

Exploring Knowledge Graphs for Medieval and Renaissance Manuscripts with SPARQL

Toby Burrows, University of Western Australia, AU; University of Oxford, UK, toby.burrows@uwa.edu.au

This article reports on research into the use of SPARQL for querying knowledge graphs, in the form of Linked Open Data triple stores, which relate to the history and provenance of medieval manuscripts. It looks at several recent initiatives and projects which rely on RDF and SPARQL, including Mapping Manuscript Migrations and the Wikibase implementation of Digital Scriptorium. As well as comparing and evaluating different ways of using SPARQL with manuscript data, the article looks at possible future directions in this field, notably the development of visual interfaces for SPARQL queries and the potential use of AI chatbots to formulate SPARQL queries.



Introduction

SPARQL is the query language used with knowledge graphs constructed in accordance with the World Wide Web Consortium's standards for Linked Open Data (LOD) (W3C SPARQL Working Group, 2013; DuCharme, 2013). These knowledge graphs consist of sentence-like statements (RDF triples) about entities and their properties (in the form: subject – predicate – object), defined in terms of one or more declared ontologies and vocabularies. These graphs represent a complex body of interconnected knowledge and require a relatively complex methodology for querying and exploring them, which is what SPARQL is designed to provide.

There are various other ways of structuring this kind of knowledge, ranging from relational databases to other kinds of graph databases, document databases, and key-value databases. The advantages and disadvantages of RDF triple stores, in comparison with these other approaches, have been widely discussed (Hogan et al., 2022; Groth et al., 2023). The assumption here is that RDF and SPARQL offer a sophisticated and valuable way of representing and querying data about medieval and Renaissance manuscripts, which can serve to answer complex research questions about the history and provenance of manuscripts. The Mapping Manuscript Migrations project, discussed in more detail below, found that SPARQL queries could answer a set of specific research questions considerably more effectively and fully than interfaces based on relational databases or TEI-encoded documents (Burrows et al., 2021).

SPARQL can be deployed in several different ways. It can be run against a SPARQL endpoint using a third-party Web service like Yasgui.¹ It can also be run using the native query interface to an endpoint, as is the case with Wikidata and other Wikibase implementations. Given the relative complexity and expertise required to construct these kinds of queries, other approaches have also been developed. One is to embed SPARQL queries into other forms of software, such as Sampo-UI which is used for the Mapping Manuscript Migrations service.² There have also been attempts to develop visual interfaces for constructing SPARQL queries. Some initial experiments have also been carried out in getting AI chatbots like ChatGPT to write SPARQL queries.

In recent years, several knowledge graphs of data about medieval and Renaissance manuscripts have been published to the Web which use the Linked Open Data and RDF framework. The most notable of these take the form of a standalone RDF triple store and graph database, with a SPARQL endpoint. The Mapping Manuscript Migrations

¹ The Yasgui service can be found at: <https://yasgui.triply.cc/> [Last Accessed 14 July 2024].

² The Mapping Manuscript Migrations site is at: <https://mappingmanuscriptmigrations.org/en/> [Last Accessed 14 July 2024].

project is the fullest purpose-built example of this approach and is discussed in more detail below. The Schoenberg Database of Manuscripts, which is a customized relational database, has also transformed its data to RDF and made them available through a public SPARQL endpoint.³ The Bibale database of manuscript provenance data has also transformed its data to RDF, but does not offer a SPARQL endpoint.⁴ The Biblissima service includes a standalone triple store, but its SPARQL endpoint is for internal use only and has not been made publicly available.⁵

There have also been various efforts to create Linked Open Data knowledge graphs of the content of some specific types of manuscripts, including notarial archives, Old English texts, books of hours, and music manuscripts. Projects along these lines include Searobend,⁶ Sphaera,⁷ MusicKG,⁸ and NotaryPedia.⁹ With the exception of MusicKG (Eyharabide et al., 2019), they have not made a public endpoint available for SPARQL queries.

Mapping Manuscript Migrations

The Mapping Manuscript Migrations (MMM) project aggregated manuscript provenance data exported from the *Schoenberg Database*, *Bibale*, and *Medieval Manuscripts in Oxford Libraries*. They were converted to RDF triples and mapped to the MMM Data Model, which was based on a combination of the FRBROO and CIDOC-CRM ontologies, with some additional entity classes and properties. The aggregated data contained more than 900,000 provenance events and over 20 million RDF triples, which could be explored through a SPARQL endpoint as well as a browsing and filtering interface (Koho et al., 2022).

This interface uses the Sampo-UI software developed at Aalto University in Finland (Ikkala et al., 2022). It leverages the relationships in the MMM Data Model to enable an extensive range of filter combinations. These might include finding manuscripts of St Augustine's *De civitate Dei* produced in France before 1200 and last recorded in

³ The SPARQL endpoint for the Schoenberg Database can be accessed at: <https://sdbm.library.upenn.edu/sparql-space> [Last Accessed 14 July 2024].

⁴ The Bibale database can be accessed at: <https://bibale.irht.cnrs.fr/> [Last Accessed 14 July 2024].

⁵ Information about the Biblissima SPARQL endpoint can be found at: <https://doc.biblissima.fr/api/sparql/> [Last Accessed 14 July 2024].

⁶ Information about the Searobend project can be found at: <https://searobend.adaptcentre.ie> [Last Accessed 14 July 2024].

⁷ The Sphaera database can be accessed at: <https://db.sphaera.mpiwg-berlin.mpg.de/resource/Start> [Last Accessed 14 July 2024].

⁸ Information about MusicKG can be found on GitHub at: <https://github.com/victoriaeyharabide/MusicKG> [Last Accessed 14 July 2024].

⁹ The NotaryPedia project can be found at: <https://notarypedia.opendatamalta.com/> [Last Accessed 14 July 2024].

North America. The results can be displayed as lists and in the form of map-based visualizations of production places, last-known locations, and the movements between them.¹⁰ Sampo-UI is built on embedded SPARQL queries, which are not initially visible to the user. It is possible, however, to inspect the underlying queries as well as to amend them and re-run them directly.

The MMM project also tested the capabilities of direct queries against its SPARQL endpoint, both in comparison with the Sampo-UI interface, and in comparison with the native browsing and querying interfaces of the three source datasets (Burrows, Cleaver et al., 2021 ; Burrows, Cleaver et al., 2022). In particular, the SPARQL endpoint enabled more quantitative and longitudinal queries, which could not be done through the Sampo-UI interface, let alone in the original databases. These queries included tracing variations in the physical dimensions of different types of liturgical manuscripts over the centuries, which showed that the ratio between height and width varied most in Books of Hours, and that this variation increased over time. Breviaries and missals were more consistent and uniform in their layout. Another query measured rates of stock retention by manuscript dealers in 20th-century Great Britain, revealing that the antiquarian bookseller Bernard Quaritch Ltd. kept manuscripts in stock for a much longer period, on average, and advertised them less frequently, than the rival London firm owned by James Tregaskis.

Queries run against the MMM SPARQL endpoint were also extended with contextual data from Wikidata to show the occupations and birthplaces of 19th-century British manuscript collectors. Birthplaces were visualized on a map to give an indication of the number of collectors who came from North-Western industrial centres rather than London and the South-East. Visualizations of queries like these are dependent on the SPARQL query tool used; the MMM project used the Yasgui tool to produce charts, graphs, and bubble charts, as well as timelines.¹¹ The results can also be exported to interfaces which can produce map-based and network visualizations. The MMM project tested the export of SPARQL query data to software like ResearchSpace and nodegoat; the latter environment includes a time slider which can show change over time, a feature not available in Sampo-UI or Yasgui (Burrows et al., 2020).

An important aspect of the MMM project's experience with SPARQL was the time and training required to produce these results. Taking real research queries, members of the project team worked through them at weekly online workshops over a period of 18 months, with guidance of a SPARQL expert. Related outputs include a SPARQL tutorial

¹⁰ The Mapping Manuscript Migrations site is at: <https://mappingmanuscriptmigrations.org/en/> [Last Accessed 14 July 2024].

¹¹ TheYasgui service can be found at: <https://yasgui.triply.cc/> [Last Accessed 14 July 2024].

based on the MMM data, and published versions of the SPARQL queries themselves as well as the result sets.¹²

Wikidata

Wikidata is one of the world’s biggest public knowledge graphs, with more than 124 million items and more than 1.44 billion statements or triples (Vrandečić et al., 2023).¹³ The Wikidata Query Service is one of the most active SPARQL endpoints, with more than 10 million queries per day as of April 2021.¹⁴ Its coverage of manuscripts is still relatively limited, although some institutions—notably the Bodleian Library, the Koninklijke Bibliotheek (KB) of the Netherlands, and the National Library of Wales—have been running projects to upload manuscript records (about 1,300, 1,260 and 590 respectively). The KB has also created Wikidata records for about 490 *alba amicorum* (“friendship books”).¹⁵ The schema for manuscript records is somewhat confused and inconsistent, but a recent project is aiming to improve this (Poulter, 2021).¹⁶ In any case, most existing records are far from being full catalogue records; their main contents are inventory numbers (including shelf marks), external identifiers, and links to digitized versions.

The existing corpus of records supports only limited in-depth SPARQL-based semantic exploration and reasoning, especially of a quantitative kind, since the content of the records varies considerably. Most of the entries from the National Library of Wales, for example, include production (‘inception’ in Wikidata terminology) dates, languages, dimensions, number of pages, genre, and main subject. Most of the Bodleian Library entries do not give any of these, with the exception of production dates and languages. A small group of the Bodleian records shows previous owners, while only one of the Welsh manuscripts (Black Book of Carmarthen) does. The KB helpfully provides a series of sample SPARQL queries for exploring its manuscript entries, including a timeline visualization using Histropedia.¹⁷

¹² A SPARQL tutorial can be found on GitHub at: https://mapping-manuscript-migrations.github.io/sparql/sparql_tutorial.html [Last Accessed 14 July 2024]. The SPARQL data can be accessed on Zenodo at: <https://zenodo.org/record/5796988> [Last Accessed 14 July 2024].

¹³ Wikidata statistics can be found at: <https://www.wikidata.org/wiki/Wikidata:Statistics> [Last Accessed 14 July 2024].

¹⁴ Statistics for the number of Wikidata triples can be found on Wikimedia at: <https://grafana.wikimedia.org/d/000000489/wikidata-query-service?viewPanel=7&orgId=1&refresh=1m> [Last Accessed 14 July 2024].

¹⁵ The Alba Amicorum project can be found on Wikidata at: https://www.wikidata.org/wiki/Wikidata:WikiProject_Alba_amicorum_National_Library_of_the_Netherlands [Last Accessed 14 July 2024].

¹⁶ The Wikidata Manuscripts Project can be accessed at: https://www.wikidata.org/wiki/Wikidata_talk:WikiProject_Manuscripts [Last Accessed 14 July 2024].

¹⁷ SPARQL information for the National Library’s Medieval Manuscripts project can be found on Wikidata at: https://www.wikidata.org/wiki/Wikidata:WikiProject_Medieval_manuscripts_National_Library_of_the_Netherlands#S-

In combination with other SPARQL endpoints, Wikidata can be queried for contextual data about the people and organizations connected with manuscripts, as demonstrated in the MMM investigations above, as well as for the works, texts, and images the manuscripts carry, and their topics, subjects, and knowledge structures. Wikidata's ability to bring together different identifiers relating to the same manuscript could also be leveraged in the future to join up multiple descriptions of the same manuscript, representations (digital or physical) of that manuscript, and the provenance evidence and scholarship relating to that manuscript.

Wikibase

The MediaWiki software used for Wikidata has been made available as an Open Source download, as well as through the Wikibase Cloud service hosted by the German Wikimedia community.¹⁸ Each Wikibase includes a SPARQL endpoint and a query interface. Among the groups experimenting with the Wikibase Cloud service are the British Library's Archives and Manuscripts, the Koninklijke Bibliotheek, the Medieval Mining Texts project, and the Mapping Manuscript Migrations project.¹⁹ None of these projects have gone much beyond creating lists of properties and some basic manuscript records, though most provide one or more sample SPARQL queries for testing.

The exception is the redeveloped Digital Scriptorium, a union catalogue of medieval and Renaissance manuscripts in North American libraries (Koho et al., 2023). The data model used for this project is object-centric (rather than event-centric like Mapping Manuscript Migrations) and focuses on the manuscript metadata record. As part of the evaluation of the prototype of this service, a set of 24 SPARQL queries was developed and run. These were intended to mimic searching and browsing, either by finding manuscript records with a specific value for a given metadata element (name, place, subject, and so on), or by finding all manuscript records with any value for that element. The SPARQL queries used for this prototype testing have been made available through a GitHub repository.²⁰

PARQL [Last Accessed 14 July 2024]. The timeline for the National Library's Alba Amicorum project can be found on Wikidata at: https://www.wikidata.org/wiki/Wikidata:WikiProject_Alba_amicorum_National_Library_of_the_Netherlands#Timeline [Last Accessed 14 July 2024].

¹⁸ Wikibase Cloud can be found at: <https://www.wikibase.cloud/> [Last Accessed 14 July 2024].

¹⁹ The list of Wikibase Cloud sites can be accessed at: <https://www.wikibase.cloud/discovery> [Last Accessed 14 July 2024].

²⁰ SPARQL queries for Digital Scriptorium can be found on GitHub at: <https://github.com/DigitalScriptorium/ds-testing/tree/main/sparql> [Last Accessed 14 July 2024].

Visual Interfaces for SPARQL Queries

As the amount of available Linked Open Data increases, and the number of SPARQL endpoints grows, there is a growing need for advanced methods of searching and browsing across these knowledge graphs. Only in this way can the full richness and sophistication of the data models be unlocked. Two developments for the future refinement and improvement of SPARQL querying are of significance here. The first involves visual interfaces, and the second focuses on AI-based approaches to query formulation. Of these, the former is more advanced, while the latter has only really emerged since ChatGPT was publicly released in November 2022.

In this context, the term ‘visual interfaces’ refers to visual methods for constructing and conceptualizing SPARQL queries, as distinct from the visual presentation of knowledge graphs through software like Sampo-UI. Most of these interfaces work specifically with the Wikidata SPARQL endpoint. Wikidata itself offers a Query Helper, which provides a way of creating queries using drop-down menus to find and combine properties and entity classes.²¹ More recently, it also added a Query Builder for creating simple SPARQL queries by combining properties from drop-down menus.²²

In November 2023, two leading Semantic Web companies—metaphacts and Ontotext—combined to launch a free public interface to the Wikidata knowledge graph, combining their respective metaphactory and GraphDB products.²³ Starting with a keyword search for a specific entity or class, it is then possible to explore graphically the various connections step-by-step. Although there are limitations in the number of connections which can be displayed, and familiarity with the Wikidata entity and property schemas is very helpful, this service does provide a useful visual approach to researching manuscript collections (as far as they are documented in Wikidata). You can build provenance trails for individual manuscripts and collections, for example, as well as explore the overlapping connections between manuscript owners and their membership of relevant organizations. Results can also be displayed as a list or explored for direct connections using the ‘Pathfinder’ feature.

This product bears some resemblance to the ResearchSpace software, originally developed by metaphacts in association with the British Museum. Designed specifically to work with the CIDOC-CRM ontology, ResearchSpace also enables queries to be built

²¹ The Wikidata Query Helper can be accessed at: https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/Query_Helper [Last Accessed 14 July 2024].

²² The Query Service for Wikidata can be accessed at: <https://query.wikidata.org/querybuilder/?uselang=en> [Last Accessed 14 July 2024].

²³ The Metaphacts demonstrator for Wikidata can be found at: <https://metaphacts.com/metaphacts-ontotext-interface-for-wikidata-demo-system> [Last Accessed 14 July 2024].

up visually from an entity or class by prompting the user with the available properties at each step of the graph. The Mapping Manuscript Migrations project experimented with a version of ResearchSpace, importing all the MMM data, setting up some basic query paths, and producing some visualizations. The ResearchSpace community is primarily associated with museum collections, though there have been a couple of manuscript-related projects.²⁴

There have been various other efforts to develop visual query builders for SPARQL, sometimes as research projects and sometimes for specific institutions or collections. An interesting recent example from the library world is SPARNATURAL, which includes demonstrators for the Bibliothèque nationale de France and the Archives nationales de France (Francart, 2023). The look of the interface derives ultimately from ResearchSpace but it can be configured to work with various data models and endpoints. A useful feature is the ability to display the actual SPARQL query.

Using AI for SPARQL Queries

The public release of chatbots based on Large Language Models (LLMs), beginning with ChatGPT in November 2022, has prompted a great deal of discussion and experimentation around their implications for building and exploring knowledge graphs (Idehen, 2023; Petkova, 2023). Since these AI services have been, amongst many other things, promoted as a means of carrying out elaborate and complex human tasks like writing, composing music, and image-making, it might also be feasible for a chatbot to enable a non-expert user to construct SPARQL queries, without the degree of knowledge required to create such queries from scratch. One of the examples given on the ChatGPT ‘Overview’ page is ‘write an SQL query’, so we might expect AI chatbots to tackle SPARQL queries too.²⁵

GPT-4 (the most sophisticated version of the GPT family, available in a free public version in Microsoft’s Copilot)²⁶ can be prompted to construct SPARQL queries for the Wikidata Query Service. I asked GPT-4 to ‘write a Wikidata SPARQL query to find manuscripts owned by persons who became members of the Roxburghe Club in the 19th century, displaying their names and membership dates where available’. In the resulting SPARQL query (Appendix 1), the entity reference (Wikidata QID) given for the Roxburghe Club was wrong, as was one of the properties, and the query asked for dates of birth rather than dates of membership. The same request was run four times in Copilot/GPT-4, which returned a different query structure each time. None of these

²⁴ Users of ResearchSpace are listed at: <https://researchspace.org/who-is-using-researchspace/> [Last Accessed 14 July 2024].

²⁵ The ChatGPTPlus service can be accessed at: <https://openai.com/chatgpt> [Last Accessed 14 July 2024].

²⁶ Copilot can be accessed at: <https://copilot.microsoft.com/> [Last Accessed 14 July 2024].

queries ran successfully against the Wikidata Query Service, even after the entity references and properties were corrected. This experiment showed that while GPT-4 knows enough about SPARQL and Wikidata to write a plausible query, it seems to have a problem with identifying QIDs for specific entities, as well as with the Wikidata query syntax. The resulting queries needed thorough debugging which required a significant level of SPARQL knowledge and was not something that a SPARQL beginner could simply run successfully.

Trying to extend GPT-4's SPARQL knowledge to endpoints other than Wikidata also proved problematic. I asked Copilot to 'write a SPARQL query to run against the 'Mapping Manuscript Migrations' SPARQL endpoint, to find all manuscripts (manifestation singletons) owned by Henry Yates Thompson'. I also told it to use the classes, properties, and namespace prefixes listed in the MMM schema.²⁷ The response (Appendix 2) was a SPARQL query which failed to define two of its prefixes (ecrm and mmm). When these definitions were added, the query still failed to produce any results, partly because the wrong property was being used to link owner and manuscript, and partly because the wrong method was being used to specify Thompson as the owner.

I also asked GPT-4 to write a SPARQL query to run against the Digital Scriptorium Wikibase, giving it the address of the SPARQL endpoint and asking for a query to find all manuscripts held by the University of Pennsylvania. I also referred it to the classes, properties, and namespace prefixes listed in Koho et al., 2023 for the query syntax. The response (Appendix 3) was unsatisfactory in various ways and produced no results:

- The QID for the University of Pennsylvania was incorrect;
- The property used was incorrect;
- The URL defining the 'ds:' prefix does not exist.

At present, then, while GPT-4 can provide a starting-point for constructing Wikidata SPARQL queries, these still require careful review and debugging. It appears to have difficulty formulating workable queries for other SPARQL endpoints. These were only basic, preliminary investigations, however, and better results might well be achieved using other LLMs like Mistral or Claude, through advanced prompting techniques and customization, or by fine-tuning the LLMs for specific tasks, perhaps in conjunction with a SPARQL validator. The level of technical knowledge involved in implementing

²⁷ The Mapping Manuscript Migrations schema can be found at: https://mapping-manuscript-migrations.github.io/data_model/mmm-schema [Last Accessed 14 July 2024].

these approaches is likely to be beyond the capability of the typical humanities researcher or librarian and would probably require investigation by appropriate experts.

Given that incorporating knowledge graphs into LLM training is a hot topic in generative AI research, it is highly likely that future iterations of LLMs will be able to respond more effectively to this kind of task. A more immediately promising development would seem to be the integration of LLMs into software like metaphactory, a commercial solution for building and exploring LOD knowledge graphs. In October 2023, a metaphactory app was released for beta testing, which translates natural language queries into SPARQL queries.²⁸ While it uses the public version of ChatGPT for this purpose, the translation process can be controlled from within the metaphactory software.

Issues Affecting the Usefulness of SPARQL

The Mapping Manuscript Migrations (MMM) experience demonstrated how the effectiveness and usefulness of SPARQL queries can be affected by uncertainties in data and data structures relevant to medieval and Renaissance manuscripts. There are many ambiguities inherent in manuscript descriptions: production dates are often only estimates, for example, while production places can usually only be assigned at a regional or country level, sometimes involving multiple possibilities. Other uncertainties relate to gaps in the chain of provenance evidence, such as markings in a manuscript which record only a name or family crest. Dealing with fragments of manuscripts is another challenge, as is the splitting and combining of manuscript volumes over time.

Ambiguity in data models is another critical factor. The *Schoenberg Database* records the appearance of manuscripts in sales and auction catalogues but does not usually include information about whether the manuscript was sold, or to whom. This inherent ambiguity meant that the CIDOC-CRM ‘Transfer of Custody’ event class (E10) could not be automatically applied to these occurrences. A generic MMM-specific entity class ‘ManuscriptActivity’ had to be created specifically for these events; this affected about two-thirds of all provenance events in the MMM data. Similarly, MMM managed to extract more than 20,000 provenance statements from the Bodleian Library’s catalogue (Burrows et al., 2021). But only the TEI <origin> element could be mapped to a specific event type in the CIDOC-CRM ontology: E12_Production. All other provenance events had to be mapped to the ‘ManuscriptActivity’ class. This significantly limited the degree of specificity with which the Oxford data could be analysed and visualized.

²⁸ Documentation for Metaphactory 5.1 can be accessed at: <https://metaphacts.com/metaphactory-5-1> [Last Accessed 14 July 2024].

In Wikidata's data model, authors should not be attached directly to manuscripts, but be linked indirectly via their works contained in a manuscript. This approach has not been adopted consistently, however. The relationship between works and authors is also problematic in the *Schoenberg Database*, where each of the multiple works in a single manuscript is not specifically linked to an author, rather a list of the multiple authors merely sits alongside the list of works. These approaches make it significantly more complicated to construct SPARQL queries involving authors of works carried by manuscripts.

There are a number of possible ways to improve the value of manuscript data for analysis through SPARQL queries (Burrows, 2023). Using a suitable vocabulary for different types of provenance event would be valuable, though the typology provided by the CIDOC-CRM ontology might need to be extended to cover situations such as 'offered for sale' (Bekiari et al., 2021:64–7). Adding roles to the occurrences of persons and organizations in provenance statements would support more granular and nuanced analyses. For CIDOC-CRM, some additions might be needed to extend its limited set of properties, which currently include 'P51_has_former_or_current_owner' and 'P52_has_current_owner'.

Other improvements might include using date ranges (1300–1325) rather than text-based statements such as 'circa 1310', or using an uncertainty model like CIDOC CRM's Temporal Relation Primitives (Bekiari et al., 2021:43–6). Embedding Uniform Resource Identifiers (URIs) from widely used vocabularies for entities like persons, organizations, and places—such as VIAF, Wikidata, and GeoNames—could also enrich queries around authors, collectors and owners. Publishing Linked Open Data vocabularies for medieval and Renaissance persons, places, and works would help with this (Burrows, 2022). A universal approach to assigning manuscript identifiers (such as that proposed by the ISMI initiative) is vital for linking disparate data relating to a specific manuscript (Bougard et al., 2020).

Conclusion

Knowledge graphs constructed with RDF triples and the Linked Open Data framework have enabled the development of rich and complex datasets about medieval and Renaissance manuscripts in recent years. While these datasets may have user-friendly interfaces for searching and browsing, exploring their full potential ultimately needs a SPARQL endpoint and the SPARQL query language. SPARQL queries can explore sophisticated relationships within the data as well as enabling linked queries across multiple datasets of this kind. The complexity of the data structures is likely to require a complex query, however, and

SPARQL is somewhat notorious for its apparent level of difficulty. A significant amount of training is recommended, together with access to expert advice.

Several projects—notably Mapping Manuscript Migrations and Digital Scriptorium 2.0—have used SPARQL queries systematically and successfully for testing, exploration, and diagnosis. These approaches have demonstrated SPARQL’s potential value for manuscript studies. Various initiatives are underway to improve the usability of SPARQL through the development of visual query interfaces and the embedding of queries into other software environments. While AI chatbots like GPT-4 may be able to help with writing SPARQL queries, they are currently quite limited in their ability to do this.

Wikidata and its associated Wikibase projects are the most significant area of development at present, although the relatively limited amount of manuscript data currently only supports fairly basic SPARQL queries. Once the Wikidata data model for manuscripts is improved and more libraries start contributing data, there is significant potential for sophisticated SPARQL queries through the Wikidata Query Service and its Wikibase equivalents.

Appendices


COPILOT
NOTEBOOK
Get the app

You

write a Wikidata SPARQL query to find manuscripts owned by persons who became members of the Roxburghe Club in the 19th century, displaying their names and membership dates where available

Copilot

Certainly! The **Roxburghe Club**, established in 1812, is the oldest and most distinguished society dedicated to printing unpublished documents and reprinting rare printed texts, including lesser-known works of English literature and history ¹. Let's construct a **Wikidata SPARQL query** to identify manuscripts owned by individuals who were members of the Roxburghe Club during the 19th century. We'll display their names and membership dates where available.



```

PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX p: <http://www.wikidata.org/prop/>
PREFIX ps: <http://www.wikidata.org/prop/statement/>
PREFIX pq: <http://www.wikidata.org/prop/qualifier/>

SELECT ?memberLabel ?manuscriptLabel ?membershipDate
WHERE {
  ?member wdt:P31 wd:Q5 ; # Instance of human
  wdt:P463 wd:Q10850700 ; # Member of the Roxburghe Club
  wdt:P569 ?birthDate . # Date of birth
  FILTER(YEAR(?birthDate) >= 1800 && YEAR(?birthDate) < 1900)
  OPTIONAL {
    ?manuscript wdt:P195 ?member ; # Manuscript owned by the member
    rdfs:label ?manuscriptLabel .
    ?member wdt:P571 ?membershipDate . # Date of membership
  }
  SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }
}
ORDER BY ?memberLabel

```

This query retrieves information about Roxburghe Club members born between 1800 and 1899, along with any manuscripts they owned and their membership dates (if available). The results will display the member's name, manuscript name, and membership date. Feel free to execute this query on Wikidata to explore further ²!

Learn more [1](#) [roxburgheclub.org.uk](#) [2](#) [roxburgheclub.org.uk](#) [3](#) [anthempres.com](#) [4](#) [core.ac.uk](#) [5](#) [academic.oup.com](#) +8 more

Appendix 1: GPT-4 MMM query 1.

You

Write a SPARQL query to find and list all manuscripts (manifestation singletons) owned by Henry Yates Thompson, using the Mapping Manuscript Migrations (MMM) SPARQL endpoint and the the classes, properties, and namespace prefixes listed at https://mapping-manuscript-migrations.github.io/data_model/mmm-schema

Copilot

Certainly! Let's construct a **SPARQL query** to find manuscripts (manifestation singletons) owned by **Henry Yates Thompson** using the **Mapping Manuscript Migrations (MMM)** SPARQL endpoint. We'll follow the MMM Data Model and utilize the provided namespace prefixes.

- First, we'll define the necessary prefixes:
 - PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema>
 - PREFIX efrbroo: <http://erlangen-crm.org/efrbroo/>
 - PREFIX ecrm: <http://erlangen-crm.org/current/>
 - PREFIX mms: <http://ldf.fi/mms/schema/>
- Next, we'll query for manuscripts owned by Henry Yates Thompson:

```

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema>
PREFIX efrbroo: <http://erlangen-crm.org/efrbroo/>
SELECT ?manuscript
WHERE {
  ?manuscript a efrbroo:F4_Manifestation_Singleton .
  ?manuscript ecrm:P30_transferred_custody_of|mms:observed_manuscript ?owner.
  ?owner rdfs:label "Henry Yates Thompson" .
}

```

- Execute this query against the MMM triplestore at the **MMM SPARQL endpoint**. It will retrieve manuscripts associated with Henry Yates Thompson. Feel free to explore the fascinating world of manuscripts! ²!

For more details, you can refer to the [MMM Data Model](#) and the [MMM Schema](#). Happy querying! ³!

Learn more [1](#) [mapping-manuscript-migrations.github.io](#) [2](#) [blog.mappingmanuscriptmigrations.org](#) [3](#) [openhumanitiesdata.metajni.com](#) +11 more

Appendix 2: GPT-4 MMM query 2.

You

Write a SPARQL query to run against the Digital Scriptorium Wikibase (<https://catalog.digital-scriptorium.org/query/>), to find all manuscripts held by the University of Pennsylvania. Use the classes, properties, and namespace prefixes listed at <https://dl.acm.org/doi/10.1145/3594723> for the query syntax.

Copilot

Certainly! To find all manuscripts held by the **University of Pennsylvania** in the **Digital Scriptorium Wikibase**, you can use the following SPARQL query:

```
PREFIX ds: <https://digital-scriptorium.org/ontology#>
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>

SELECT ?manuscript
WHERE {
  ?manuscript ds:holdingInstitution wd:Q4917 .
}
```

In this query:

- We use the `ds` namespace prefix to refer to the Digital Scriptorium ontology.
- The `wd` and `wdt` prefixes are used for Wikidata entities and properties, respectively.
- We select the `?manuscript` variable representing the manuscripts.
- The `ds:holdingInstitution` property is used to filter manuscripts held by the University of Pennsylvania (Wikidata ID: Q4917).

You can run this query on the [Digital Scriptorium Query Service](#) to retrieve the relevant results¹. The Digital Scriptorium is a valuable resource for exploring and querying manuscript records from multiple sources using linked open data principles². If you have any further questions or need additional assistance, feel free to ask!

Learn more [1](#) [2](#) [3](#) [4](#)

👍 🗑️ 📄 📄 🗑️ 1 of 5

Appendix 3: GPT-4 – Digital Scriptorium query.

Acknowledgements

The Mapping Manuscript Migrations project was funded by the Trans-Atlantic Partnership under its Digging into Data (Round 4) program. The partners in this project were the University of Oxford, the University of Pennsylvania, Aalto University in collaboration with University of Helsinki (HELDIG), and the Institut de recherche et d'histoire des textes. Funding was provided by the UK Economic and Social Research Council, the Institute of Museum and Library Services, the Academy of Finland, and the Agence nationale de la recherche.

Competing Interests

The author has no competing interests to declare.

References

Bekiari, C, Bruseker, G, Doerr, M, Ore, C, Stead, S and Velios, A 2021 *Definition of the CIDOC Conceptual Reference Model, Version 7.1.1*. [Paris]: International Council of Museums, International Committee for Documentation. https://www.cidoc-crm.org/sites/default/files/cidoc_crm_v.7.1.1_0.pdf [Last Accessed 3 February 2024].

Bougard, F, Cassin, M, Duba, W, Fabian, C, Flüeler, C and Turcan-Verkerk, A 2020 International Standard Manuscript Identifier (ISMI): pour un registre électronique des identifiants des livres manuscrits. *Digitalia*, 15(1): 45–53. <https://doi.org/10.36181/digitalia-00003> [Last Accessed 3 February 2024].

- Burrows, T** 2022 Linked Open Data and Medieval Studies: Some Lessons from the Mapping Manuscript Migrations Project. *International Journal of Humanities and Arts Computing*, 16(1): 64–77. <https://doi.org/10.3366/ijhac.2022.0277> [Last Accessed 14 July 2024].
- Burrows, T** 2023 Computational Study of Medieval Manuscript Provenance. In: Klemettilä, H, Niskanen, S and Willoughby, J (eds.) *Routledge Resources Online: Medieval Studies*. London: Routledge. Published online 19 June. <https://doi.org/10.4324/9780415791182-RMEO379-1> [Last Accessed 3 February 2024].
- Burrows, T, Cleaver, L, Emery, D, Hyvönen, E, Koho, M, Ransom, L, Thomson, E and Wijsman, H** 2021 Medieval Manuscripts and their Migrations: Using SPARQL to Investigate the Research Potential of an Aggregated Knowledge Graph. [Dataset]. *Zenodo*. Published online 21 December. <https://doi.org/10.5281/zenodo.5796988> [Last Accessed 3 February 2024].
- Burrows, T, Cleaver, L, Emery, D, Hyvönen, E, Koho, M, Ransom, L, Thomson, E and Wijsman, H** 2022 Medieval Manuscripts and Their Migrations: Using SPARQL to Investigate the Research Potential of an Aggregated Knowledge Graph. *Digital Medievalist* 15(1): 1–48. <https://doi.org/10.16995/dm.8064> [Last Accessed 3 February 2024].
- Burrows, T, Emery, D, Fraas, M, Hyvönen, E, Ikkala, E, Koho, M, Lewis, D, Morrison, A, Page, K, Ransom, L, Thomson, E, Tuominen, J, Velios, A and Wijsman, H** 2020 Mapping Manuscript Migrations Knowledge Graph: Data for Tracing the History and Provenance of Medieval and Renaissance Manuscripts. *Journal of Open Humanities Data*, 6: 3. <https://doi.org/10.5334/johd.14> [Last Accessed 3 February 2024].
- Burrows, T, Holford, M, Lewis, D, Morrison, A, Page, K and Velios, A** 2021 Transforming TEI Manuscript Descriptions into RDF Graphs. In: *Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing*. Norderstedt: BoD, pp. 145–156.
- DuCharme, B** 2013 *Learning SPARQL*. 2nd ed. Sebastopol, CA: O'Reilly.
- Eyharabide, V, Lully, V and Morel, F** 2019 MusicKG: Representations of Sound and Music in the Middle Ages as Linked Open Data. In: Acosta, M, Cudré-Mauroux, P, Maleshkova, M, Pellegrini, T, Sack, H and Sure-Vetter, Y (eds.) *Semantic Systems: The Power of AI and Knowledge Graphs: SEMANTICS 2019* (Lecture Notes in Computer Science, vol. 11702). Cham: Springer, pp. 57–63. https://doi.org/10.1007/978-3-030-33220-4_5 [Last Accessed 3 February 2024].
- Francart, T** 2023 Sparnatural: A Visual Knowledge Graph Exploration Tool. In: *The Semantic Web: ESWC 2023 Satellite Events: Hersonissos, Crete, Greece, May 28–June 1, 2023, Proceedings*. Berlin: Springer-Verlag, pp. 11–15. https://doi.org/10.1007/978-3-031-43458-7_2 [Last Accessed 3 February 2024].
- Groth, P, Simperl, E, van Erp, M and Vrandečić, D** 2023 Knowledge Graphs and their Role in the Knowledge Engineering of the 21st Century (Dagstuhl Seminar 22372). *Dagstuhl Reports*, 12(9): 60–120 <https://doi.org/10.4230/DagRep.12.9.60> [Last Accessed 3 February 2024]
- Hogan, A, Gutierrez, C, Cochez, M, de Melo, G, Kirrane, S, Polleres, A, Navigli, R, Ngonga Ngomo, A, Rashid, S, Schmelzeisen, L, Staab, S, Blomqvist, E, d'Amato, C, Labra Gayo, J, Neumaier, S, Rula, A, Sequeda, J and Zimmermann, A** 2022 *Knowledge Graphs* (Synthesis Lectures on Data, Semantics, and Knowledge, No. 22). Cham: Springer. <https://doi.org/10.1007/978-3-031-01918-0> [Last Accessed 3 February 2024].

Idehen, K 2023 ChatGPT and Semantic Web Symbiosis. *OpenLink Virtuoso Weblog*, June 20. <https://medium.com/virtuoso-blog/chatgpt-and-semantic-web-symbiosis-1fd89df1db35> [Last Accessed 3 February 2024].

Ikkala, E, Hyvönen, E, Rantala, H and Koho, M 2022 Sampo-UI: A full stack JavaScript framework for developing semantic portal user interfaces. *Semantic Web*, 13(1): 69–84.

Koho, M, Burrows, T, Hyvönen, E, Ikkala, E, Page, K, Ransom, L, Tuominen, J, Emery, D, Fraas, M, Heller, B, Lewis, D, Morrison, A, Porte, G, Thomson, E, Velios, A and Wijsman, H 2022 Harmonizing and Publishing Heterogeneous Premodern Manuscript Metadata as Linked Open Data. *Journal of the Society for Information Science and Technology*, 73(2): 240–257. <https://doi.org/10.1002/asi.24499> [Last Accessed 3 February 2024].

Koho, M, Coladangelo, L, Ransom, L and Emery, D 2023 Wikibase Model for Premodern Manuscript Metadata Harmonization, Linked Data Integration, and Discovery. *Journal of Computing and Cultural Heritage*, 16(3): 1–25. <https://doi.org/10.1145/3594723> [Last Accessed 3 February 2024].

Petkova, T 2023 Do Large Language Models Dream of Knowledge Graphs – Impressions from Day 2 At SEMANTiCS 2023. *Ontotext blog*, October 13. <https://www.ontotext.com/blog/do-large-language-models-dream-of-knowledge-graphs-impressions-from-day-2-at-semantics-2023/> [Last Accessed 3 February 2024].

Poulter, M 2021 Manuscripts on Wikidata: the state of the art? *Medium*, October 14. <https://medium.com/@infobomb/manuscripts-on-wikidata-the-state-of-the-art-7aeab63e0d56> [Last Accessed 3 February 2024].

Vrandečić, D, Pintscher, L and Krötzsch, M 2023 Wikidata: The Making Of. In: *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion)*, April 30–May 04, 2023, Austin, TX, USA. New York: ACM, pp. 615–24. <https://doi.org/10.1145/3543873.3585579> [Last Accessed 3 February 2024].

W3C SPARQL Working Group 2013 SPARQL 1.1 Overview: W3C Recommendation 21 March 2013. <https://www.w3.org/TR/2013/REC-sparql11-overview-20130321/> [Last Accessed 3 February 2024].

