Transforming the Digital Landscape: Towards a Medieval Knowledge Graph

Dr Toby Burrows

University of Oxford; University of Western Australia

toby.burrows@uwa.edu.au

Transforming the Digital Landscape: Towards a Medieval Knowledge Graph

Digitization and digital methods are now pervasive in medieval studies. The Medieval Academy of America has acknowledged these developments in various ways. In 2017, *Speculum* published a special Open Access issue showcasing examples of research based on digital methodologies,[1] and the Academy has supported a range of specific initiatives, including the Medieval Digital Resources site, the annual Digital Humanities and Multimedia Studies Prize, the Digital Latin Library, and the Library of Digital Latin Texts.

Numerous claims have been made for the transformative effect of all this digital activity, often using the rhetoric of "democritization" of access.[2] My aim here is to consider the reality behind this rhetoric, especially in relation to Western manuscripts and the texts they carry, and to speculate on what kinds of digital initiatives might be truly transformative for medieval research in the future. The focus of such developments, I argue, should be on Linked Open Data (LOD) and shareable unique identifiers, with support for collaborative content creation and analysis.[3] A Medieval Knowledge Graph constructed on this basis would be a good example of the kind of "Smart Data," which scholars like Marcia Lei Zeng have identified as bringing "extraordinary value" to the humanities.[4]

Access to medieval manuscripts is undoubtedly much easier when high-quality digital colour facsimiles are freely available over the Web, instead of scholars having to find the funds to travel considerable distances and be authenticated as a suitable person to consult manuscripts face-to-face. While A. S. G. Edwards has pointed out some ways in which in-person access is still important, Andrew Prescott has deftly analysed some of the myths attached to consulting manuscripts in person, and the mystique attached to individual "celebrity manuscripts."[5] This "celebrity" approach persists in the digital world, as Michelle Warren has pointed out. Noting

[1] David J. Birnbaum, Sheila Bonde, Mike Kestemont (eds), "The Digital Middle Ages: a *Speculum* Supplement," *Speculum* 92/S1 (October 2017). https://doi.org/10.1086/695389

[2] Irene van Renswoude, "After Digitization, What More?: The Touch and Feel of Virtual Manuscripts," *Quaerendo* 54/2-3 (2024): 210-225 (at 211-212). https://doi.org/10.1163/15700690-bja10026

[3] For an explanation of this approach, see: Eero Hyvönen, "How to Create and Use a National Cross-Domain Ontology and Data Infrastructure on the Semantic Web," *Semantic Web – Interoperability, Usability, Applicability* 15/4 (2024):1499-1513. https://doi.org/10.3233/SW-243468

[4] Marcia Lei Zeng, "Smart Data for Digital Humanities," *JDIS: Journal of Data and Information Science* 2/1 (2017): 1-12.

[5] A. S. G. Edwards, "The Digital Archive, Scholarly Enquiry, and the Study of Medieval English Manuscripts," *Archive Journal* Sept. 2018. https://www.archivejournal.net/essays/digital-archive-scholarly-enquiry-and-the-study-of-medieval-english-manuscripts/ ; Andrew Prescott, "Ways of Seeing Manuscripts: Exploring Parker 2.0," in *Medieval Manuscripts in the Digital Age*, ed. Benjamin Albritton, Georgia Henley, and Elaine Treharne (Oxford, 2020), 37- 54 (at 39-40).

that “famous literary authors” and “dazzling documents” have been relied on to attract funding for manuscript digitization, she concludes that “digitization has left the literary canon largely intact – along with the nationalist values that built the canon.”[6] Tjamke Snijders makes a similar point, identifying three main emphases in digitization projects: “richly illuminated manuscripts, ‘iconic’ manuscripts (often the oldest ones in the collection), and those containing texts that can be attributed to well-known authors.”[7]

In reality, digitization has only been applied to a small proportion of the surviving medieval manuscripts and documents, and these have tended to be celebrity manuscripts and collections, as well as focusing on thematic areas which are of interest to specific funders. Precise statistics are hard to come by, but it appears that Oxford University’s Bodleian Libraries have fully digitized only about 500 of their 10,000 medieval manuscripts, while the Vatican Library has digitized about 29,000 of its 80,o00 mainly medieval and Renaissance manuscripts, the Bibliothèque nationale de France nearly 23,000 medieval manuscripts, and the British Library around 4,800 “Ancient, Medieval, and Early Modern” manuscripts (as of January 2022).[8] The overall proportion of surviving medieval manuscripts which have been digitized might amount to no more than 5-10% overall – and even lower if medieval documents are added to the calculation.

The growing corpora of digital images of manuscripts have inspired some innovatiuve large-scale research. The images can be compared and analysed for the letter-forms they exhibit, as Peter Stokes and others have shown in their work on DigiPal.[9] Analysing page layouts using digital images of manuscripts has been carried out by the HORAE project across five hundred books of hours.[10] Digital images also serve as the basis for the automatic transcription of texts using tools like eScriptorium and Transkribus.[11] Much of this work is underpinned by the International Image Interoperability Framework (IIIF), which enables

---

[6] Michelle Warren, *Holy Digital Grail: a Medieval Book on the Internet* (Stanford, 2022), 6-8.
[7] Unpublished paper quoted in van Renswoude, “After Digitization,” 217.
[8] Calum Cockburn, “Our Digitised Collection Keeps on Growing,” British Library, *Medieval Manuscripts Blog*, Jan. 5, 2022. https://blogs.bl.uk/digitisedmanuscripts/2022/01/our-digitised-collection-keeps-on-growing.html Other figures come from the relevant library Web sites.
[9] Peter A. Stokes, “The Problem of Digital Dating: a Model for Uncertainty in Medieval Documents,” *Digital Humanities 2015*, Sydney, Australia. http://peterstokes.org/pubs/Stokes_digital_dating.pdf
[10] Mélodie Boillet, Marie-Laurence Bonhomme, Dominique Stutzmann, and Christopher Kermorvant, “HORAE: an Annotated Dataset of Books of Hours,” in *Proceedings of The 5th International Workshop on Historical Document Imaging and Processing (HIP ’19)* (New York, 2019). 6 pages. https://doi.org/10.1145/3352631.3352633
[11] P. A. Stokes, B. Kiessling, D. Stökl Ben Ezra, R. Tissot, and H. Gargem, “The eScriptorium VRE for Manuscript Cultures,” in: *Ancient Manuscripts and Virtual Research Environments*, ed. Claire Clivaz and Garrick V. Allen. Special issue of *Classics@* 18 (2021). https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/

digital images from different manuscripts to be displayed together on a single screen.[12]

Applying computational tools to the *semantic content* of these digitized manuscripts (as well as to undigitized ones) is a different matter. This requires access to descriptions of the manuscripts, as well as to the texts carried by the manuscripts. These descriptive metadata can be found in library and museum catalogues, together with notes on the history and provenance of manuscripts. At present, there are numerous databases and (sometimes digitized) printed books which have to be searched separately, even just for Western European manuscript collections. The national catalogues in most European countries do at least aggregate substantial proportions of the descriptive data. But in some cases, notably the United Kingdom, there is no national catalogue of manuscripts at all, and the available databases are all at the institutional level.[13] The Manuscripts Portal of the Consortium of European Research Libraries (CERL) was discontinued in early 2020.

Interoperability between these sources has been urged for some years, with relatively little success in practice, as Bridget Whearty demonstrates in her account of the failure of the Digital Manuscripts Index project.[14] While some notable recent efforts have been made to overcome these limitations, the results have been patchy. The French *Biblissima* service aggregates information about nearly 235,000 manuscripts (almost 40% of which have been digitized), which can be browsed, searched, and filtered in various ways, including map visualizations and image-based searches.[15] The “authority files” data are also accessible through APIs (Application Programming Interfaces).[16]

Europeana – the large-scale aggregator of European cultural heritage images – contains over 94,000 images related to the Middle Ages, as well as more than 500,000 images of manuscripts, but the descriptions are minimal and the ability to filter by period, form, and topic is limited. Support for analysis and visualization is limited to the provision of APIs to extract data; these include an endpoint for Linked Open Data (LOD) and knowledge graph queries, but this relies on a 2017 export of the *Europeana* data.[17]

---

[12] International Image Interoperability Framework, “How It Works: a Plain-Language Guide to How the APIs Work.” https://iiif.io/get-started/how-iiif-works/ Accessed March 9, 2025.
[13] Neil Ker’s *Medieval Manuscripts in British Libraries* (5 vols., Oxford, 1969-2002) deliberately omits the collections of the larger libraries, on the grounds that these have their own detailed catalogues.
[14] Bridget Whearty, *Digital Codicology: Medieval Books and Modern Labor* (Stanford, 2023), especially Ch. 4 “Interoperable Metadata and Failing Towards the Future” (168-211).
[15] https://portail.biblissima.fr/en Accessed March 9, 2025.
[16] https://data.biblissima.fr/w/Accueil/en Accessed March 9, 2025.
[17] Europeana, “SPARQL API Documentation.” Last updated Nov. 4,2024. https://europeana.atlassian.net/wiki/spaces/EF/pages/2385870903/SPARQL+API+Documentation

Two services developed in recent years have demonstrated the potential value of an approach based entirely on LOD and knowledge graphs. The Mapping Manuscript Migrations Project brought together metadata for more than 220,000 manuscripts from three different data sources. It also demonstrated the analytical value of using the SPARQL query language to reason across a knowledge graph with more than twenty million statements expressed as RDF (Resource Description Framework) triples.[18] The redeveloped version of Digital Scriptorium applies a similar approach to manuscript metadata, covering more than 18,000 manuscripts from 39 North American institutions and enriching the descriptions with contextual semantic content drawn from LOD sources.[19] LOD-based initiatives can also be found in the field of Classics and Ancient History.[20]

This kind of approach can be scaled up to incorporate data from a range of sources and disciplinary areas relevant to medieval studies. The result will be an assemblage of knowledge in the form of a large and expanding Medieval Knowledge Graph, constructed from semantic building-blocks (entities like people, places, organizations, manuscripts, and objects) together with their properties and the relationships between them. Computational reasoning can then be applied to explore this graph for both quantitative and qualitative research questions. Various kinds of interfaces could be built on top of the graph, ranging from graph-specific software like SampoUI, ResearchSpace, or metaphactory[21] to a data exploration service like Yale University's LUX collections discovery platform.[22]

Rather than requiring a single giant database, however, the LOD and knowledge graph approach can be implemented as a distributed framework, in which the components can be separate but linked through identifiers which refer to the same people, places, manuscripts, and other entities. Nor would it require every contributing component to adopt the same standards or the same data model. As long as individual approaches are documented carefully and followed consistently, they can be linked by mapping from one to another. Including identifiers from services like the Virtual International Authority File (VIAF), Wikidata, the

---

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

[19] Mikko Koho, L. P. Coladangelo, Lynn Ransom, & Doug Emery, "A Wikibase Model for Premodern Manuscript Metadata Harmonization, Linked Data Integration, and Discovery," *Journal of Computing and Cultural Heritage* 16/3 (2023), article 56, 1-26. https://doi.org/10.1145/3594723

[20] Sarah Middle, *Investigating Linked Data Usability for Ancient World Research*. PhD thesis, The Open University, 2022. https://doi.org/10.21954/ou.ro.00014b1f

[21] https://seco.cs.aalto.fi/tools/sampo-ui/ ; https://researchspace.org/ ; https://metaphacts.com/solutions/products/metaphactory accessed March 9, 2025.

[22] Emma Metcalfe Hurst and Karina Wratschko, "LUX: Yale Collections Discovery," *ARLIS/NA Multimedia & Technology Reviews*, Dec. 2023. https://doi.org/10.17613/3hy1-pv45

Getty Thesaurus of Geographical Names (TGN), and GeoNames and publishing the data in LOD-compatible formats will enable them to be incorporated into larger knowledge graph structures.[23]

This would not be a Google-like service which searches huge amounts of textual materials for specific words or phrases. Nor would it be a "generative AI" service of the kind designed to produce sophisticated summaries drawing on textual material from across the Web.[24] Instead of assembling content from sentence structure patterns found in textual materials, as ChatGPT does, the Medieval Knowledge Graph would enable reasoning across semantic patterns created with human input. AI tools could still be used to help construct and populate this knowledge graph, however; Machine Learning can identify references to entities in textual and other sources (using Named Entity Recognition) and link them with other nodes in the knowledge graph.[25]

This kind of service opens up a range of possibilities. For a start, it can be extended outward beyond the initial focus on manuscripts and their texts, to join them up to other kinds of evidence about the medieval period: objects, artworks, archaeological sites, buildings, and so on. Publishing a range of vocabularies for people, places, works, and organizations as Linked Open Data can provide the nodes which glue together the Medieval Knowledge Graph.[26] For too long, the various different disciplinary perspectives on medieval studies have lived in their own separate worlds of databases and scholarly literature. A unifying knowledge graph can bridge these gaps and enable exploration, analysis, and visualization in a cross-disciplinary setting.

Such a knowledge graph can also be designed to make possible participatory services like annotation, collaboration, and contestation. Most existing databases and catalogues are built around what libraries call "authority control": approved forms of names and terminology which can only be edited or added to by a small number of authorized people.[27] But as initiatives like Australia's HuNI (Humanities Networked Infrastructure) have demonstrated, it is not only possible but worthwhile

[23] See Hyvönen," How to Create" for an overview of these processes.
[24] Space prevents a more extensive discussion of this point, but asking a service like ChatGPT questions like "How did Isidore of Seville influence Rabanus Maurus?" or " Who were the most important American manuscript collectors of the 20th century?" will illustrate both the abilities and the limitations of "generative AI": https://chatgpt.com/ accessed March 9, 2025.
[25] Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet, "Named Entity Recognition and Classification in Historical Documents: A Survey," *ACM Computing Surveys* 56/2, Article 27 (Feb. 2024), 47 pages. https://doi.org/10.1145/3604931
[26] Toby Burrows, "Linked Open Data and Medieval Studies: Some Lessons from the Mapping Manuscript Migrations Project," *International Journal of Humanities and Arts Computing* 16/1 (2022): 64-77.
[27] Toby Burrows, Deb Verhoeven, and Mike Jones, "Selling Our Soul (For Total Control)? Linked Open Data and GLAM," in *The Routledge Companion to Libraries, Archives, and the Digital Humanities*, ed. Isabel Galina Russell & Glen Layne-Worthey (London, 2024), 187-203.

to build digital infrastructure which can capture different perspectives on knowledge – reflecting the kind of contestation and collaboration which is at the heart of the humanities.[28] A truly collaborative knowledge graph would be a transformative achievement for medieval studies.

The barriers to this kind of development are more cultural and organizational than technical. Existing digital research projects have tended to be small-scale, self-contained efforts, with little reference to the larger research landscape and more general digital infrastructure. The current orientation in both funding and research is heavily weighted towards supporting the kinds of projects which Michael Gervers memorably described as a "dead end" involving endless reinvention of the same digital wheel.[29] The sustainability of these digital products beyond the initial project can often be tenuous and limited.[30] Library and museum databases, while generally much larger in scale and with better ongoing institutional support, are usually limited to the institution's own collections. Commercial databases like those published by Brepols and ProQuest, amongst others, exist behind subscription portals and have little if any outward links.[31]

A Medieval Knowledge Graph would encourage larger-scale research and more collaborative projects, but would also enable more specific investigations to draw on a wider range of contextual material. Funding criteria for projects would need to include a demonstrable orientation towards reusable and linkable data and a willingness to work in partnership with libraries and museums in publishing data. Equally important would be a commitment to sustainability and to minimizing the carbon footprint of large-scale data centres and proliferating cloud-based digital services.[32]

A range of interested parties would need to be brought together to organize this kind of collaboration for medieval studies: scholarly societies like the Medieval Academy of America, research library groups like the Consortium of European Research Libraries (CERL), funding bodies like the Mellon Foundation and the European Commission, and commercial publishers like Brepols. Without this kind of large-scale

[28] In some ways, Wikidata provides a similar approach, but it works best as a means for connecting different databases and LOD services by linking common identifiers for people, places, concepts, and objects.
[29] Michael Gervers, "DigiDeeds: Linking Databases and Developing Sustainability for Data," *International Medieval Congress*, Leeds, July 1-4, 2024. [unpublished paper]
[30] For example: James Smithies et al., "Managing 100 Digital Humanities Projects: Digital Scholarship & Archiving in King's Digital Lab," *DHQ: Digital Humanities Quarterly* 13/1 (2019). https://www.digitalhumanities.org/dhq/vol/13/1/000411/000411.html
[31] For example: https://about.proquest.com/en/products-services/patrologia_latina/ ; https://www.brepols.net/series/LLT-O accessed March 9, 2025.
[32] Keith L. Pendergrass, Walker Sampson, Tessa Walsh, and Laura Alagna, "Toward Environmentally Sustainable Digital Preservation," *The American Archivist* 82/1 (March 2019): 165–206. https://doi.org/10.17723/0360-9081-82.1.165 Joanna Tucker, "Facing the Challenge of Digital Sustainability as Humanities Researchers," *Journal of the British Academy* 10 (2022): 93–120. https://doi.org/10.5871/jba/010.093

collaborative approach, medieval studies is increasingly being swamped by the sheer numbers of different digital initiatives and projects which do not talk to each other, are difficult to track down, and are often not reusable or sustainable. A truly transformative approach will involve more than experimenting with the latest digital techniques or producing yet more silos of digital materials. Rather, it will require some serious joined-up thinking about how and why to bring individual initiatives together in the form of a Medieval Knowledge Graph for the benefit of future researchers.