Linked Open Data Vocabularies and Identifiers for Medieval Studies

Toby Burrows[1] [0000-0002-0469-7584], Antoine Brix[2] [0000-0001-6532-7416], Douglas Emery[3] [0000-0002-5147-7736], Arthur Mitchell Fraas[3] [0000-0003-3228-9876], Eero Hyvönen[4] [0000-0003-1695-5840], Esko Ikkala[4] [0000-0002-9571-7260] Mikko Koho[4] [0000-0002-7373-9338] David Lewis[1] [0000-0003-4151-0499], Synnøve Myking[2], Kevin Page[1] [0000-0002-1668-6540], Lynn Ransom[3] [0000-0002-5231-3602], Emma Cawlfield Thomson[3] [0000-0002-5896-7922], Jouni Tuominen[4] [0000-0003-4789-5676], Hanno Wijsman[2] and Pip Willcox[5] [0000-0003-4060-3255]

[1] University of Oxford, 7 Keble Road, Oxford OX1 3QG, United Kingdom
{toby.burrows,david.lewis,kevin.page}@oerc.ox.ac.uk
[2] Institut de la recherche et d'histoire des textes, Campus Condorcet, 14 cours des Humanités, 93322 Aubervilliers cedex, France
{antoine.brix,synnove.myking,hanno.wijsman}@irht.cnrs.fr
[3] University of Pennsylvania, 3420 Walnut Street, Philadelphia PA 19104, United States
{emeryr,fraas,lransom,emmacaw}@upenn.edu
[4] Aalto University, Konemiehentie 2, 02150 Espoo, Finland
{eero.hyvonen,esko.ikkala,mikko.koho,jouni.touminen}@aalto.fi
[5] The National Archives, Kew TW9 4DU, United Kingdom
pip.willcox@nationalarchives.gov.uk

Abstract. This paper examines the use of Linked Open Data in the research field of medieval studies. We report on a survey of common identifiers and vocabularies used across digitized medieval resources, with a focus on three internationally significant collections in the field. This survey has been undertaken within the "Mapping Manuscript Migrations" (MMM) project since 2017, aimed at aggregating and linking disparate datasets relating to the history of medieval manuscripts. This has included reconciliation and matching of data for five main classes of entities: Persons, Places, Organizations, Works, and Manuscripts. For each of these classes, we review the identifiers used in MMM's source datasets, and note the way in which they tend to rely on generic vocabularies rather than specialist medieval ones. As well as discussing some of the major issues and difficulties involved in conceptualizing each of these types of entity in a medieval context, we suggest some possible directions for building a more specialized Linked Open Data environment for medieval studies in the future.

Keywords: Linked Open Data, Medieval Studies, Data Reconciliation.

1 Context

Medieval studies is a large field of research, covering the entire history and culture of Western Europe for more than a thousand years, from the fifth to the fifteenth centuries.

One of its major conferences, the annual International Medieval Congress in Leeds, contains more than 750 sessions with over 2,500 participants. A wide range of academic disciplines intersect in their interest in medieval Europe, including literary studies, linguistics, history, music, art, philosophy, theology, archaeology, and so on. There is a vast literature of printed and electronic publications, and an increasingly complex environment of digital resources of many different kinds, including digitized manuscripts, digital editions, and specialized databases.

In that digital environment, however, there is nothing to match the recent initiatives in classics and ancient history to develop tools and resources for data linkage. Pelagios is the focus for several of these, including gazetteers, annotation tools, and the Distributed Text Services specification. Perseus is another important node, providing a comprehensive corpus of Open Greek and Latin texts together with Canonical Text Services and collaborative editing tools. In early modern studies, there are similar developments, with Early Modern Letters Online launching its Early Modern Places, People, and Dates vocabularies. At the heart of these initiatives is Linked Open Data – both as a paradigm and as a set of specifications, tools, and methodologies, loosely coordinated by the Linked Pasts Network. [1]

In this paper, we survey the current situation for data in medieval studies and present proposals for future directions in the application of Linked Open Data to this research field. Our proposals build on work being carried out as part of the "Mapping Manuscript Migrations" (MMM) project, an international collaboration between the University of Oxford, the University of Pennsylvania, Aalto University, and the Institut de recherche et d'histoire des textes (IRHT). [2] Funded by the Trans-Atlantic Platform under its Digging into Data Challenge for 2017-2020, the MMM project focuses on aggregating data from three major data sources: the Schoenberg Database of Manuscripts, Medieval Manuscripts in Oxford Libraries, and Bibale. Two of these sources are bespoke relational databases, while the Oxford catalogue consists of Text Encoding Initiative (TEI) XML documents. The methodologies being deployed include a unified data model derived from the CIDOC-CRM and $FRBR_{OO}$ ontologies [3], and a matching and reconciliation process based on Linked Open Data identifiers and vocabularies. Our method is to create a single, repeatable data pipeline, which first converts the source datasets into the unified data model, and after that reconciles and merges the core entities (if possible) using generic Linked Open Data vocabularies. In this way, updates in the source datasets can be handled by running the whole pipeline again.

These identifiers and vocabularies are the best available form of data linkage for humanities research fields. A variety of tools, services, and interfaces can be built on top of a Linked Open Data environment which covers the main types of entity classes present in data models. For the MMM project, these classes are Manuscripts, Persons, Organizations, Places, and Works. In the field of medieval studies as a whole, the Manuscripts class would need to be expanded to cover Objects more broadly. While manuscripts are the most common surviving type of object from the medieval period and the most valuable evidence from a range of interdisciplinary research perspectives, they are not the only surviving form of evidence. They need to be supplemented by other types of object – notably artworks of various kinds, buildings, coins, ceramics, textiles, and so on. For the purposes of this paper, we will be focusing on manuscripts. As well

as being the main focus of medieval research, they exemplify most of the issues connected with the development of identifiers and vocabularies for medieval objects.

2 Places

The majority of places in the source datasets for the MMM project can be matched to identifiers from one or more of the most widely used generic Linked Open Data vocabularies: the Getty Thesaurus of Geographical Names (TGN) and GeoNames. TGN is much smaller than GeoNames but more historically oriented and actively curated. Neither has anything like the coverage of the gazetteers produced by mapping agencies. For Great Britain, for example, TGN has only half as many places names as GeoNames, but this in its turn has only a fraction of the names provided by the Ordnance Survey. [4]

The MMM datasets, in which medieval, early modern, and modern place names are all present, have their preferred default vocabulary: GeoNames for Bibale, and TGN for Schoenberg and Oxford. The case of the Oxford dataset is particularly interesting; about 92% of its place names have been given TGN identifiers, with a further 2% in GeoNames. The remaining names – about 6% of the total – have not been found in either of these sources. Many of them have been given identifiers from a range of other sources, not all of which are Linked Open Data vocabularies; they include the Historical Gazetteer of England, Pleiades, Trismegistos, VIAF, Vision of Britain, and Gatehouse-Gazetteer.

None of these datasets makes use of a gazetteer specifically designed to cover medieval places and their medieval names. This is partly, at least, because much of the evidence in the MMM source datasets relates to the history of manuscripts in the post-medieval world. The places referred to are usually (though not always) the same as places which exist today, and they are usually presented in modern languages, rather than in Latin. For medieval places in England, the best source of data is the *Survey of English Place Names*, published by the English Place Name Society and digitized as part of the *Historical Gazetteer of England's Place-Names*. This is not yet available as Linked Data, though its creators have claimed that it would be "relatively easy to convert it to RDF". [5] The *Historical Gazetteer* itself was archived by the King's Digital Lab in 2018 and its data are only available on request, though some static pages can be found in the Internet Archive. [6]

3 Persons

In the source datasets for the MMM project, "person" entities can cover everything from medieval authors and early modern owners to modern buyers and sellers of manuscripts. The three datasets are consistent in using VIAF (Virtual International Authority File) identifiers as their primary source for persons. In the Schoenberg Database of Manuscripts, VIAF is the preferred source. In the Bodleian catalogue, VIAF identifiers are supplemented by other sources, including the Library of Congress, Wikidata, Sudoc, Gemeinsame Normadatei (GND), and the Oxford Dictionary of National

Biography. Bibale also prefers VIAF, but adds identifiers from Wikidata and the ARK identifiers from the Bibliothèque nationale de France; Biblissima identifiers are also referenced.

Personal names in the Schoenberg Database cover a wide range of roles and activities, including artist, scribe, selling agent, buyer, and seller, as well as author. Since VIAF is ultimately drawn from library bibliographical catalogues, its coverage of people other than published authors is limited. Despite its focus on the authority files of national libraries, VIAF does include some other name sources, such as Perseus, Wikidata, and the Getty's Union List of Artist Names (ULAN). Although the Schoenberg Database has offered to contribute a significant number of non-VIAF names to augment the VIAF vocabulary, VIAF has a defined set of "Admission Criteria", which prefers national and international programmes and seems to exclude this kind of one-off, smaller-scale contribution.

None of the MMM datasets makes reference to any specifically medieval name authorities. There are many lists and catalogues of medieval persons and their names; Wright's bibliography lists more than fifty, most of which are available in digital form. [7] It is worth noting that several of the digital sources are commercial products, distributed by publishers like Brepols. These include Europa Sacra (with data on 30,000 church prelates) and the author indexes to the Library of Latin Texts. Given the proliferation of this kind of digital resource, an initial assessment to determine which might most fruitfully be transformed into Linked Open Data ought to be carried out. Some specific initiatives are already under way, including the work by the "Medieval Publishing" project at the University of Helsinki to transform Sharpe's *Handlist of the Latin Writers of Great Britain and Ireland before 1540* into database form. [8] Publishing this dataset as Linked Open Data is envisaged as a second stage for that project.

4 Organizations

Organizations are not always distinguished from places in the source datasets for the MMM project. The interface to the Oxford manuscripts catalogue, for instance, simply includes organizations in its browsable list of "places". This is presumably because the majority of these organizations are medieval religious houses of various kinds, which are usually known by a place name, such as "Abingdon, Benedictine abbey" or "Admont, Austria, Benedictine abbey". A minority are recorded as named events, which were held in specific places, such as "Agde, Council of (506)". Behind the scenes, however, they are encoded as organization names in the TEI files. They are also given identifiers from VIAF rather than TGN. Other identifiers are also given, including Wikidata, the Library of Congress, Sudoc, and ARK identifiers from the Bibliothèque nationale de France.

In the Schoenberg Database of Manuscripts, on the other hand, organizations are combined with persons into a "Names" file. These are both medieval and modern, and include a significant number of religious houses. A large proportion of the organizational names have been given VIAF identifiers.

Bibale's organizations consist mostly of libraries and other institutional owners of manuscripts (especially medieval and early modern religious houses). VIAF is again consistently used, accompanied by identifiers from the Bibliothèque nationale de France, Biblissima, and Wikidata.

Organizations as a class raise various questions of scope and definition in a medieval context. As the conflation of religious houses with places suggests, it is difficult to separate what we might call a corporate body from its medieval location. In medieval England, for instance, Anglo-Saxon administrative units like "hundreds" could be considered primarily as places, since they were defined as geographical areas within a county or shire. But the Norman system of government overlaid on top of this framework after 1066 was less conceptually clear; the "honours" and "baronies" of the great lords consisted of various estates scattered across the country, which did not correspond to a specific geographical area. They were primarily organizational and administrative in nature. [9] A specialist Linked Open Data vocabulary for medieval organizations in England could be drawn from existing lists of baronies and religious houses, together with data relating to offices within the structure of royal government and to the administrative hierarchy of the Church.

5 Works

The Works class is more problematic than persons and places. The source datasets for the MMM project contain very little in the way of links to external Linked Open Data identifiers for works, and some do not even distinguish "works" – or even titles – as a separate element in their data. The Schoenberg Database, for example, does not contain separate records for works or titles; they are simply recorded as part of the observation record.

In the Oxford manuscript catalogue, on the other hand, the Bodleian Library provides a unique ID number for more than 12,000 works. It records the "uniform title" (author + work), the "manuscript title", and the "normalised manuscript title". There are more than 360 links to external identifiers – primarily Pinakes for Greek works and Mirabile for Latin works, as well as a small number of instances for the German Handschriftencensus and the French ARLIMA.

Bibale distinguishes between "Oeuvre" (i.e., work) and the "Text/Edition" (i.e., manifestation). There are unique identifiers for both of these, but the only links to external identifiers come from the IRHT's Fama database, which provides identifiers for more than 1,600 works. These were loaded into Bibale during 2019.

There are hundreds of lists and catalogues of medieval religious and literary writings, and many of these are available in some kind of digital form. [7] Among the issues to be tackled here are the Bible and its derivatives (the focus of much medieval writing, especially in the earlier Middle Ages); the multiplicity of anonymous works – often without any kind of title and often identifiable only by their opening words (their "incipit"); and the way in which many medieval "works" actually consist of annotations, commentaries, and glosses on earlier works, as well as re-workings and abridgements of them. The work of the "Medieval Publishing" project is relevant here too, since

Sharpe's *Handlist* identifies and provides titles for the writings of medieval British and Irish authors.

6 Manuscripts

Each of the MMM source datasets attempts to go beyond simply using the shelf-mark and repository for identifying individual manuscripts. Bibale includes identifiers from another IRHT database, Medium, which lists more than 109,000 manuscripts. Each of these has its own ARK identifier and is also linked, where appropriate, to the other main IRHT databases, such as Pinakes and Jonas. Links to Pinakes and Jonas are also included in some of the Oxford manuscript descriptions; each manuscript has a unique ID number, but there is no authority list for manuscripts (unlike persons, places, organizations, and works). The creation of ARK identifiers for Oxford manuscripts is currently being investigated by the MMM project and the Bodleian Library. [10] In the Schoenberg Database, each of the nearly 25,000 manuscript records has a unique URL, but they are not accompanied by links to any external identifiers or to an authoritative shelf-mark or title.

The need for a system of Linked Open Data identifiers for medieval manuscripts is now broadly accepted, and has been the subject of several meetings over the last two years. Under the auspices of the IRHT, a project has been established to develop and test a specification for an International Standard Manuscript Identifier or ISMI. [11] Building on work already done in the Pinakes database of Greek manuscripts, the ISMI project envisages an ISO standard implemented as a register of Western European manuscripts. The Medium database is a possible starting-point for this implementation.

There is still considerable work to be done in defining the scope and purpose of ISMI, as well as in designing and building the technical solution. Even when the ISMI registry is operational, it will still take a major effort to create and embed these identifiers in local databases for manuscripts with known and established current institutional locations and shelf-marks. Matching other occurrences of a manuscript (e.g., in medieval library catalogues or modern sales catalogues) to its ISMI identifier will be an even greater task – but it has the potential to revolutionize medieval manuscript studies through large-scale inter-linking of heterogeneous manuscript data.

All this assumes that ISMI will focus initially on those manuscripts which have a modern location and shelf-mark. As Cassin notes [11], there are a range of questions around the margins of this focus. Most obviously, what about medieval manuscripts which have been scattered into multiple modern locations? An extreme example of this is the result of the work of Otto Ege in the 1940s and 1950s in breaking up manuscripts into single leaves and selling them to multiple owners. [12] Composite modern manuscripts pose the reverse kind of problem. And what about manuscripts which were known to exist in the nineteenth and twentieth centuries but can no longer be connected with a specific current owner or location? What about modern collections of disparate cuttings or single leaves?

At the root of these questions, ultimately, is the purpose of an identifier like ISMI. Does it limit itself to a manuscript as a single discrete physical object which exists today

in a known location, or should it also try and identify "attested" manuscripts – ones which existed in the past but are no longer extant or locatable today?

7 Conclusion

In this paper we have summarised our survey of identifiers and vocabularies used in three internationally significant digital repositories of medieval data, examined the implications and context for other medieval datasets, and proposed how these might be extended and combined for the benefit of manuscript studies. In the MMM project, this survey has formed the basis for an ongoing technical integration of the three datasets into a unified web of Linked Open Data, which will be reported in future work. The project plans to publish its own LOD vocabularies later in 2020. They will include identifiers for 215,000 manuscripts, 424,000 works, 51,000 persons and organizations, and 4,000 places.

We believe that it is critical to develop, publish and re-use Linked Open Data vocabularies in order to further the interoperability of digital resources for medieval studies. The more generalist vocabularies like VIAF and TGN already have a reasonably good coverage of persons, places, and organizations relevant to this field, as the MMM project has been able to demonstrate. But the level of that coverage has been inflated to some extent by the scope of the datasets relevant to manuscript research, since they include many early modern and modern names as well as medieval ones.

We suggest the next priority for medieval studies more broadly should be the conversion of a selection of existing specialist vocabularies to Linked Open Data formats, accompanied by an investigation of ways in which they can be associated with the more general vocabularies. The work being done by the "Medieval Publishing" project provides a possible model here. While persons, places, and organizations might be considered the "low-hanging fruit" in this context, it will be more important and valuable in the longer term to tackle the more complex issues around works and manuscripts – the latter within the framework envisaged by the ISMI initiative.

References

1. Linked Pasts, http://linkedpasts.org, last accessed 2020/01/16.
2. Hyvönen, E., Ikkala, E., Tuominen, J. et al.: A Linked Open Data service and portal for pre-modern manuscript research. In: DHN 2019 Conference (2019), http://ceur-ws.org/Vol-2364/20_paper.pdf, last accessed 2020/01/16.
3. Le Boeuf, P.: Modeling rare and unique documents: using $FRBR_{OO}$/CIDOC CRM, Journal of Archival Organization 10(2), pp. 96–102 (2012).
4. De Sabbata, S., Acheson, E.: Geographies of gazetteers in Great Britain. In: 24th GIS Research UK (GISRUK 2016) Conference (2016), http://hdl.handle.net/2381/38182, last accessed 2020/01/16.
5. Grover, C., Tobin, R.: A Gazetteer and Georeferencing for English Historical Documents. In: Proceedings of the 8th Workshop on Language Technology for Cultural

Heritage, Social Sciences, and Humanities (LaTeCH) @ EACL 2014, pp. 119–127 (2014), http://www.aclweb.org/anthology/W14-0617, last accessed 2020/01/16.
6. http://deep.kdl.kcl.ac.uk, last accessed 2020/01/16.
7. Wright, C.: Medieval History and Historical Sources, https://bibliography.arc-humanities.org/medieval-history-and-historical-sources/, last accessed 2020/01/16.
8. Medieval Publishing from c. 1000 to 1500, https://www.helsinki.fi/en/researchgroups/medieval-publishing, last accessed 2020/01/16.
9. Sanders, I.: English Baronies: a Study of their Origin and Descent, 1086−1327. Oxford University Press, Oxford (1960).
10. Burns, H., Burrows, T., Downie, J., Lewis, D., Page, K., Velios, A.: "Assessing the practicality of ARK identifier usage in a catalogue of medieval manuscripts", poster, iConference 2019, Washington, DC, http://hdl.handle.net/2142/103380, last accessed 2020/01/16.
11. Cassin, M.: ISMI: International Standard Manuscript Identifier: Project of Unique and Stable Identifiers for Manuscripts (2018), https://www.manuscript-cultures.uni-hamburg.de/files/mss_cataloguing_2018/Cassin_pres.pdf, last accessed 2020/01/16.
12. Gwara, S.: Otto Ege's Manuscripts: a Study of Ege's Manuscripts, Portfolios, and Retail Trade, with a Comprehensive Handlist of Manuscripts Collected or Sold. De Brailes Publishing, Cayce, SC (2013).