Using SPARQL to investigate the research potential of a Linked Open Data knowledge graph: the Mapping Manuscript Migrations project

The Mapping Manuscript Migrations (MMM) project has developed a large Linked Open Data knowledge graph relating to the histories of 220,000 medieval and Renaissance manuscripts: https://mappingmanuscriptmigrations.org/ This aggregates data from three sources: the *Schoenberg Database of Manuscripts*, the *Bibale* database, and Oxford University's catalogue of medieval manuscripts. The project produced a Web portal where users can browse, search, and visualize the data. With a combination of filtering and searching, this portal supports relatively complex queries. But the project team wanted to explore still more ambitious and analytical queries, using the SPARQL query language directly with the MMM triple store.

This paper reports on how the MMM project team used SPARQL to address research questions, enhance and expand the context of the MMM data, and carry out diagnostic exploration of the contents of the source datasets. It includes worked examples of specific queries, and compares them with other ways of querying the same dataset. It also reports on the lessons learned from this process.

Weekly SPARQL training sessions were implemented as a way to transfer knowledge from the technical experts on the project to the other groups of staff involved. These included manuscript researchers, librarians, and dataset curators, as well as computer scientists who had expertise with relational databases but not with Linked Open Data and graph databases. We were also joined by a researcher running a different, large-scale manuscript provenance project, who was making extensive use of the Schoenberg Database. After regular weekly sessions for nearly two years, this kind of learning-by-doing has enabled attendees to become more confident, not only in using SPARQL syntax and commands, but also in understanding how to use the entity and property structures in the MMM data model to answer complex queries.

Initially the focus was on 25 specific research questions developed for the MMM project, which were used in designing the project's data model and building the user interface to the portal. We were especially interested in questions which could not be easily or fully answered through the browsing, searching, and filtering functions of the portal. Working through them raised several issues around translating the terms used to frame the questions into SPARQL search strategies. A research question like "What was the most popular text by a medieval author in France in the seventeenth century?", for example, forced a reconsideration what was meant by "popular" in this context. Was it "the text with the most recorded manuscript copies" or "the text found in manuscripts which changed hands most often"? Did "in France" mean "manuscripts located in France" or "manuscripts which changed hands in France" during the seventeenth century?

After working through a selection of these queries, the team then moved on to other research questions outside the original set. These included questions relating to the trade in medieval manuscripts in the later 19th and earlier 20th centuries. While they revealed the limitations of the price data in the source datasets, they also showed how SPARQL could be used to compare patterns in the buying and selling behaviour of specific manuscript dealers. We also looked at ways of extending the scope of the queries beyond the boundaries of the MMM

knowledge graph. While MMM contains data about the activities of manuscript owners, buyers, and sellers, it is limited in its contextual information about the people and organizations themselves. It may include birth and death dates, but does not usually include places of birth or death, or occupations. By using the SPARQL "Service" command to search Wikidata as well as MMM, we were able to analyse the occupations of 19th and 20th century manuscript owners, and map their places of birth, in order to see what kinds of people were collecting manuscripts and what their social background was.

Among the lessons learned was that the answers to quantitative and descriptive questions alike were shaped by the nature and scope of the source datasets. The *Schoenberg Database* is much fuller for certain decades and selling agents than others, for example, while the Oxford catalogue does not yet have full coverage of all the collectors involved (despite an enhancement project carried out by MMM). One application of this work will be to identify priority areas for future data input and enhancement in the source datasets.

The importance of unique identifiers was also emphasized, since these are the most efficient way of joining MMM queries with other sources like Wikidata and VIAF (as well as for aggregating the data in MMM itself). Analysing the existing use of identifiers within MMM can serve as a diagnostic tool for understanding which kinds of identifiers would be the most valuable additions in the future, as well as contributing to discussions about where identifiers should most effectively be created and propagated: the source datasets, MMM itself, or one of the major external sources.

SPARQL is less familiar as a tool for humanities scholars than it ought to be. While it is often mentioned as being supported by digital humanities projects which use the Linked Open Data framework, there are very few specific evaluations of SPARQL in a humanities setting.[1] The training materials on the *Programming Historian* site have been "retired" because the British Museum's SPARQL end-point, on which they rely, is no longer publicly available.[2] The work of the MMM project offers some realistic insights into the value and potential of the SPARQL query language in exploring a large humanities knowledge graph, as well as the time and resources required for learning and practicing SPARQL.

[1] One example is: Ichinose S., Kobayashi I., Iwazume M., Tanaka K. (2014) "Ranking the Results of DBpedia Retrieval with SPARQL Query." In: Kim W., Ding Y., Kim HG. (eds) *Semantic Technology: JIST 2013* (Lecture Notes in Computer Science, vol. 8388), pp. 306-319. Cham: Springer. https://doi.org/10.1007/978-3-319-06826-8_23

[2] https://programminghistorian.org/en/lessons/retired/graph-databases-and-SPARQL