From libraries as patchwork to datasets as assemblages?

Dr Mia Ridge, Digital Curator, British Library

The British Library's collections are vast, and vastly varied, with 180-200 million items in most known languages. Within that, there are important, growing collections of manuscript and sound archives, printed materials and websites, each with its own collecting history and cataloguing practices. Perhaps 1-2% of these collections have been digitised, a process spanning many years and many distinct digitisation projects, and an ensuing patchwork of imaging and cataloguing standards and licences. This paper represents my own perspective on the challenges of providing access to these collections and others I've worked with over the years.

Many of the challenges relate to the volume and variety of the collections. The BL is working to rationalise the patchwork of legacy metadata systems into a smaller number of strategic systems.[1] Other projects are ingesting masses of previously digitised items into a central system, from which they can be displayed in IIIF-compatible players.[2]

The BL has had an 'open metadata' strategy since 2010, and published a significant collection of metadata, the British National Bibliography, as linked open data in 2011.[3] Some digitised items have been posted to Wikimedia Commons,[4] and individual items can be downloaded from the new IIIF player (where rights statements allow). The BL launched a data portal, https://data.bl.uk/, in 2016. It's work-in-progress - many more collections are still to be loaded, the descriptions and site navigation could be improved - but it represents a significant milestone many years in the making. The BL has particularly benefitted from the work of the BL Labs team in finding digitised collections and undertaking the paperwork required to make the freely available. The BL Labs Awards have helped gather examples for creative, scholarly and entrepreneurial uses of digitised collections collection re-use, and BL Labs Competitions have led to individual case studies in digital scholarship while helping the BL understand the needs of potential users.[5] Most recently, the BL has been working with the BBC's Research and Education Space project,[6] adding linked open data descriptions about articles to its website so they can be indexed and shared by the RES project.

In various guises, the BL has spent centuries optimising the process of delivering collection items on request to the reading room. Digitisation projects are challenging for systems designed around the 'deliverable item', but the digital user may wish to access or annotate a specific region of a page of a particular item, but the manuscript itself may be catalogued (and therefore addressable) only at the archive box or bound volume level. The visibility of research

[1] The British Library, 'Unlocking The Value: The British Library's Collection Metadata Strategy 2015 - 2018'.
[2] The International Image Interoperability Framework (IIIF) standard supports interoperability between image repositories. Ridge, 'There's a New Viewer for Digitised Items in the British Library's Collections'.
[3] Deloit et al., 'The British National Bibliography: Who Uses Our Linked Data?'
[4] https://commons.wikimedia.org/wiki/Commons:British_Library
[5] http://www.bl.uk/projects/british-library-labs, http://labs.bl.uk/Ideas+for+Labs
[6] https://bbcarchdev.github.io/res/

activities with items in the reading rooms is not easily achieved for offsite research with digitised collections. Staff often respond better to discussions of the transformational effect of digital scholarship in terms of scale (e.g. it's faster and easier to access resources) than to discussions of newer methods like distant reading and data science.

The challenges the BL faces are not unique. The cultural heritage technology community has been discussing the issues around publishing open cultural data for years,[7]in part because making collections usable as 'data' requires cooperation, resources and knowledge from many departments within an institution. Some tensions are unavoidable in enhancing records for use externally - for example curators may be reluctant or short of the time required to pin down their 'probable' provenance or date range, let alone guess at the intentions of an earlier cataloguer or learn how to apply modern ontologies in order to assign an external identifier to a person or date field.

While publishing data 'as is' in CSV files exported from a collections management system might have very little overhead, the results may not be easily comprehensible, or may require so much cleaning to remove missing, undocumented or fuzzy values that the resulting dataset barely resembles the original. Publishing data benefits from workflows that allow suitably cleaned or enhanced records to be re-ingested, and export processes that can regularly update published datasets (allowing errors to be corrected and enhancements shared), but these are all too rare. Dataset documentation may mention the technical protocols required but fail to describe how the collection came to be formed, what was excluded from digitisation or from the publishing process, let alone mention the backlog of items without digital catalogue records, let alone digitised images. Finally, users who expect beautifully described datasets with high quality images may be disappointed when their download contains digitised microfiche images and sparse metadata.

Rendering collections as datasets benefits from an understanding of the intangible and uncertain benefits of releasing collections as data and of the barriers to uptake, ideally grounded in conversations with or prototypes for potential users. Libraries not used to thinking of developers as 'users' or lacking the technical understanding to translate their work into benefits for more traditional audiences may find this challenging. My hope is that events like this will help us deal with these shared challenges.

[7] For example, the 'Museum API' wiki page listing machine-readable sources of open cultural data was begun in 2009 http://museum-api.pbworks.com/w/page/21933420/Museum%C2%A0APIs following discussion at museum technology events and on mailing lists.