Evaluating a Semantic Portal for the "Mapping Manuscript Migrations" Project

Toby Burrows - *Oxford e-Research Centre, University of Oxford*
Nicole Bergk Pinto – Mahaut Cazals – Alexandre Gaudin – Hanno Wijsman
Institut de recherche et d'histoire des textes (CNRS-IRHT)

Questo articolo riporta una valutazione del portale semantico Mapping Manuscript Migrations (MMM), che combina in una struttura Linked Open Data dati su manoscritti medievali e rinascimentali provenienti da diverse fonti. Il gruppo di ricerca ha esaminato inizialmente l'esperienza utente per i nuovi fruitori del portale:

– *Il progetto MMM è presentato con sufficiente chiarezza per i nuovi utenti?*
– *Le istruzioni per navigare nel portale sono reperibili con facilità?*
– *L'interfaccia è intuitiva? La navigazione è facile? Quanto è semplice fare ricerche?*

Sono stati poi valutati i modi in cui le ricerche potevano essere inquadrate ed eseguite, insieme alla presentazione e all'uso dei risultati, incluso l'esame dei problemi derivanti dal collegamento incrociato di dati disparati e la ricerca di eventuali errori o imprecisioni evidenti nei dati. È stata inoltre esaminata l'efficacia delle visualizzazioni basate su mappe prodotte dal software Sampo-UI. Le raccomandazioni derivanti dalla valutazione sono state utilizzate come base per migliorare la funzionalità di Sampo-UI, aggiungendosi alla guida in linea fornita agli utenti del portale e aumentando la quantità e il tipo di informazioni disponibili agli utenti avanzati, soprattutto in relazione all'ambito e copertura dei dati MMM.

Context

Mapping Manuscript Migrations (MMM) is a project funded between 2017 and 2020 by the Digging into Data Challenge of the Trans-Atlantic Platform research funding consortium. The main goal of the project is to combine data from several disparate sources about medieval and Renaissance manuscripts, and to use the aggregated data to explore a range of research questions about manuscript history and provenance. The project took data from three existing databases (the Schoenberg Database of Manuscripts[1], Medieval Manuscripts in Oxford Libraries[2], and Bibale[3]) and turned them into Linked Open Data (LOD)[4]. This involved transforming them into RDF triples and mapping them to a newly developed unified data model, drawing on the CIDOC-CRM and $FRBR_{OO}$ ontologies. Vocabularies for the main classes of entity (manuscripts, actors, places, and works) were reconciled across the three data sources using a mixture of automatic and semi-automatic methods.

[1] https://sdbm.library.upenn.edu/.
[2] https://medieval.bodleian.ox.ac.uk/.
[3] http://bibale.irht.cnrs.fr/.
[4] Full details of the technical aspects of the MMM project can be found in the MMM *White Paper*, available here: <https://diggingintodata.org/file/1281/download?token=x59u8fFQ>.

The aggregated data (nearly 22.5 million RDF triples) have been made available in several different ways. A copy of the dataset has been published through the Zenodo repository[5]. The data are hosted on the Linked Data Finland platform, from which they can be queried through a SPARQL endpoint or inspected directly[6]. A semantic portal has also been implemented using the Sampo-UI framework, through which the 217,000 manuscripts and other entities can be searched and browsed, using a combination of filters and map-based visualizations[7]. Result sets from the portal can also be downloaded in the form of CSV files via the SPARQL query service Yasgui[8].

The project convened an initial focus group of Oxford University researchers in 2017 and asked them to identify desirable features and functionality for the kind of data discovery service envisaged. It also gathered a set of research questions for testing, data modelling, and evaluation purposes. A User Group was established within the project, consisting of librarians, curatorial staff, and manuscript researchers from the project, with the aim of guiding and testing the implementation and customization of the semantic portal, as the user interface to the aggregated data. An early version of the portal was presented at a workshop during the 2019 "Digital Humanities" conference in Utrecht, and useful feedback was obtained from the researchers, librarians, and DH specialists who attended.

Once the User Group had signed off on the portal in early 2020, a more formal evaluation was carried out. The three researchers who carried out this evaluation were employed by the Institut de recherche et d'histoire des textes (IRHT) for the final months of the project, but had not been involved in the design or implementation of the portal and therefore brought fresh eyes to the way in which it worked. Their report was then used to make additional modifications and improvements to the portal, and also served as the basis for developing a set of Frequently Asked Questions for the project's Web site[9].

Methodology

The evaluation team consisted of three early career researchers: two historians and one philologist. Their profile was generally similar to that of the intended users of the MMM portal: qualifications and research interests in the field of manuscript studies, with varying degrees of computer skills, but not specialists in manuscript curation or Digital Humanities. They did have some experience in adding records to the Bibale database, and consequently made some use of Bibale as a point of comparison. They were not given any specific training in the use of the MMM portal, and were asked to approach it as new users, drawing only on the "Help" information available on the site itself. They were asked to comment on what worked in the portal, what did not appear to work, and what was unclear about the design and the instructions.

Their starting-point was the set of 26 research questions previously assembled by the project team (see Table 1). Some of these

[5] https://zenodo.org/record/4019643.

[6] http://ldf.fi/mmm/sparql.

[7] Esko Ikkala – Eero Hyvönen – Heikki Rantala – Mikko Koho, *Sampo-UI: A Full Stack JavaScript Framework for Developing Semantic Portal User Interfaces*, submitted to «Semantic Web Journal», (2020), <http://www.semantic-web-journal.net/content/sampo-ui-full-stack-javascript-framework-developingsemantic-portal-user-interfaces>.

[8] https://yasgui.triply.cc/.

[9] http://blog.mappingmanuscriptmigrations.org/frequently-asked-questions/.

questions were very specific, reflecting the interests of MMM project team members or the Oxford focus group, while others were more generic. The source for this latter group was a list of "Requêtes intéressantes" produced by the French Biblissima project[10]. The evaluation team also searched the portal on the basis of their own professional and personal interests. They investigated topics that they were specifically familiar with, such as the collection of Claude Fauchet; the manuscripts of the "Tournoiment Antechrist" and "Le Roman de le Rose"; and the Montpellier area in the medieval period.

[A1] How many manuscripts from pre-1600 produced in European countries survive?
[A2] How many manuscripts survive that contain Spanish texts written in gothic rotunda were produced in Castile for an abbey or convent? Then show me those which were owned during the nineteenth century by English private collectors; Then show me those which are now owned by an institution in North America.
[A3] What French collectors purchased manuscripts since the end of the Wars of Religion (after 1598)? Where are their manuscripts now?
[B1] How many manuscripts containing texts by Ramon Llul were sold in the 19th century?
[B2] Where are Ramon Llul manuscripts today?
[B3] Who collects manuscripts with texts by Ramon Llul?
[B4] How many times do texts by Ramon Llul's appear with texts by Albertus Magnus in the same manuscript?
[C1] What was the most popular text by a medieval author in France in the 17th Century?
[C2] Did Sir Thomas Phillipps own a 13th-century Bible with historiated initials?
[F1] Combien de manuscrits enluminés se trouvent dans une collection particulière? (volumétrie)
[F2] Quelle est la dynamique dans l'évolution des acquisitions et des dons? (étude diachronique)
[F3] Qui sont les donateurs et les propriétaires d'une collection?
[F4] Faire des recherches par sujet, par technique, par langue, par artiste voire par pigments (plus d'encre d'or, argent et pourpre) dans une collection.
[F5] Particularités d'une collection (sujet, technique, lieu de production etc.)? Quelles en sont les lacunes? Quelles en sont les dominantes?
[F6] Vie d'une collection, vie d'un livre enluminé?
[F7] Quels manuscrits sont probablement perdus?
[F8] Quel manuscrit a été vendu et n'est pas identifié au sein d'une collection à l'heure actuelle? (catalogue de vente)
[G1] Quelles copies d'un texte sont enluminées?
[G2] Quelle position occupe une copie dans l'histoire de la transmission d'un texte? Y a-t-il des exemplaires uniques des oeuvres?
[G3] Histoire de la transmission des images? [Outside the scope of the MMM dataset]
[G4] Quelles sont les versions existantes d'une oeuvre? Qui a fait une traduction française d'un texte ancien? Quand?
[G5] Quelles sont les différentes publications existantes [manuscript copies] d'un texte? (date, lieu de production, personne(s) responsable(s) etc.)
[H1] How many manuscripts were produced in Northern Italy and/or Lombardy?
[H2] How many manuscripts were produced in the Low Countries?
[H3] How many manuscripts were produced in London in the 15th century?
[H4] How many manuscripts formerly owned by Sir Thomas Phillipps are in British libraries?
[H5] What is the average number of folios in a Book of Hours?
[H6] Which collectors bought manuscripts from Wilfrid Voynich? Where were they located? What do we know about the kind of manuscripts he sold, and their earlier histories?

Table 1. *Mapping Manuscript Migrations research questions*

[10] https://doc.biblissima.fr/ontologie-biblissima - méthodologie.

The evaluation team focused initially on the user experience for new users of the portal:
– Is the MMM project presented with sufficient clarity for new users?
– Can one easily find instructions to browse the portal?
– Is the interface intuitive? Is browsing easy? How simple is it to make queries?

They then went on to evaluate the ways in which queries can be framed and executed, together with the presentation and use of the results. This included examining issues arising from the cross-linking of disparate data, and looking for any noticeable errors or inaccuracies in the data. They also examined the effectiveness of the map-based visualizations produced by the Sampo-UI software.

Findings

The findings from this evaluation were, in general, very positive and enthusiastic. Browsing the portal is easy and does not require specific computer skills. Basic queries (e.g., finding a manuscript, a collection, a place, etc.) are very straightforward. Using the filters and the map visualizations is very intuitive, even enjoyable. But the findings of most interest to the project team concerned areas where improvements could be made, or where clearer explanations were required.

Some of these related to specific features of the Sampo-UI software. Moving the date slider with precision was felt to be difficult, for example. Instead of displaying all the many columns of information about each manuscript, by default, it would be helpful to be able to select which columns to display. Pie charts are provided as a way of visualizing the results of some filters (such as owners), but not for others (such as production dates). But their value might be limited when the percentages are too small to display. The percentage of French-language manuscripts in the Yale University Library is labelled as 6.6% in the appropriate pie chart, for instance. But the percentages for the least-represented languages in the Yale collections are too small and cannot be displayed.

Selecting and combining different elements within a filter was thought to lack certain options. Filtering by "Owners" to find manuscripts owned by either Thomas Phillipps or Thomas Thorpe, for example, was possible, but not manuscripts owned by both Phillipps and Thorpe. Some combinations for filtering by the hierarchical lists of places were not possible, especially with overlapping regions or terms like Languedoc, Occitanie, and Southern France. The other approach to filtering – drawing and using a "Bounding Box" on the map – was not necessarily an effective alternative.

In the tabular display of information relating to manuscripts, there are separate columns for "Owner" and "Transfer of Custody". This type of presentation was felt to make it difficult to connect these provenance events with the relevant owners of a manuscript. Because the connections between columns are not displayed – except inasmuch as they relate to the same manuscript – the only way to see these relationships is to click on a specific element, such as a "Transfer of Custody" event. This will display all the available information about that event, including the persons and organizations involved, in the "Custody surrendered by" and "Custody received by" fields.
The export tool based on the underlying SPARQL queries was found to be relatively simple to use, but the resulting CSV file was less easy to understand. A manuscript like "Montpellier (F), BU Historique de Médecine, H 069", for example, produces a spreadsheet with six lines rather than one, together with a total of 84 columns. Because there is so much information about each manuscript, often

with multiple values for the same element (e.g., production date), it would be difficult and possibly misleading to combine all the information on to a single row. Once the CSV file has been exported and downloaded, it is possible to combine multiple rows using software like OpenRefine or Google Sheets.

A number of the findings appeared to relate to the Sampo-UI software but were actually connected to the nature and structure of the source data, and the ways in which the data had been mapped and harmonized. When the manuscripts are filtered by production place and the results are presented in a table, the total number of results may differ. Filtering for "Europe" as "Production Place" shows a total of 86,332 in the filter, but 73,665 in the table of results. This is because the first is the number of places and the second is the number of manuscripts. A manuscript may have multiple production places assigned to it, either because of disagreements between the data sources or because of uncertainty about its origins: Southern France or Northern Italy?

Manuscripts belonging to a specific person or organization can be filtered by "Collections" as well as by "Owners", but these produce different results. Claude Fauchet as an "Owner" owns nine manuscripts but his "Collection" contains only one manuscript; Vossius owns seventy manuscripts, though his collection only shows five of them. This is because the "Collection" information is derived exclusively from the Bibale database; the other data sources do not include the separate concept of a collection in their data model for provenance histories. "Owner" information, on the other hand, may come from any of the three data sources and will invariably have the fullest picture of a specific person's manuscript holdings.

The different stages in the migration of a manuscript are not shown on the map visualizations. They only show the place of production and the last-known location, together with the arc joining these two places. This is because the information in the source datasets is simply not full enough or sufficiently well-structured to arrange the intervening places in the history of a manuscript's ownership over the centuries in chronological order. The available information is displayed in the events and owners columns in the tabular presentation of data for each manuscript, however, and can also be found through more complex SPARQL queries. The use of last-known location rather than current location was also queried. Although the current location is obvious for manuscripts from the Oxford catalogue, Bibale does not necessarily record a current location, while the Schoenberg Database focuses on last-known locations associated with or recorded in catalogue entries.

The evaluation noted that some authors had multiple entries, while other authors had only one entry. This is because many authors appear in different forms in the data sources. The MMM project tried to harmonize as many of these as possible, but there are still some authors who have not been harmonized. As a result, they appear as two or more different names in the portal. Searching the "Author" filter and selecting the different forms there will produce a combined set of results. Authors may be linked to many works, even though they are the author of only one of these works. This arises from the structure of some entries in the Schoenberg Database, where multiple works and multiple authors in a single manuscript are not individually linked. In these cases, the author is labelled "possible author".

The evaluation also noted some repetitions and inconsistencies in some elements of the manuscript descriptions. This is usually because the MMM project has combined data

about the same feature from two or three different sources. For production dates and places, in particular, it is quite often the case that (for example) the Schoenberg Database has multiple differing values for the same manuscript from different sales catalogues. The MMM project chose to display all these variations, together with their source, rather than attempting to merge them or identify a single "most-accurate" value.

More generally, some of the findings related to the overall scope and purpose of the MMM project and its data. While the researchers found the portal's design to be attractive and inviting for new users, they also commented that it was hard to know, at a glance, what the aims of the portal were, what data were incorporated in it, and what its limits were. They were unsure to what extent the data had been harmonized, how complete or incomplete the manuscript data were, and what information had not been mapped from the source datasets.

These issues were demonstrated through specific research inquiries devised during the evaluation. Manuscripts with texts relating to the Carolingian Duke Guillaume d'Orange cannot easily be found in a single search of works, partly because the MMM data are not oriented towards mapping textual traditions specifically and partly because it proved difficult to reconcile the titles of works across the data sources. There are relatively few results with manuscripts produced in the south of France during the Middle Ages, simply because this reflects the coverage in the data sources.

The evaluation report suggested the provision of fuller information to guide users in their expectations of the portal and the data. This might include explaining the nature of the data, and the possibilities and limitations arising from such factors as the nature of the data in the various constituent databases, the processes for harmonization and reconciliation of the data, and the way in which the portal was built, including the data model.

Outcomes

The evaluation report by the three IRHT researchers presented specific recommendations and suggestions for addressing the issues identified. The most important of these concerned the need to provide more explanatory and contextual information about the portal. As a result, the "Info" section of the portal has been expanded with links to the project's technical documentation on GitHub[11] and to the project's Web site and blog[12], where a set of Frequently Asked Questions has been added. The FAQ covers general questions about the scope and purpose of MMM and the nature of its data, as well as more specific questions about the use and functionality of the MMM portal[13]. The GitHub documentation includes a SPARQL tutorial based on the MMM data and research questions[14].

Several recommendations suggested specific improvements to the functionality of the portal. Where feasible, these were added to the workplan for the Sampo-UI software, and most have subsequently been implemented. They included the following items:

– Combining values from within one filter: an enhancement will give users the choice of using "AND" within a filter (to find manuscripts owned by both Phillipps and Beatty, for example) instead of the default "OR"

[11] https://mapping-manuscript-migrations.github.io/.
[12] http://blog.mappingmanuscriptmigrations.org/.
[13] http://blog.mappingmanuscriptmigrations.org/frequently-asked-questions/.
[14] https://mapping-manuscript-migrations.github.io/sparql/sparql_tutorial.html.

combination (manuscripts owned by either Phillipps or Beatty);
– Reducing the number of columns which display in the results table: an enhancement will allow users to select columns for display, as an alternative to the default display of all columns;
– Extending the display of percentages in pie charts to all the values;
– Replacing the date slider with "from" and "to" boxes, for selecting date ranges for manuscript dates of production;
– Providing an option to bypass the automatic selection of all subsidiary places when filtering manuscripts by place of production.

Other issues raised in the evaluation report were related to the nature of manuscript provenance data more generally, and the ways in which such data are recorded and structured. These arose partly because of the inherent incompleteness and uncertainty of the data, and partly because of limitations and assumptions in the data models of the source datasets, which made some elements of them difficult to map with sufficient granularity or specificity. Some compromises had to be made in the MMM data modelling and mapping processes in order to find common semantic ground across three very different sets of complex data.

The project has produced a White Paper which surveys these issues in more detail, and has published its data model and its data for other projects or users to work with and improve on[15]. It is also developing specific recommendations for enhancing the TEI encoding of provenance in manuscript catalogues, with the aim of improving the mapping process to event-based data models like that employed by MMM.

Conclusion

The overall conclusion of the evaluation report was that the MMM portal is an excellent tool, and very easy to use. Nevertheless, the report also showed that it is important to understand and acknowledge the scope and parameters of the MMM data underlying the portal, and the inherent limitations which result from them. The source data have not been corrected or amended, and they have been harmonized rather than merged. Some elements, especially those unique to one of the sources, have not been mapped. But links to the original records in the data sources have been provided, so that the full information about any given manuscript in that source can be easily found.
The portal is a very rich tool for exploring the history of books, of cultural exchanges, of libraries, and of collections. This tool also makes it possible to follow the history of a given manuscript, and to conceive of the manuscript as a source (text additions, transformation, reason for its preservation, and so on). It is easy to create corpora organized by collection or owner, but more difficult with regard to geographical areas or historical periods. It is also possible to create corpora according to production places (when this information exists), and to study the vitality of specific scriptoria.

[15] https://diggingintodata.org/file/1281/download?token=x59u8fFQ.

This paper reports on an evaluation of the semantic portal Mapping Manuscript Migrations (MMM), which combines data from several different sources about medieval and Renaissance manuscripts, in a Linked Open Data framework. The evaluation team looked initially at the user experience for new users of the portal:

– *Is the MMM project presented with sufficient clarity for new users?*
– *Can one easily find instructions to browse the portal?*
– *Is the interface intuitive? Is browsing easy? How simple is it to make queries?*

They then went on to evaluate the ways in which queries could be framed and executed, together with the presentation and use of the results. This included examining issues arising from the cross-linking of disparate data, and looking for any noticeable errors or inaccuracies in the data. They also examined the effectiveness of the map-based visualizations produced by the Sampo-UI software. The recommendations arising from the evaluation were used as the basis for improving the Sampo-UI functionality, adding to the online help given to users of the portal, and increasing the amount and type of information available to advanced users, especially in relation to the scope and coverage of the MMM data.

L'ultima consultazione dei siti web è avvenuta nel mese di dicembre 2020