Digital Humanities in the Nordic Countries 2018

Cultural heritage collections as research data

Cultural heritage materials held in institutional collections are crucial sources of evidence for many disciplines, ranging from history and literature to anthropology and art. They are also the subjects of research in their own right – encompassing their form, their history, and their content, as well as their places in broader assemblages like collections and ownership networks. They can be studied for their unique and individual qualities, as Neil McGregor demonstrated in his *History of the World in 100 Objects,* but also as components within a much larger quantitative framework.

Large-scale research into the history and characteristics of cultural heritage materials is heavily dependent on the availability of collections data in appropriate formats and sufficient quantities. Unfortunately, this kind of research has been seriously limited, for the most part, by lack of access to suitable curatorial data. In some instances this is simply because collection databases have not been made fully available on the Web. This is particularly the case with art galleries and some museums. Even where databases are available, however, they often cannot be downloaded in their entirety or through bulk selections of relevant content. Data downloads are frequently limited to small selections of specific records.

Collections data are often available only in formats which are difficult to re-use for research purposes. In the case of libraries, the only export formats tend to be proprietary bibliographic schemas such as EndNote or RefCite. Even where APIs are made available, they may be difficult to use or limited in their functionality. CSV or XML downloads are relatively rare. Data licensing regimes may also discourage re-use, either by explicit limitations or by lack of clarity about terms and conditions.

Even where researchers are able to download usable data, it is very rare for them to be able to feed back any cleaning or enhancing they may have done. The cultural heritage institutions supplying the data may be unable or unwilling to accept corrections or improvements to their records. They may also be suspicious of researchers developing new digital services which appear to compete with the original database.

As a result, there has been a significant disconnect between curatorial databases and researchers, who have struggled to make effective use of what is potentially a very rich source of computationally usable evidence. One important consequence is that re-use of curatorial data by researchers often focuses on the data which are the easiest to obtain. The results are neither particularly representative nor exhaustive, and may weaken the validity of the conclusions drawn from the research.

Some recent “collections as data” initiatives (such as collectionsasdata.github.io) have started to explore approaches to best practice for “computationally amenable collections", with the aim of "encouraging cultural heritage organizations to develop collections and systems that are more amenable to emerging computational methods and tools". Under the auspices of the Library of Congress and the Institute of Museum and Library Services, the Collections as Data programme "aims to foster a strategic approach to developing, describing, providing access to, and encouraging reuse of collections that support computationally-driven research" (Always Already Computational 2017). One of the drivers for this initiative is the perception that, as Miriam Posner argues, "Libraries and archives [and museums] are increasingly making their materials available online, but, as a general rule, these materials aren't of much use for computational purposes" (Posner 2017).

This paper focuses on three case studies of projects which are addressing these issues. The first project is "Collecting the West", in which Western Australian researchers are working with the British Museum to deploy and evaluate the ResearchSpace software, which is designed to integrate heterogeneous collection data into a cultural heritage knowledge graph in a Linked Data environment. The second project is HuNI – the Humanities Networked Infrastructure – which has been building a "virtual laboratory" for the humanities by reshaping collections data into semantic network graphs. The third project – "Reconstructing the Phillipps Collection", funded by the European Union under its Marie Curie Fellowships scheme – involved combining collections data from a range of digital and physical sources to reconstruct the histories of manuscripts in the largest private collection ever assembled.

To make services like these possible, collections data need to be made available in certain ways and under certain conditions. Recommendations for best practice, at the moment, tend to be focused mostly on processes and procedures, encompassing download formats, licensing, and availability in particular (Fitzpatrick 2017). These are undoubtedly important; having collections data easily accessible in bulk on the Web, under a Creative Commons licence which permits free reuse, is essential. Download formats are more debatable: APIs are not necessarily the best approach, given that their use is likely to require a significant level of technical expertise (Tauberer 2014). XML dumps and CSV files are easier to use, but may not contain all the elements in the source database.

As the interest of researchers in reusing collections data continues to grow, however, cultural heritage institutions increasingly need to start looking beyond simply making their data available for bulk downloading or via an API. One of the major use cases is to link together data from different institutions, without diminishing the semantic richness, in order to ask questions on a larger scale. At the moment, researchers are having to do much of this work themselves. This raises two important questions: should institutions help this process, and what kind of infrastructure might be built as a result?

The prominence of Linked Data in the solutions being adopted by researchers strongly suggests that institutions should make their data available in formats suitable for incorporation into Linked Data environments. While many institutions might not yet see a ‘business case’ for this approach, others like the British Library and the British Museum have already followed this route. Making available an RDF version of a relational database would be a significant contribution. But even embedding into that database identifiers which point to widely-used Linked Data ontologies and vocabularies like VIAF, GeoNames and Wikidata would be valuable. So too would taking a critical look at ways of improving the computational value of ownership and provenance data in these records. Enabling researchers and curators to annotate and add to the data is also emerging as an important requirement.

Beyond this, though, lies the wider landscape of digital infrastructure. The *Santa Barbara Statement on Collections as Data* (2017) observes that “Working toward interoperability entails alignment with emerging and/or established community standards and infrastructure.” At present, the Linked Data landscape is largely being built by research groups rather than cultural institutions, which still tend to focus on their own collections. In this context, an initiative like “Linked Pasts”, which has emerged from the Pelagios Commons, is an important development, offering a vision of joining up disparate Linked Data projects in the humanities to create a “wider ecosystem” (Grossner and Hill 2017).

As long as these kinds of initiatives remain tied to research projects, their future sustainability will be reliant on the uncertainty of grant funding. Collecting institutions should look closely at them as outcomes of the reuse of collections data, and consider seriously the value of partnerships with the researchers involved. They should recognize that there is a growing group of researchers who do not simply want to search or browse a collections database. There is an increasing demand for access to collections data for downloading and re-use, in suitable formats and on non-restrictive licensing terms. In return, researchers will be able to offer enhanced and improved ways of analyzing and visualizing data, as well as correcting and amplifying collection database records on the basis of research results. There are significant potential benefits for both sides of this partnership.

Bibliography

Always Already Computational: Library Collections as Data. 2017. https://collectionsasdata.github.io/

Burrows, Toby. 2017. “The History and Provenance of Manuscripts in the Collection of Sir Thomas Phillipps: New Approaches to Digital Representation,” *Speculum* 92 S1:S39-S64

Burrows, Toby, and Deb Verhoeven. 2015. “Aggregating Cultural Heritage Data for Research Use: The Humanities Networked Infrastructure (HuNI).” In *Metadata and Semantics Research, 9th Research Conference, MTSR 2015, Manchester, UK, September 9–11, 2015: Proceedings*, ed. Emmanouel Garoufallou, Richard J. Hartley, Panorea Gaitanou (Communications in Computer and Information Science, 544), 417-423. Cham: Springer.

Fitzpatrick, L. Kelly. 2017. “Shared Practices in Museum Open Collections Data,” *Medium*, February 22. https://medium.com/berkman-klein-center/shared-practices-in-museum-open-collections-data-72e924c4849a

Flanders, Julia. 2014. “Rethinking Collections.” In *Advancing the Digital Humanities*, edited by Paul Longley Arthur and Katherine Bode, 163-174. London: Palgrave MacMillan.

Grossner, Karl, and Timothy Hill. 2017. “From Linking Places to a Linked Pasts Network.” http://kgeographer.com/pubs/LinkedPastsNetwork_7Dec.pdf

Hyvönen, Eero. 2012. *Publishing and Using Cultural Heritage Linked Data on the Semantic Web*. San Rafael, CA: Morgan & Claypool.

MacGregor, Neil. 2010. *A History of the World in 100 Objects*. London: Penguin.

Posner, Miriam. 2017. “Actually Useful Collection Data: Some Infrastructure Suggestions.” In *Always Already Computational: Library Collections as Data: National Forum Position Statements.* https://github.com/collectionsasdata/collectionsasdata.github.io/raw/master/aac_positionstatements.pdf

Santa Barbara Statement on Collections as Data. 2017. https://collectionsasdata.github.io/statement/

Tauberer, Joshua. 2014. “Bulk data or an API?” https://opengovdata.io/2014/bulk-data-an-api/