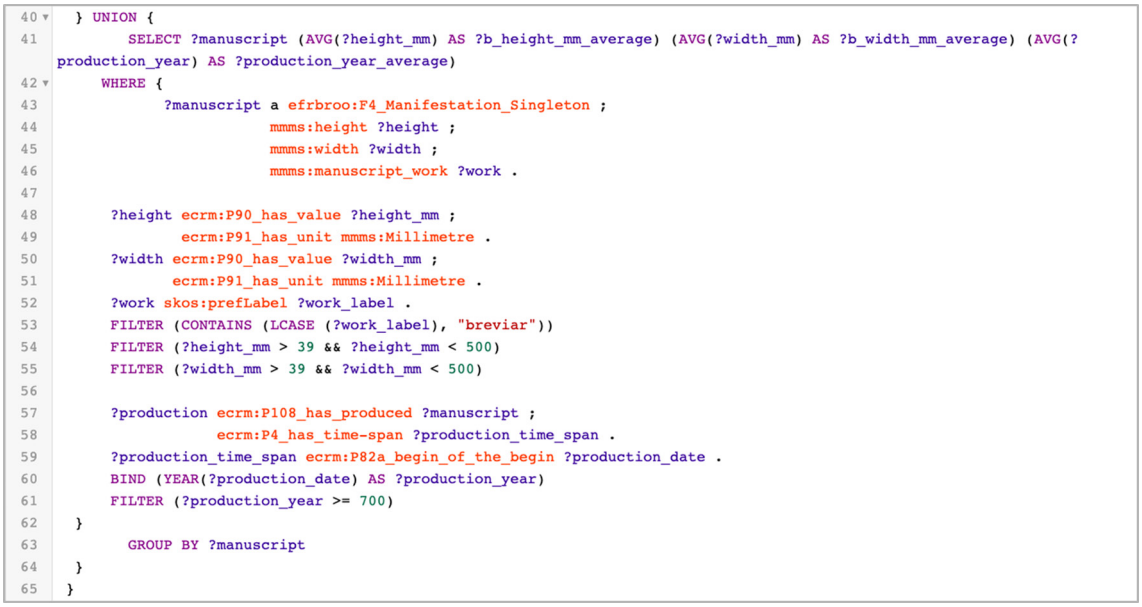
```
10 SELECT
11 ?production_year_average
12 (?height_mm_average / ?width_mm_average AS ?missal_ratio)
13 (?b_height_mm_average / ?b_width_mm_average AS ?breviary_ratio)
14
```

Figure 13a: SPARQL query for Alternative Query 4 (lines 10–14).



```
15 WHERE {
16   {
17     SELECT ?manuscript (AVG(?height_mm) AS ?height_mm_average) (AVG(?width_mm) AS ?width_mm_average) (AVG(?production_year) AS ?
   production_year_average)
18      WHERE {
19            ?manuscript a efrbroo:F4_Manifestation_Singleton ;
20                        mmms:height ?height ;
21                        mmms:width ?width ;
22                        mmms:manuscript_work ?work .
23
24       ?height ecrm:P90_has_value ?height_mm ;
25               ecrm:P91_has_unit mmms:Millimetre .
26       ?width ecrm:P90_has_value ?width_mm ;
27              ecrm:P91_has_unit mmms:Millimetre .
28       ?work skos:prefLabel ?work_label .
29       FILTER (CONTAINS (LCASE (?work_label), "missal"))
30       FILTER (?height_mm > 39 && ?height_mm < 500)
31       FILTER (?width_mm > 39 && ?width_mm < 500)
32
33       ?production ecrm:P108_has_produced ?manuscript ;
34                   ecrm:P4_has_time-span ?production_time_span .
35       ?production_time_span ecrm:P82a_begin_of_the_begin ?production_date .
36       BIND (YEAR(?production_date) AS ?production_year)
37       FILTER (?production_year >= 700)
38   }
39     GROUP BY ?manuscript
40   } UNION {
```

Figure 13b: SPARQL query for Alternative Query 4 (lines 15–40).



```
40   } UNION {
41         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)
42      WHERE {
43            ?manuscript a efrbroo:F4_Manifestation_Singleton ;
44                        mmms:height ?height ;
45                        mmms:width ?width ;
46                        mmms:manuscript_work ?work .
47
48       ?height ecrm:P90_has_value ?height_mm ;
49               ecrm:P91_has_unit mmms:Millimetre .
50       ?width ecrm:P90_has_value ?width_mm ;
51              ecrm:P91_has_unit mmms:Millimetre .
52       ?work skos:prefLabel ?work_label .
53       FILTER (CONTAINS (LCASE (?work_label), "breviar"))
54       FILTER (?height_mm > 39 && ?height_mm < 500)
55       FILTER (?width_mm > 39 && ?width_mm < 500)
56
57       ?production ecrm:P108_has_produced ?manuscript ;
58                   ecrm:P4_has_time-span ?production_time_span .
59       ?production_time_span ecrm:P82a_begin_of_the_begin ?production_date .
60       BIND (YEAR(?production_date) AS ?production_year)
61       FILTER (?production_year >= 700)
62   }
63     GROUP BY ?manuscript
64   }
65 }
```

Figure 13c: SPARQL query for Alternative Query 4 (lines 40–65).

§59 This alternative query copies the basic structure of the previous query to produce results that compare the average ratios of missals to the average ratios of breviaries. The `SELECT` statement includes two different sets of ratios, one for missals (line 12) and one for breviaries (line 13).

§60 To calculate these two different ratios, we use the same subquery strategy as employed previously, but a `UNION` clause (line 40) allows the results to be displayed together. The first subquery (beginning on line 17) calculates the data for missals, using the `FILTER` function to isolate those manuscripts that have the characters "missal" in their work label (line 29).

§61 This exact structure is copied in the second subquery (beginning on line 41), except in this case the `FILTER` function finds works containing the characters "breviar" (line 53). To distinguish the two results, the averages related to breviaries are called `?b_height_mm_average` and `?b_width_mm_average`.

5.4.3 Results

§62 Step 1 of the original query visualizes the height-to-width ratios for 4,513 liturgical manuscripts (**Figure 14**). It includes missals, graduals, breviaries, and antiphonaries, but the ratios are not distinguished by type of manuscript. There are no limits on the date of production, or on the size of the ratios. Because there are four outlying ratios between 8.636 and 30.831, as well as a small number of early production dates, the other results are heavily compressed, and the details of the other ratios cannot easily be seen.

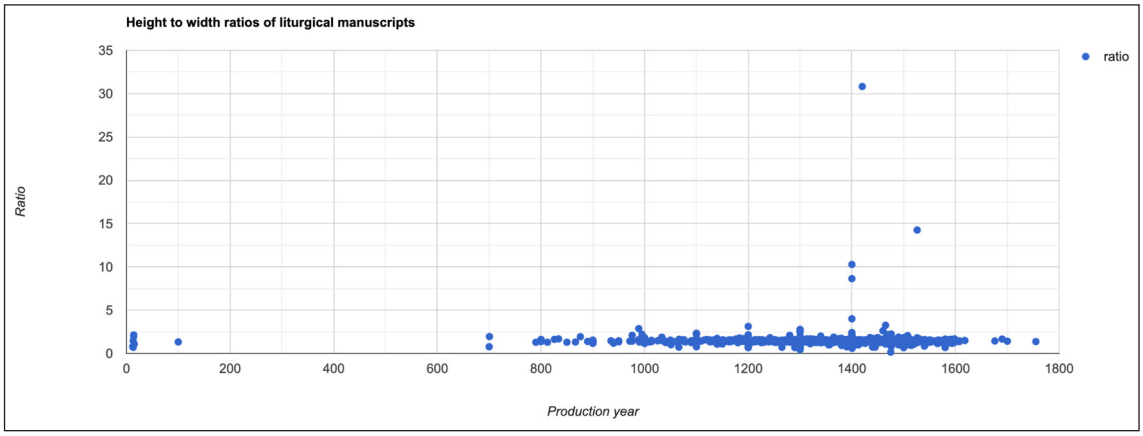


Figure 14: Height-to-width ratios of liturgical manuscripts.

§63 Step 2 of the original query visualizes the height-to-width ratio of 4,030 liturgical manuscripts by limiting the production period to 700 to 1800 CE (**Figure 15**). The variation in ratios is also limited by the exclusion of manuscripts larger than 500mm or smaller than 39mm. It includes missals, graduals, breviaries, and antiphonaries, but the ratios are not distinguished by type of manuscript. The results are clustered around 1.25 to 1.6; most manuscripts were produced in the 14th or 15th centuries. The clusters of

results for the years 900, 1000, 1100, 1200, 1300, and 1400 reflect the use of start dates for estimated production year ranges. Another version of this query makes use of end dates as well, to smooth out this kind of clustering: https://api.triplydb.com/s/uG86O-AIC.

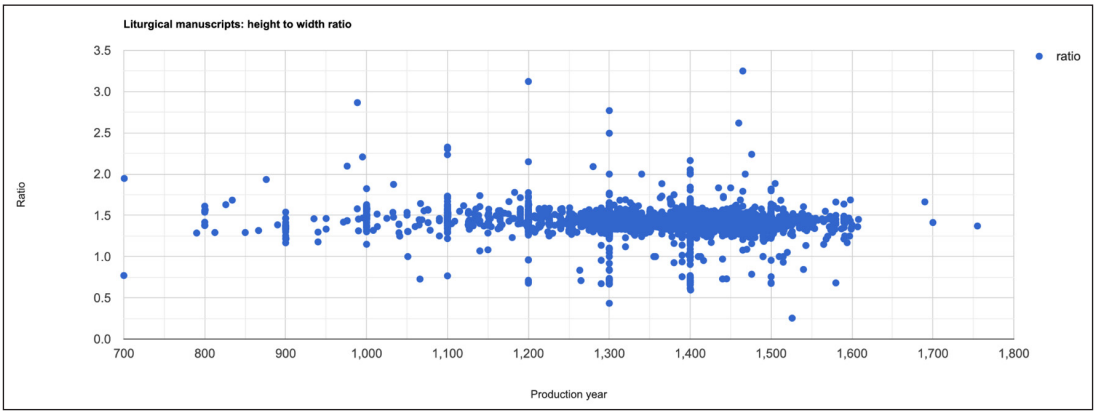


Figure 15: Height-to-width ratios of liturgical manuscripts produced between 700-1800 AD.

§64 The alternative query compares the height-to-width ratio of two different types of liturgical manuscripts produced during the period 700 to 1700 CE (**Figure 16**). The total number of manuscripts involved is 12,169. Missals are shown as blue dots and breviaries appear in red. Most of the manuscripts fall within the range 1.0 to 2.0, though the majority fall between 1.25 and 1.6. There is considerable similarity between the two different types. Relatively few manuscripts have ratios less than 1.0 (i.e., with their width greater than their height).

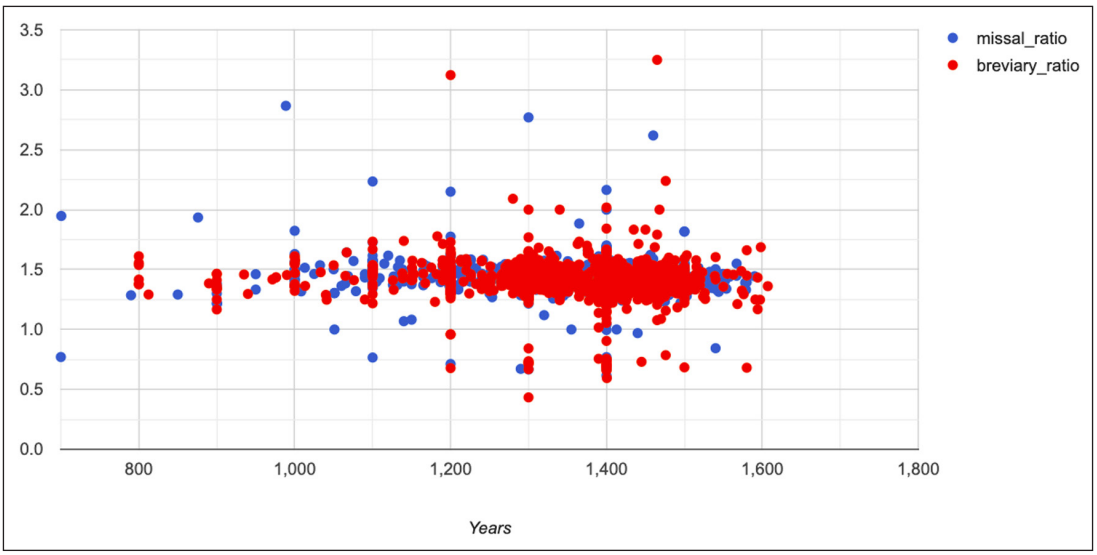


Figure 16: Height-to-width ratios of breviaries and missals (outliers removed).

5.4.4 Lessons learned

§65 Neither MMM nor the source datasets provide information about the categories or subjects of works, so liturgical manuscripts had to be identified by keyword searches on uniform titles. Fortunately, these are generally common to Latin, English, and French, such as missal/missale, antiphonal/antiphonarium, breviary/breviarium, and so on. The initial query produced a single set of ratios regardless of the specific type of liturgical manuscript; later refinement visualized the ratios for the specific types separately, enabling comparisons between them.

§66 Dimensions are likely to have multiple values in the SDBM, reflecting different descriptions from different observations of the same manuscript. The same kind of variation can also be found for the same manuscript in two or three of the data sources. We dealt with this by averaging the height and the width across the different values.

§67 Some problems were identified with the source data, including records that had height but not width, and some cases where mm and cm measurements were mixed together. These could produce incorrect ratios, since the query works by adding up the raw figures and then dividing by the number of values.

§68 Production date ranges are often approximate, for example, "1300–1400" or "1225–1250." We dealt with this initially by taking the earliest date in the date range, that is, "1300" and "1225" in these two cases. Further refinement of this query involved calculating an average for production date ranges (e.g., 1400–1450 as 1425), to avoid results bunching together at 1400 for 15th-century manuscripts.

§69 Several outliers were noticeable in **Figure 16**, including one with a ratio of 30 (not shown). These were checked to see if they reflected an error in the source data, but the extreme outlier was found to be a roll rather than a codex, an unexpected result that could challenge assumptions about the use and readership of liturgical manuscripts in the Middle Ages. Our choice to remove outliers from the results meant that a more granular display of results in Yasgui became possible, but at the expense of a fuller and more accurate representation of variations in the data as the roll breviary indicates. Further, excluding outlying values for height and width actually affected the ratio calculations for some manuscripts and produced incorrect values. Excluding outlying ratios might be a better way of achieving this goal.

§70 As originally formulated, the query obscured whether height or width was the larger dimension, since the ratio was constructed by dividing the larger dimension by the smaller one, regardless of which was the height or width. The resulting ratios were

always 1.0 or greater. A different formulation of the query was required to show the ratio of height to width consistently; the results then included ratios lower than 1.0, in cases where a codex was wider than it was long. Choosing between these queries depends on the ultimate goal of the research: is it simply to find the average relative proportions of a manuscript, or is it examining the orientation and layout of the pages as well?

§71 The resulting scatter plot showing ratios for 12,169 individual manuscripts, coloured according to their type, provided a very effective visual representation of a relatively large body of data. But these queries also made clear the importance of consistent approaches to recording this kind of data and documenting the assumptions made in analyses of the data.

5.5 Query 5: How long did the bookseller James Tregaskis keep manuscripts in his stock?

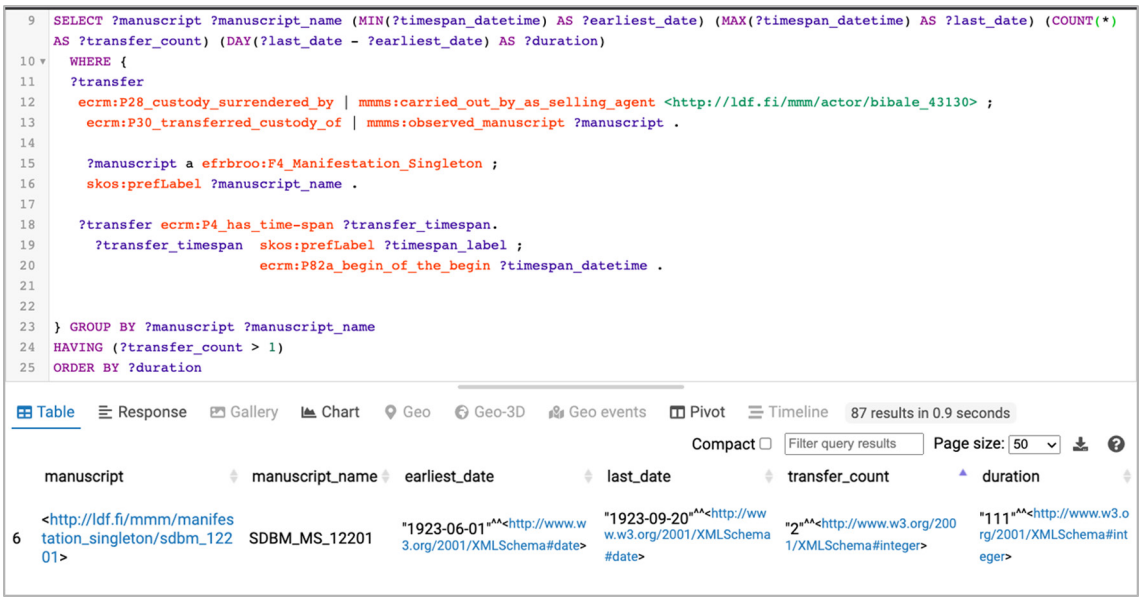
§72 The next query, derived from the work of the Cultivate MSS project, considers the length of time books remained in the stock of a particular dealer. In this case, we looked at the London dealer James Tregaskis, who was a prolific producer of catalogues, many of which have been entered into the SDBM as part of the project and are now searchable as LOD within the MMM portal (Worms 2016).

§73 A manuscript might appear in Tregaskis' catalogues multiple times a year, allowing the duration of a manuscript's time in his stock to be traced with a relatively high degree of precision. This is particularly valuable because, unlike some other firms (notably J. & J. Leighton), no sales records are known to survive. Tregaskis' activities therefore have to be reconstructed from his catalogues and records of his purchases at auctions. The SDBM allows for records pertaining to a single manuscript to be linked to a manuscript record, but it is difficult to compare those manuscript records. SPARQL provides the potential to calculate the length of time a manuscript remained in Tregaskis' stock and to compare these figures. Comparing this with the same information for a larger and longer-lived firm like Bernard Quaritch Ltd. would help to assess the significance of the Tregaskis data.

§74 Tregaskis' catalogues provide price data for each manuscript as it is offered for sale. It is therefore possible to track changes in the prices asked for a manuscript over time. However, in the time period covered by the catalogues in the SDBM (1892-1936), Great Britain was not using a decimal currency. Moreover, Tregaskis expressed prices in both pounds, shillings and pence, and guineas (a guinea was £1 1s). Using SPARQL to query price movements over time is therefore not feasible without some normalization of the raw price data.

5.5.1 Query explanation

5.5.1.1 Step 1 (Figure 17): Manuscripts sold by Tregaskis, their dates of transfer, their transfer counts, and the number of days they stayed in Tregaskis' stock: https://api.triplydb.com/s/euJRw2LfK



```
SELECT ?manuscript ?manuscript_name (MIN(?timespan_datetime) AS ?earliest_date) (MAX(?timespan_datetime) AS ?last_date) (COUNT(*)
AS ?transfer_count) (DAY(?last_date - ?earliest_date) AS ?duration)
  WHERE {
  ?transfer
   ecrm:P28_custody_surrendered_by | mmms:carried_out_by_as_selling_agent <http://ldf.fi/mmm/actor/bibale_43130> ;
    ecrm:P30_transferred_custody_of | mmms:observed_manuscript ?manuscript .

    ?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 ?manuscript ?manuscript_name
HAVING (?transfer_count > 1)
ORDER BY ?duration
```

Table | Response | Gallery | Chart | Geo | Geo-3D | Geo events | Pivot | Timeline | 87 results in 0.9 seconds

Compact | Filter query results | Page size: 50

	manuscript	manuscript_name	earliest_date	last_date	transfer_count	duration
6	<http://ldf.fi/mmm/manifestation_singleton/sdbm_12201>	SDBM_MS_12201	"1923-06-01"^^<http://www.w3.org/2001/XMLSchema#date>	"1923-09-20"^^<http://www.w3.org/2001/XMLSchema#date>	"2"^^<http://www.w3.org/2001/XMLSchema#integer>	"111"^^<http://www.w3.org/2001/XMLSchema#integer>

Figure 17: SPARQL query for Query 5: Step 1.

§75 This query includes several calculations in its `SELECT` statement to determine the amount of time manuscripts remained in Tregaskis' stock. The `MIN` aggregate function extracts the earliest date in a manuscript's transfer timespan `(MIN(?timespan_datetime) AS ?earliest_date)`. An identical strategy calculates the last date in the same timespan with the `MAX` function `(MAX(?timespan_datetime) AS ?last_date)`. With these two new variables, `?earliest_date` and `?last_date`, the `DAY` function can calculate the duration of time a manuscript remained in Tregaskis' possession `(DAY(?last_date - ?earliest_date) AS ?duration)`. The `COUNT` function calculates the number of times each manuscript appeared in a Tregaskis catalogue as the `?transfer_count` variable.

5.5.1.2 Step 2 (Figure 18): Duration and transfer count of Quaritch stock https://api.triplydb.com/s/0BBppvlWj

§76 Step 2 mirrors the process used in Step 1 to find transfers associated with Bernard Quaritch Ltd., but reduces the amount of information displayed so that a scatter-plot visualization becomes possible. The `SELECT` statement is reduced to two calculated

variables: duration and transfer count (line 9). The MAX and MIN calculations are included within the DAY function to calculate duration. The URI for Quaritch is swapped for Tregaskis in lines 12–13, and the transfer count is limited to those manuscripts with 2 or more transfers (line 25).

```
9  SELECT  (DAY(MAX(?timespan_datetime) - MIN(?timespan_datetime)) AS ?duration) (COUNT(*) AS ?transfer_count)
10   WHERE {
11   ?transfer
12     ecrm:P28_custody_surrendered_by | mmms:carried_out_by_as_selling_agent <http://ldf.fi/mmm/actor/bibale_30748> ; #quaritch
13     ecrm:P30_transferred_custody_of | mmms:observed_manuscript ?manuscript .
14
15     ?manuscript a efrbroo:F4_Manifestation_Singleton ; #excludes collections
16     skos:prefLabel ?manuscript_name .
17
18    ?transfer ecrm:P4_has_time-span ?transfer_timespan.
19      ?transfer_timespan  skos:prefLabel ?timespan_label ;
20                          ecrm:P82a_begin_of_the_begin ?timespan_datetime .
21
22 } GROUP BY ?manuscript_name
23 HAVING (?transfer_count > 1)
24 ORDER BY ?duration
```

Figure 18: SPARQL query for Query 5: Step 2.

5.5.1.3 Step 3 (Figures 19a–b): Tregaskis and Quaritch duration and transfers compared https://api.triplydb.com/s/RY-FOOqM4

```
9  SELECT  ?duration ?tregaskis_transfer_count  ?quaritch_transfer_count
10 WHERE {
11   { SELECT (COUNT (?tregaskis_transfer) AS ?tregaskis_transfer_count) (DAY(MAX(?timespan_datetime) - MIN(?timespan_datetime)) AS ?
   duration)
12     WHERE {
13       ?tregaskis_transfer
14         ecrm:P28_custody_surrendered_by | mmms:carried_out_by_as_selling_agent <http://ldf.fi/mmm/actor/bibale_43130> ; #tregaskis
15         ecrm:P30_transferred_custody_of | mmms:observed_manuscript ?manuscript .
16       ?manuscript a efrbroo:F4_Manifestation_Singleton ;
17         skos:prefLabel ?manuscript_name .
18
19       ?tregaskis_transfer ecrm:P4_has_time-span ?transfer_timespan.
20       ?transfer_timespan  skos:prefLabel ?timespan_label ;
21                           ecrm:P82a_begin_of_the_begin ?timespan_datetime .
22
23     } GROUP BY ?manuscript
24     HAVING (?tregaskis_transfer_count > 1)
25   }
```

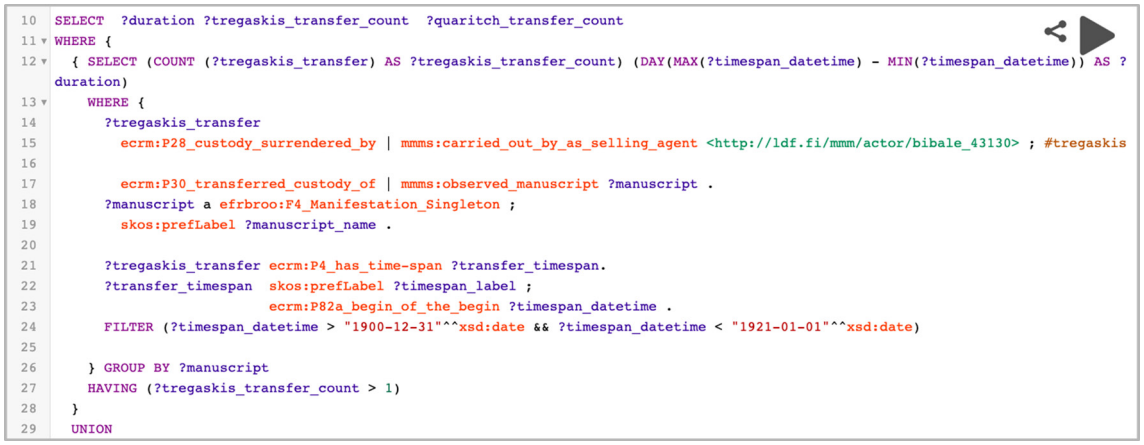
Figure 19a: SPARQL query for Query 5: Step 3 (lines 9–25, relating to Tregaskis).

```
24     HAVING (?tregaskis_transfer_count > 1)
25   }
26   UNION
27   { SELECT (COUNT (?quaritch_transfer) AS ?quaritch_transfer_count) (DAY(MAX(?timespan_datetime) - MIN(?timespan_datetime)) AS ?
   duration)
28     WHERE {
29       ?quaritch_transfer
30         ecrm:P28_custody_surrendered_by | mmms:carried_out_by_as_selling_agent <http://ldf.fi/mmm/actor/bibale_30748> ; #quaritch
31         ecrm:P30_transferred_custody_of | mmms:observed_manuscript ?manuscript .
32       ?manuscript a efrbroo:F4_Manifestation_Singleton ;
33         skos:prefLabel ?manuscript_name .
34
35       ?quaritch_transfer ecrm:P4_has_time-span ?transfer_timespan.
36       ?transfer_timespan  skos:prefLabel ?timespan_label ;
37                           ecrm:P82a_begin_of_the_begin ?timespan_datetime .
38     } GROUP BY ?manuscript
39     HAVING (?quaritch_transfer_count > 1 )
40   }
41
42 }
43 ORDER BY ?duration
```

Figure 19b: SPARQL query for Query 5: Step 3 (lines 24–43, relating to Quaritch).

§77 Step 3 of the query brings together the results of the previous two queries for easier comparison. This involves creating two similar sub-queries—one for Tregaskis (lines 11 to 27) and one for Quaritch (lines 29 to 43), combining them with a `UNION` command (line 28), and displaying the duration and the transfer counts from the two sub-queries in an overarching `SELECT` statement (line 9). The transfer count in each sub-query is limited to those manuscripts with 2 or more transfers (lines 26 and 42).

5.5.1.4 Step 4 (Figures 20a–b): Comparison of Tregaskis and Quaritch stock between 1901–1920
https://api.triplydb.com/s/syyzeyQ_q



```
10 SELECT  ?duration ?tregaskis_transfer_count  ?quaritch_transfer_count
11 WHERE {
12   { SELECT (COUNT (?tregaskis_transfer) AS ?tregaskis_transfer_count) (DAY(MAX(?timespan_datetime) - MIN(?timespan_datetime)) AS ?
   duration)
13     WHERE {
14       ?tregaskis_transfer
15         ecrm:P28_custody_surrendered_by | mmms:carried_out_by_as_selling_agent <http://ldf.fi/mmm/actor/bibale_43130> ; #tregaskis
16
17         ecrm:P30_transferred_custody_of | mmms:observed_manuscript ?manuscript .
18       ?manuscript a efrbroo:F4_Manifestation_Singleton ;
19         skos:prefLabel ?manuscript_name .
20
21       ?tregaskis_transfer ecrm:P4_has_time-span ?transfer_timespan.
22       ?transfer_timespan  skos:prefLabel ?timespan_label ;
23                           ecrm:P82a_begin_of_the_begin ?timespan_datetime .
24       FILTER (?timespan_datetime > "1900-12-31"^^xsd:date && ?timespan_datetime < "1921-01-01"^^xsd:date)
25
26     } GROUP BY ?manuscript
27     HAVING (?tregaskis_transfer_count > 1)
28   }
29   UNION
```

Figure 20a: SPARQL query for Query 5: Step 4 (lines 10–29, relating to Tregaskis).

```
28   }
29   UNION
30   { SELECT (COUNT (?quaritch_transfer) AS ?quaritch_transfer_count) (DAY(MAX(?timespan_datetime) - MIN(?timespan_datetime)) AS ?
   duration)
31     WHERE {
32       ?quaritch_transfer
33         ecrm:P28_custody_surrendered_by | mmms:carried_out_by_as_selling_agent <http://ldf.fi/mmm/actor/bibale_30748> ; #quaritch
34         ecrm:P30_transferred_custody_of | mmms:observed_manuscript ?manuscript .
35       ?manuscript a efrbroo:F4_Manifestation_Singleton ;
36         skos:prefLabel ?manuscript_name .
37
38       ?quaritch_transfer ecrm:P4_has_time-span ?transfer_timespan.
39       ?transfer_timespan  skos:prefLabel ?timespan_label ;
40                           ecrm:P82a_begin_of_the_begin ?timespan_datetime .
41       FILTER (?timespan_datetime > "1900-12-31"^^xsd:date && ?timespan_datetime < "1921-01-01"^^xsd:date)
42     } GROUP BY ?manuscript
43     HAVING (?quaritch_transfer_count > 1 )
44   }
45
46 }
47 ORDER BY ?duration
```

Figure 20b: SPARQL query for Query 5: Step 4 (lines 28–47, relating to Quaritch).

§78 Step 4 is designed to limit the comparison between Tregaskis and Quaritch to a period when they were both active: between 1901 and 1920. This is done by adding

a statement to each sub-query (at lines 25 and 43) to filter the timespan for values after 31 December 1900 and before 1st January 1921: `FILTER (?timespan_datetime > "1900-12-31"^^xsd:date && ?timespan_datetime < "1921-01-01"^^xsd:date)`

5.5.1.5 Step 5 (Figure 21): An improved scatter-plot visualization https://api.triplydb.com/s/qyGoY07li

```
9  SELECT ("" AS ?manuscript_name_sample) ?duration ?transfer_count ?seller (COUNT(?manuscript_name) AS ?manuscript_count)
10   WHERE {
11   {
12 SELECT ?manuscript_name (DAY(MAX(?timespan_datetime) - MIN(?timespan_datetime)) AS ?duration)  (COUNT(*) AS ?transfer_count) ?
   seller
13   WHERE {
14  { ?transfer
15    ecrm:P28_custody_surrendered_by | mmms:carried_out_by_as_selling_agent <http://ldf.fi/mmm/actor/bibale_43130> ; #tregaskis
16    ecrm:P30_transferred_custody_of | mmms:observed_manuscript ?manuscript .
17     BIND ("tregaskis" AS ?seller)
18   }
19   UNION
20   { ?transfer
21     ecrm:P28_custody_surrendered_by | mmms:carried_out_by_as_selling_agent <http://ldf.fi/mmm/actor/bibale_30748> ; #quaritch
22     ecrm:P30_transferred_custody_of | mmms:observed_manuscript ?manuscript .
23     BIND ("quaritch" AS ?seller)
24   }
25
26     ?manuscript a efrbroo:F4_Manifestation_Singleton ;
27     skos:prefLabel ?manuscript_name .
28
29    ?transfer ecrm:P4_has_time-span ?transfer_timespan.
30      ?transfer_timespan  skos:prefLabel ?timespan_label ;
31                          ecrm:P82a_begin_of_the_begin ?timespan_datetime .
32
33 } GROUP BY ?seller ?manuscript_name
34 HAVING (?transfer_count > 1) (?duration > 0)
35     }
36 } GROUP BY ?seller ?duration ?transfer_count
```

Figure 21: SPARQL query for Query 5: Step 5.

§79 Step 5 of this query is designed to address a significant limitation in the scatter-plot visualizations: one coloured dot could hide several manuscripts with the same duration and number of transfers (e.g., two transfers and zero days duration). We wanted to use a bubble chart to show the relative frequency of each combination of duration and transfers. This involved re-working the query to match the pattern of variables required for a bubble chart: (1) Text – the label for each bubble; (2) Numeric – X axis; (3) Numeric – Y axis; (4) Text – determines the colour of bubbles; (5) Numeric – determines the relative size of bubbles.

§80 The query uses two sub-queries to find transfers associated with Tregaskis or Quaritch, and binds the relevant name as the seller (lines 14 to 19; 21 to 26). The sub-queries are joined with a `UNION` command (line 20). We then find the manuscripts involved in these transfers and the dates of the transfers (lines 28 to 33). The results are limited to those with a transfer count greater than one, and a duration greater than zero days (line 36). The calculation of transfer counts and durations is done in a `SELECT` statement at line 12.

§81 To construct the pattern of variables required for the bubble chart, an outer `SELECT` statement is added (line 9). This also counts the number of manuscripts with the

same combination of duration and number of transfer counts. The manuscript names, although required for the bubble chart, have been replaced with a blank space enclosed between quotation marks, since their inclusion would make the chart unreadable.

5.5.2 Results

§82 Step 1 of the query produces 87 results, with transfer counts ranging from 2 to 8, and durations ranging from zero to 3,927 days. The transfer dates range from 1900 to 1935.

§83 Step 2 of the query produces 750 results, with transfer counts ranging from 2 to 12, and durations ranging from zero to 36,159 days (99 years). The results can now be visualized as a scatter plot (**Figure 22**).

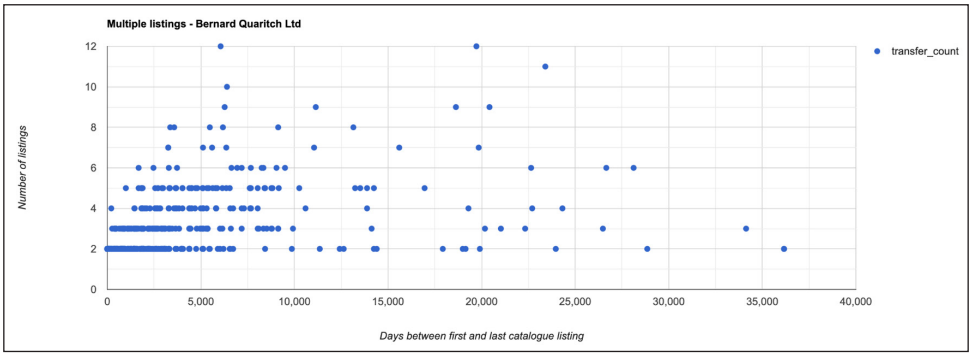


Figure 22: Multiple catalogue listings by Bernard Quaritch Ltd.

§84 Step 3 of the query produces 837 results, with transfer counts ranging from 2 to 12, and durations ranging from zero to 36,159 days (99 years). This is a simple addition of the separate Tregaskis and Quaritch results, which can now be distinguished and compared on the same visualization, with Tregaskis manuscripts shown in blue and Quaritch in red (**Figure 23**).

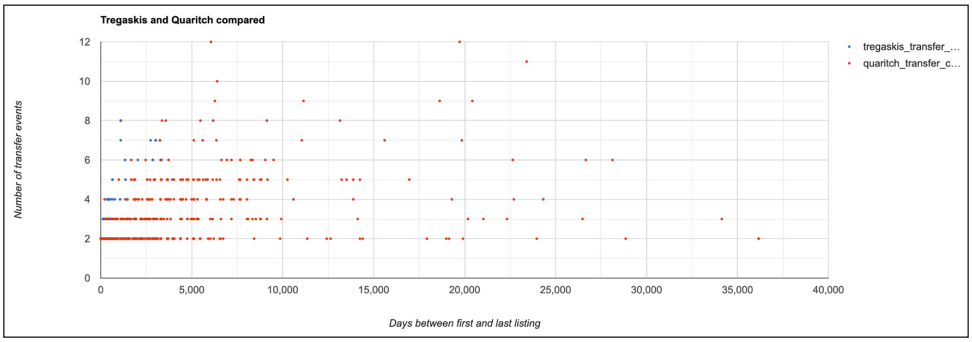


Figure 23: Tregaskis and Quaritch transfer counts and durations compared.

§85 The durations between first and last listings are much greater for Quaritch, as are the number of listings, but does this reflect anything more than the much longer time period over which this firm has operated? In some cases, Quaritch had bought back (and re-sold) a manuscript originally sold by the firm some decades earlier, so the manuscript was not actually kept in stock for the whole period in question.

§86 Step 4 of the query produced 203 results for manuscript transfers between 1901 and 1920. The maximum duration was 6,605 days (18 years), and the maximum number of transfers during these 20 years was eight. This visualization makes it clear that Tregaskis was likely to list the same manuscript many more times than Quaritch during this period, and usually within a significantly shorter period of time (**Figure 24**).

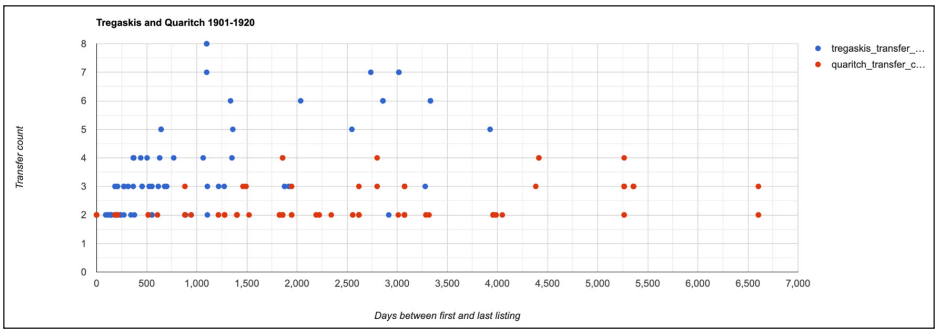


Figure 24: Comparison between Tregaskis and Quaritch 1901–1920.

§87 Step 5 of this query produces a bubble chart in which 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 (**Figure 25**).

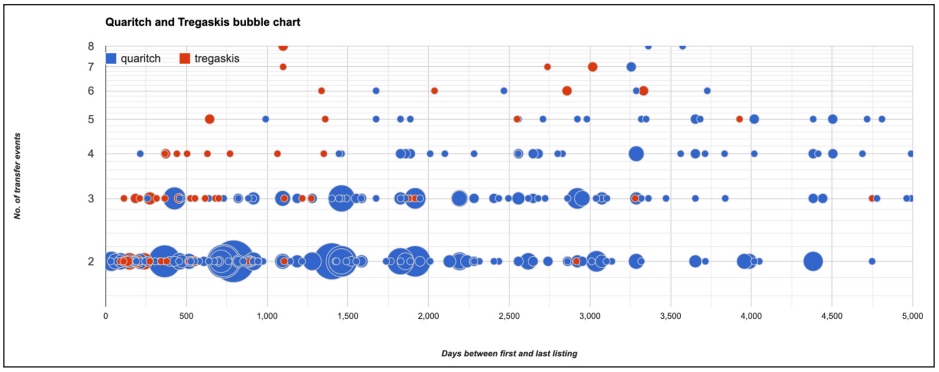


Figure 25: Bubble chart comparing Quaritch and Tregaskis.

5.5.3 Lessons learned

§88 SPARQL can be used with the MMM data to find and compare patterns in the stock retention and catalogue listings of manuscripts by dealers like Tregaskis and Quaritch over multiple years. Visualizations in the form of scatter plots and bubble charts are a valuable way of displaying this information; in the case of bubble charts, four different variables can be combined in the same chart. This cannot be done with the Sampo-UI interface to MMM, nor in the interfaces of the three source datasets.

§89 Nevertheless, these visualizations—and the underlying SPARQL results—need to be treated with some caution, since they conceal various assumptions about the data. There is no way of distinguishing manuscripts that were sold, bought back, and sold again by Quaritch from those that were kept in stock for a number of consecutive years and advertised in multiple catalogues during that period. The duration in stock is simply calculated from the earliest listing to the last recorded listing. Manuscripts with a duration of zero days between two listings may have been advertised twice in the same year, without a specific day or month being recorded, but these entries may also reflect two versions of the same catalogue or stock list entered separately in the Schoenberg Database.

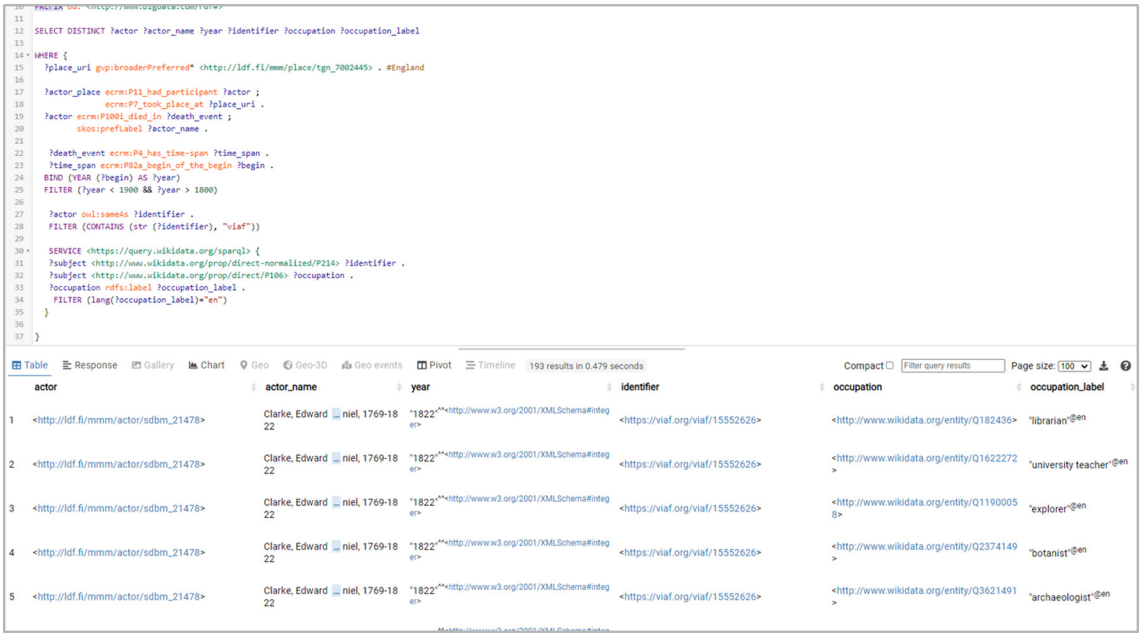
5.6 Query 6: What is known about the social backgrounds of 19th- and 20th-century British collectors?

§90 Interested in researching the social backgrounds of 19th- and 20th-century manuscript collectors, author Toby Burrows raised the possibility of using a federated query to add information from external sources like Wikidata to an MMM SPARQL query. A federated query in SPARQL connects one endpoint to any other openly available endpoint, thus greatly expanding the possibilities for finding associations among datasets not otherwise obviously connected by topic or content. For Burrows' purposes, the MMM data on its own could not provide information about the occupation, gender, life events, and other data that would build a fuller picture about the lives of these collectors and provide more insight into their habits and motivations for collecting. Mechanisms such as the shared use of identifiers from resources like the Virtual International Authority File (VIAF; http://viaf.org/) that are included when available in the name authority metadata associated with people and institutions in both MMM and Wikidata provide one of the easier ways to execute a federated search and present an opportunity to pull the personal data of persons and institutions provided in Wikidata entries together into MMM query results.

5.6.1 Query explanation

§91 Query 6 (**Figure 26**): https://api.triplydb.com/s/44t3wQf0g. This federated query combines information from the MMM and Wikidata datasets with a single SPARQL

query sent to the MMM endpoint. Line 15 limits the geographical scope of this query to England. We then find the actors associated with events occurring in England (lines 17–18), together with their death events and names (lines 20–21). The dates of these death events are then filtered for those occurring before 1900 and after 1800 (lines 23 to 26). Actors are then limited to those with VIAF identifiers (lines 28–29). Line 28 equivocates the `?actor` and `?identifier` variables by using the `owl:sameAs` predicate, which is how the query will connect the VIAF data between the MMM and Wikidata datasets. These VIAF identifiers are then passed to Wikidata using the `SERVICE` keyword. To find the corresponding "person" record in Wikidata, lines 31–32 return Wikidata resources that have VIAF identifiers that match the identifiers returned for the MMM actors above, along with their occupations. The `FILTER` on line 35 returns the occupation labels in English, rather than any language that may appear in the Wikidata data.



```
11
12 SELECT DISTINCT ?actor ?actor_name ?year ?identifier ?occupation ?occupation_label
13
14 WHERE {
15   ?place_uri gvp:broaderPreferred* <http://ldf.fi/mmm/place/tgn_7002445> . #England
16
17   ?actor_place ecrm:P11_had_participant ?actor ;
18                ecrm:P7_took_place_at ?place_uri .
19   ?actor ecrm:P100i_died_in ?death_event ;
20          skos:prefLabel ?actor_name .
21
22    ?death_event ecrm:P4_has_time-span ?time_span .
23    ?time_span ecrm:P82a_begin_of_the_begin ?begin .
24   BIND (YEAR (?begin) AS ?year)
25   FILTER (?year < 1900 && ?year > 1800)
26
27    ?actor owl:sameAs ?identifier .
28    FILTER (CONTAINS (str (?identifier), "viaf"))
29
30    SERVICE <https://query.wikidata.org/sparql> {
31     ?subject <http://www.wikidata.org/prop/direct-normalized/P214> ?identifier .
32     ?subject <http://www.wikidata.org/prop/direct/P106> ?occupation .
33     ?occupation rdfs:label ?occupation_label .
34      FILTER (lang(?occupation_label)="en")
35    }
36
37 }
```

Table | Response | Gallery | Chart | Geo | Geo-3D | Geo events | Pivot | Timeline | 193 results in 0.479 seconds | Compact | Filter query results | Page size: 100

	actor	actor_name	year	identifier	occupation	occupation_label
1	<http://ldf.fi/mmm/actor/sdbm_21478>	Clarke, Edward Daniel, 1769-1822	"1822"^^<http://www.w3.org/2001/XMLSchema#integer>	<https://viaf.org/viaf/15552626>	<http://www.wikidata.org/entity/Q182436>	"librarian"@en
2	<http://ldf.fi/mmm/actor/sdbm_21478>	Clarke, Edward Daniel, 1769-1822	"1822"^^<http://www.w3.org/2001/XMLSchema#integer>	<https://viaf.org/viaf/15552626>	<http://www.wikidata.org/entity/Q1622272>	"university teacher"@en
3	<http://ldf.fi/mmm/actor/sdbm_21478>	Clarke, Edward Daniel, 1769-1822	"1822"^^<http://www.w3.org/2001/XMLSchema#integer>	<https://viaf.org/viaf/15552626>	<http://www.wikidata.org/entity/Q11900058>	"explorer"@en
4	<http://ldf.fi/mmm/actor/sdbm_21478>	Clarke, Edward Daniel, 1769-1822	"1822"^^<http://www.w3.org/2001/XMLSchema#integer>	<https://viaf.org/viaf/15552626>	<http://www.wikidata.org/entity/Q2374149>	"botanist"@en
5	<http://ldf.fi/mmm/actor/sdbm_21478>	Clarke, Edward Daniel, 1769-1822	"1822"^^<http://www.w3.org/2001/XMLSchema#integer>	<https://viaf.org/viaf/15552626>	<http://www.wikidata.org/entity/Q3621491>	"archaeologist"@en

Figure 26: Query 6.

§92 The results show each actor's MMM ID and name, together with their year of death, their VIAF identifier, and their occupation identifier and occupation (the latter two from Wikidata).

5.6.2 Results

§93 The query produces 205 results with 91 distinct names of people with death dates in the 19th century, who have an average of two occupations, though some have

many more than this: 19 in the case of William Morris! There are a total of 83 different occupations. Politicians (21) and writers (21) are the most common, though there is also an astrologer, a brewer, and at least three slave holders. (Readers can find these results by clicking on the Yasgui link for Query 6.) We thus are presented with a cross section of occupations associated with those with the means and motivations to collect manuscripts in the 19th century.

5.6.3 Lessons learned

§94 As this query shows, Wikidata can be a valuable source of additional information about people and institutions in the MMM dataset that is not otherwise captured by the source datasets, in this case, the occupations of individual collectors. The results show that the personal, professional, and academic interests of collectors of premodern European manuscripts in the 19th century are diverse and sometimes surprising, including "singer-songwriter" or "science fiction writer." The results may also show a certain bias. For example, why are there so many occupations associated with William Morris compared to other collectors? Is it because he was that much more active than anyone else, or that as a seminal figure in the Arts and Crafts movement, we have simply collected more data about him than other 19th-century manuscript collectors?

§95 The question of bias cannot be ignored as it has implications for how we collect data and, in this case, data about people. Well-known people or institutions will have more data about their lives associated with them in online resources. But it is also interesting to note that women are not included in these results, though we know that there were women involved in the book trade in Britain with death dates before 1900. (For example, Henrietta Katherine Burrell, recorded in the SDBM Name Authority: https://sdbm.library.upenn.edu/names/40365/.) Why is this so? The simple answer is that the overlap of persons with the same VIAF identifiers in both Wikidata and MMM is small. Indeed, while there are 56,685 actors (persons and organizations) in MMM, only a fraction have VIAF identifiers. At present, MMM has more than 15,300 VIAF identifiers for actors, but only 4,400 Wikidata identifiers.

§96 This lack of representation in VIAF could be due in large part to a systemic lack of recognition for the contributions that women have made in the book trade in the 19th century and in society in general. Following these results, we performed a similar query that asked to return actors with the same VIAF number but were identified in Wikidata as female. The best set of results was found among women collectors in the United States born between 1900 and 1950: https://api.triplydb.com/s/OZlC0ieHo, which returns 14 results showing 9 different women, with occupations ranging from librarian, book collector, and archaeologist to politician, statistician, and lawyer, among others.

§97 As these results show, this query strategy requires both the MMM person and the Wikidata person to have a shared VIAF identifier to return results. Our results point to the broader problem of the lack of representation of a large number of actors in available authorities and LOD resources. A systematic import of Wikidata identifiers into MMM (or into the source datasets) would increase results, but the problem will not be fully addressed until actors in underrepresented social groups and minorities are given better data representation in these resources.

6 Conclusion

§98 The weekly SPARQL workshop held by the MMM project began as a knowledge transfer activity designed to teach the practical skill of learning how to perform SPARQL queries, but gradually developed into a wider investigation of the use of SPARQL to analyze the data, explore broader types of research questions, and assess the research potential of the MMM aggregated dataset and its Knowledge Graph. The benefits of investing over 500 hours of staff time in learning and practicing SPARQL queries can be seen in various ways, beginning with a diagnostic approach to identifying limitations in the data aggregated by the MMM project. This includes areas (like the different types of events) where the data sources do not enable an optimum level of granularity in the MMM data model. The source datasets do not collect the same information or, sometimes, when they *do* collect the same information, it is not computationally accessible via the same methods. This is more than a matter of improved mapping and transformation. Information that is explicit in one dataset may be only inferred from another. Discrete pieces of information in one source may be stored in aggregated form in another.

§99 Like most collection-based humanities datasets and their interfaces, the MMM data sources are designed to produce lists of items (manuscripts) meeting certain criteria, rather than supporting statistical analyzes. The price data in the SDBM, for example, are purely descriptive and do not provide an adequate basis for quantitative analysis, even within a SPARQL query. On the other hand, some contextual information that is outside the scope of the source datasets can be added on-the-fly in SPARQL queries, as our work with person data from Wikidata shows. This also reinforced the importance of Linked Open Data identifiers in enabling this kind of approach and raised some significant questions about future strategies for including identifiers in datasets like those used by MMM.

§100 There are 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 (nodegoat 2021). 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. (See this query: https://api.triplydb.com/s/u_-KEd-US.) 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).

§101 More generally, the workshop resulted in a better understanding of how querying data in a computational context works. For the humanists on the team, learning the technical language and structures of SPARQL also showed them how to develop more ambitious approaches to the MMM data, transforming the traditional research questions that had shaped the initial data modelling work into more sophisticated and expansive queries that took full advantage of the MMM data model. As a result, the returned data from these queries better reflected the true value of the combined dataset for humanistic research. For the computer scientists, the more evolved approach to querying led to more understanding about the complex research questions that are of interest to manuscript researchers, and to better analysis to determine the success of the project.

§102 As these case studies show, querying the MMM dataset via its SPARQL endpoint does not produce perfect results, or results that provide a definitive answer in the traditional sense to the research questions. The methodology presented in these case studies follows the principles of distant reading, whereby computational aggregation and analysis of the data presented in returned results brings new insights into and raises new questions about the nature of the data and the subject it represents—in this case pre-modern manuscripts (Moretti 2013). While one would not want to draw hard conclusions from the results achieved in these queries, we hope to have shown that the process of learning and experimenting in a SPARQL environment brings three important benefits: 1) a better understanding of a complex and imperfect dataset, 2) a better understanding of how manuscript description and associated data involving the

people and institutions involved in the production, reception, and trade of premodern manuscripts needs to be presented to better facilitate computational research, and 3) an awareness of the need to further develop data literacy skills among researchers in order to take full advantage of the wealth of unexplored data now available to them in the Semantic Web (Koltay 2015).

Acknowledgements

This work was funded by the Trans-Atlantic Platform under its Digging into Data Challenge (https://diggingintodata.org) for 2017–2020. The Mapping Manuscript Migration project was led by the University of Oxford, in partnership with the University of Pennsylvania, Aalto University, and Helsinki Centre for Digital Humanities (HELDIG) at the University of Helsinki, and the Institut de recherche et d'histoire des textes (IRHT). The authors wish to acknowledge CSC–IT Center for Science, Finland, for computational resources. The transformation of the Oxford Manuscript data into RDF builds upon earlier work by the OXLOD project. The authors acknowledge the contributions of the following: Antoine Brix (IRHT), Petri Leskinen (Aalto University), Synnøve Myking (IRHT), Pierre-Louis Pinault (IRHT), and Jouni Tuominen (University of Helsinki).

Competing interests

LR currently serves as the Director of *Digital Medievalist*; her tenure on the board ends July 2022.

Contributions

Authorial contributions

Authorship is alphabetical after the drafting author and principal technical lead. Author contributions, described using the CASRAI CredIT typology, are as follows:

The corresponding author is: Lynn Ransom (lr)

List of contributors and roles in alphabetical order

- Toby Burrows: tb
- Laura Cleaver: lc
- Doug Emery: de
- Eero Hyvönen: eh
- Mikko Koho: mk
- Lynn Ransom: lr
- Emma Thomson: et
- Hanno Wijsman: hw

- Conceptualization: tb; lc; eh; de; mk; lr; et
- Methodology: tb; lc; de; eh; mk; lr; et
- Investigation: tb; lc; de; mk; lr; et; hw
- Writing – Original Draft Preparation: tb; lc; de; mk; lr; et
- Writing – Review & Editing: tb; de; eh; lr; et
- Visualization: tb; mk
- Supervision: tb; eh; lr; hw
- Project Administration: tb; eh; lr; hw
- Funding Acquisition: tb; eh; lr; hw

Editorial contributions
Recommending editors:
Mike Kestemont, University of Antwerp, Belgium

Recommending referees:
Tiziana Mancinelli, Ca' Forscari Università Venezia, Italy
Roman Bleier, University of Graz, Austria

Section/copy/layout editors:
Morgan Pearce, The Journal Incubator, University of Lethbridge, Canada
Christa Avram, The Journal Incubator, University of Lethbridge, Canada

References

Burrows, Toby, Nicole Bergk Pinto, Mahaut Cazals, Alexandre Gaudin, and Hanno Wijsman. 2020. "Evaluating a Semantic Portal for the 'Mapping Manuscript Migrations' Project." *DigItalia: Rivista del Digitale nei Beni Culturali* 2, 178–185. Accessed May 3, 2022. http://digitalia.sbn.it/article/view/2643.

Davis, Lisa Fagin (@lisafdavis). 2020. "Voilà! 3,413 data points, height/width over time for several different genres of liturgical manuscripts. Data from @schoenbergdb (Caveat! Some of these data points are duplicates, since each database record is an observation of a particular manuscript at a particular time)." Twitter, December 5, 8:09 a.m. Accessed May 3, 2022. https://twitter.com/lisafdavis/status/1335239769765392386".

Doerr, Martin. 2003. "The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata." *AI Magazine*, 24(3): 75–92. https://doi.org/10.1609/aimag.v24i3.1720.

DuCharme, Bob. 2013. *Learning SPARQL: Querying and Updating with SPARQL 1.1*. 2nd ed. Sebastopol, CA: O'Reilly.

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*, 1(1): 1–136. Accessed February 17, 2022. DOI: https://doi.org/10.2200/S00334ED1V01Y201102WBE001

Hyvönen, Eero, Esko Ikkala, Mikko Koho, Jouni Tuominen, Toby Burrows, Lynn Ransom, and Hanno Wijsman. 2021. "Mapping Manuscript Migrations on the Semantic Web: A Semantic Portal and Linked Open Data Service for Premodern Manuscript Research." In *Proceedings of the 20th International Joint Conference of Semantic Web* (ISWC 2021), virtual, October 24–28, 615–630. New York: Springer. DOI: https://doi.org/10.1007/978-3-030-88361-4_36

Ichinose, Shiori, Ichiro Kobayashi, Michiake Iwazume, and Kuogi Tanaka. 2014. "Ranking the Results of Dbpedia Retrieval with SPARQL Query." In *JIST 2013: Semantic Technology* (Lecture Notes in Computer Science, vol 8388), edited by Wooju Kim, Ying Ding, and Hong-Gee Kim. 306–319. Cham: Springer. DOI: https://doi.org/10.1007/978-3-319-06826-8_23

Ikkala, Esko, Eero Hyvönen, Heikki Rantala, and Mikko Koho. 2021. "Sampo-UI: A Full Stack JavaScript Framework for Developing Semantic Portal User Interfaces." *Semantic Web* 13(1): 69–84. DOI: https://doi.org/10.3233/SW-210428

Koho, Mikko, Toby Burrows, Eero Hyvönen, Esko Ikkala, Kevin Page, Lynn Ransom, Jouni Tuominen, Doug Emery, Arthur Mitchell Fraas, Benjamin Heller, David Lewis, Andrew Morrison, Guillaume Porte, Emma Thomson, Athanasios Velios, and Hanno Wijsman. 2021. "Harmonizing and Publishing Heterogeneous Premodern Manuscript Metadata as Linked Open Data." *Journal of the Association for Information Science and Technology* 73(2): 240–257. DOI: https://doi.org/10.1002/asi.24499

Koltay, Tibor. 2015. "Data Literacy for Researchers and Data Librarians." *Journal of Librarianship and Information Science* 49(1): 3–14. DOI: https://doi.org/10.1177/0961000615616450

Lincoln, Matthew. 2014. "SPARQL for Humanists." *Matthew Lincoln, PhD* [blog]. 10 July. Accessed February 17, 2022. https://matthewlincoln.net/2014/07/10/sparql-for-humanists.html.

———. 2015. "Using SPARQL to access Linked Open Data." *Programming Historian*. Accessed February 17, 2022. https://programminghistorian.org/en/lessons/retired/graph-databases-and-SPARQL.

Meroño-Peñuela, Albert, Ashkan Ashkpour, Marieke van Erp, Kees Mandemakers, Leen Breure, Andrea Scharnhorst, Stefan Schlobach, and Frank van Harmelen. 2015. "Semantic Technologies for Historical Research: A Survey." *Semantic Web* 6(6): 539–564. Accessed May 3, 2022. http://www.semantic-web-journal.net/sites/default/files/swj301.pdf. DOI: https://doi.org/10.3233/SW-140158

Moretti, Franco. 2013. *Distant Reading*. Verso Books.

nodegoat. 2021. "nodegoat Workshop Series Organised by the SNSF SPARK Project 'Dynamic Data Ingestion.'" *nodegoat* (blog), April 6. Accessed May 3, 2022. https://nodegoat.net/blog.p/82.m/54/nodegoat-workshop-series-organised-by-the-snsf-spark-project-dynamic-data-ingestion.

Penny, Ralph. 2002. *A History of the Spanish Language*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511992827

Rietveld, Laurens, and Rinke Hoekstra. 2017. "The YASGUI Family of SPARQL Clients." *Semantic Web* 8(3): 373–383. DOI: https://doi.org/10.3233/SW-150197

Riva, Pat, Martin Doerr, and Maja Žumer. 2009. "FRBRoo: Enabling a Common View of Information from Memory Institutions." *International Cataloguing and Bibliographic Control* 38(2), 30–34. Accessed February 17, 2022. https://archive.ifla.org/IV/ifla74/papers/156-Riva_Doerr_Zumer-en.pdf.

Schweizer, Tobias, and Benjamin Geer. 2021. "Gravsearch: Transforming SPARQL to Query Humanities Data," *Semantic Web* 12(6): 379–400. DOI: https://doi.org/10.3233/SW-200386

Sharpe, Richard. 2003. Titulus: Identifying Medieval Latin Texts, an Evidence-Based Approach. Turnhout: Brepols.

Smith, Innocent (@InnocentOP). 2020. "Leafing through Emmanuel Borque's Etude sur les sacramentaires romains (published in the 40s and 50s), I'm struck by how he neglects to give any details about the physical aspects of the manuscripts, e.g. size and number of folios." Twitter, December 5, 7:13 a.m. Accessed May 3, 2022. https://twitter.com/InnocentOP/status/1335225723859169282.

Staab, Steffan, and Studer, Rudi, eds. 2009. *Handbook of Ontologies*. 2nd ed. Berlin: Springer-Verlag. DOI: https://doi.org/10.1007/978-3-540-92673-3

Tauberer, Joshua. 2006. “What Is RDF.” *XML.com*. Accessed May 3, 2022. https://www.xml.com/pub/a/2001/01/24/rdf.html.

Worms, Laurence. 2016. “James Tregaskis.” *Antiquarian Booksellers’ Association*. https://aba.org.uk/page/james-tregaskis.