Where do we go with Union Catalogues?

The United Kingdom boasts union catalogues for its major research libraries, journal holdings, archives and, most recently, for its public library collections. For researchers wanting to locate material across the UK, such aggregations have long served as a first stop for researchers wanting to find the right material and also provided a showcase for our formidable research collections.

In the global networked environment, search engines and social networks can fulfil much of the functionality of union catalogues and have become the natural places to which our users go for search and discovery, even in academic situations.

Right now, there is a 'disconnect' between the data describing our collections and the places users first turn to start their searches. This can be fixed by exposing descriptive data to wider audiences beyond the silo of the local catalogue, but data publishing is a fast moving area with little obvious short-term institutional-level gain and some start-up barriers.

Publishing library data to the open web at the level of a national aggregation would utilize existing skill sets and infrastructure, minimize risk and maximize impact.

EDMUND CHAMBERLAIN
Systems Development Librarian / Head of Discovery & Access
Digital Services
Cambridge University Library

"The 'Magimix' cake cookbook was fantastic, the best thing Mary Berry ever wrote. My mum relied on it but now you can't find it anywhere."

Not the usual reference query I might handle in a Cambridge library, but instead something my wife mentioned during the latest baking programme on BBC One.

As a loyal systems librarian, I swiftly grabbed my laptop and turned to my institution's discovery platform with the sole intent of proving her wrong. After all, it describes, amongst other things, the collections of a national legal deposit library.

I was thwarted, however. No mention there of said cookbook. The latest webscale offering we are trialling? Again, nothing. So I tried COPAC, the UK Research Libraries union catalogue. Nothing there. Then the Library of Congress. No results. AbeBooks? Nothing.

Dejected, I turned to Google. A few mentions on forums (mostly avid bakers concurring with my wife) but the fourth or fifth result down was a reference to Facebook[1]. It turned out that the book had its own page on the popular social network (complete with two 'likes') but no entry in any UK major catalogue I checked.

> "… Google and Facebook between them found me the item I wanted, but … these were not the first places I went to."

What is surprising here is not that Google and Facebook between them found me the item I wanted, but that these were not the first places I went to. As a librarian, I'm embedded in world of academic search services and have a personal knowledge of collections described in union catalogues. I will instinctively turn to them for bibliographic queries, but these days I suspect I am in a minority, even in academic circles.

Lorcan Dempsey has succinctly described the reasons for this behaviour pattern[2]. With academic users living in an information-rich environment but often in time-poor situations, it is only natural that they gravitate to the highest possible layer of aggregation, the global search engine. In this model, local discovery services and national-level aggregations tend to get skipped over, despite their rich and valuable content.

What is possibly even more surprising is how Facebook came to know about the 'Magimix cookbook'[3] in the first place. Bibliographic data used to generate the page actually originated from the Harvard University library catalogue, which was last year published openly for anyone to use and reuse as they see fit[4]. In this case, it has been embedded into Facebook's graph search[5], a complex semantic structure showing the relationships between people, objects and events in the social network. There, my wife's approval of this work is now visible to all of her friends.

The value of opening data to the web

It was thanks to Harvard publishing their bibliographic data under a permissive licence that I was able to confirm the existence of this book. But Hollis, the Harvard Library catalogue, although a splendid resource, was not on my list of places to try, mainly because I live and work in the other Cambridge.

"It is hard not to argue with the idea that discovery of academic material is increasingly happening outside of the library search domain."

Nonetheless, the whole episode neatly demonstrates the value of libraries opening up the data describing their collections to the global network of the internet, as indexed by search engines. Whilst search engines lack the fidelity and functionality of academic-centric search offerings, getting the right data into them gives users at least a fighting chance of finding the right material without having to a) know about the catalogues in the first place, and b) trawl through several of them as I had done.

It is hard not to argue with the idea that discovery of academic material is increasingly happening outside of the library search domain. Libraries need to acknowledge this trend and work with it. Getting our collection data out there under a liberal licence is one way to help make this happen, either via dumps of data or by allowing web crawlers to access online catalogue pages. The model has worked wonders for e-commerce, which sees search engine optimization as a key means to drive business growth.

The Magimix cookbook example worked well in search engine results as it is a niche case. Library data describing more popular material would have to sit alongside results from Amazon and other popular sites.

Exposing such niche cases offers incredible opportunities for libraries. Replace obscure 1970s cookbooks with rare or early printed books, scientific data sets, rare material or unique collections of manuscripts and you move into the world of genuine academic use cases. Libraries could finally fulfil the promise of becoming 'a very long tail of scholarly and cultural materials'[6], putting the pieces in place to link up the right researchers with the right material in emerging global data structures.

Problems with institutional-level publishing

Since Cambridge University Library opened up its data in 2011 through the Open Bibliography[7] and COMET[8] projects, the UK higher education (HE) and cultural heritage sectors have taken great strides in publishing large amounts of open bibliographic data. The web pages for the Jisc-funded Discovery programme offer an insight into the advantages in creating an ecosystem of open data. Ben Showers' article in the July 2012 issue of this publication provides a great summary of this work to date[9].

It is also encouraging to see such data being reused and repurposed to new ends. The British Library's Open British National Bibliography data set has reportedly seen over two million requests for data. Clearly, we are only beginning to see the advantages this will bring.

However, there are some caveats. Very few institutions currently publishing data do so routinely on their own. Cambridge and others have mostly relied on external funding to do so. As a single institution, publishing data in house looks like a potentially complex activity

that may be seen to fail if no one immediately makes use of it. Standards and technologies are changing, and the intriguing potential of unknown medium- to long-term benefits[10] does not necessarily fit in alongside more obvious institutional demands, such as enhancing the student experience.

To a budget-conscious library director, it could be seen as an expensive luxury or technical distraction. This is not necessarily the case. There are lightweight approaches to getting data on your collections harvested and published on the web.

Most promising of these is the schema.org initiative[11], which allows structured descriptive information to be inserted into web pages in a way that can be intelligently crawled by search engines. Doing so allows them to understand the author and publisher of a work as separate data elements, not just as text on a web page.

"The bigger, more reputable and more linked to you are, the more likely search engines are to notice you …"

OCLC is leading the way in exploring how libraries can use this technology, looking at getting library descriptive data and holdings information read and understood by search engines. Imagine being able to look for a book in Google, and have it automatically show you a map with the nearest library holding the work along with information on how many copies are currently checked in. Google has changed small things in my life, like finding a local cinema that is showing the film I want to see. Imagine if we could do this for our collections.

Aggregate to disseminate

Regardless of the mechanism, publishing on your own is also not that likely to get you noticed by large consumers of data such as search engines and social networks. They tend to value larger sources.

The bigger, more reputable and more linked to you are, the more likely search engines are to notice you and promote you in their results. With this in mind, does it not make sense to look at existing aggregations as the natural platform to start publishing our data? Such aggregations arguably carry sufficient volume and authority to get noticed and harvested by big players.

Sidestep the OPAC

Another reason for using aggregators to publish data lies in the very nature of our institutional-level library systems infrastructure. In-house catalogues (OPACs) are not generally suitable for widespread crawling by search engines. A heavy level of page-crawling over an OPAC by search engines risks toppling the whole library management system that OPACs form part of. Commercial discovery platforms currently lack the flexibility and functionality to act as data publishing platforms[12].

To get web pages crawled by search engines as OCLC is currently doing, we need better more flexible infrastructure with more control, often found at the aggregation level.

"The more customers that use a product, the bigger the underlying shared data set can grow."

Library service platforms – level the playing field with open aggregations

Not only is the way we expose and share data changing, but the ways in which we as librarians create data are changing as well. The new generation of the LMS, library service platforms[13], promise to transform data creation with streamlined workflows built around global data sets. These products have models of aggregation based around the customer base. The more customers that use a product, the bigger the underlying shared data set can grow.

This situation has the potential to create a network effect[14], whereby libraries feel drawn to use the service provider with its hands on the largest and best aggregation of data. The more libraries that join the leading supplier, the harder it becomes for anyone else to challenge them. This is bad for the marketplace and ultimately bad for libraries as customers.

In the UK, the KB+ project[15] is partly addressing a parallel problem related to e-journal holdings information and licensing terms. It is aiming to create a national-level store of licence and holdings data that is owned and managed by a community. System vendors can also take and contribute data from and to the store. This allows libraries as customers to migrate from one electronic resource management system (ERM) to another with confidence that data will be uniform in quality across the marketplace.

A national-level aggregation of bibliographic data could potentially fulfil a similar role, underpinning the new generation of library service platforms as a storehouse of well-curated bibliographic data.

"Discovery is not just about knowing what material is out there, but why it is useful to you."

Social search

Discovery is not just about knowing what material is out there, but why it is useful to you. The assessment of a work has long been a vital use case for library catalogues and the data inside them. Recommendations and other social interactions also play a pivotal role in assessment, particularly those from a peer or trusted source, but also anonymously, based on aggregated activity data. They have always taken place in an academic social context, but increasingly this discourse can and does happen online. Joining up online recommendations with descriptive information seems like a sure-fit for libraries. Jisc has already funded a number of initiatives in this space[16].

If libraries were to usefully enter this area en masse, then we might also consider a neutral house to store and aggregate recommendation and usage information. As with the bibliographic data it might link to, such usage information is too valuable to be left in the hands of a single discovery platform supplier. A national-level aggregation might be valuable in this respect.

Conclusion – building the business case

Aggregations of data are a wonderful resource for libraries and remain so in the age of the internet. With the above use cases, and potentially more, it is not hard to see how current aggregations could gain new leases of life as data publishing platforms. Many of the original use cases for aggregation, including acting as a collective shop front for UK research libraries, are still just as valid now. With this suggested direction of change, aggregations have the potential to reach far greater audiences by pushing data directly to the social networks and search engines that engage our users daily. Along the way, they can solve a few problems facing libraries as well.

There are challenges. Many aggregations currently depend on sale of bibliographic data to fund their efforts, so a change in business model would be required. One option is to move to selling services based around open data, rather than simply raw data itself. Rufus Pollock, co-founder of the Open Knowledge Foundation, is particularly fond of a phrase: 'Data is a platform not a commodity: you build on it rather than sell it. And that's why it should be open'[17].

The financial and organizational barriers to actually making this change are likely to be complex and difficult to navigate. But the right combination of open data and the ability to reach new library users by getting collection data into their favourite online environments is arguably too tasty a mix to ignore.

References and notes

1. Facebook - Cake-making-with-Magimix-32-recipes-written-by-Mary-Berry-for-food-processors https://www.facebook.com/pages/Cake-making-with-Magimix-32-recipes-written-by-Mary-Berry-for-food-processors/452132341502971?fref=ts (accessed 2 April 2013).
2. Demspey, L, Thirteen Ways of Looking at Libraries, Discovery, and the Catalog: Scale, Workflow, Attention, *EDUCAUSE Review Online*, 2012: http://www.educause.edu/ero/article/thirteen-ways-looking-libraries-discovery-and-catalog-scale-workflow-attention (accessed 2 April 2013).
3. edu/?itemid=|library/m/aleph|011745164 (accessed 2 April 2013). Never published on its own and only released with food processors, hence its scarcity.
4. Harvard Library Bibliographic Dataset: http://openmetadata.lib.harvard.edu/bibdata (accessed 2 April 2013).
5. Facebook Graph Search: http://en.wikipedia.org/wiki/Facebook_Graph_Search (accessed 2 April 2013).
6. Dempsey, L, Libraries and the Long Tail – Some Thoughts about Libraries in a Network Age, *D-Lib Magazine*, 2006, 12(4).
7. Open Bibliography: http://openbiblio.net/p/jiscopenbib/ (accessed 2 April 2013).
8. Cambridge Open Metadata (COMET): http://cul-comet.blogspot.co.uk/ (accessed 2 April 2013).
9. Showers, B, Data-driven library infrastructure: towards a new information ecology, *Insights*, 2012, 25(2), 150–54: http://dx.doi.org/10.1629/2048-7754.25.2.150 (accessed 22 April 2013).
10. Open Bibliographic Data Guide – real world use cases http://obd.jisc.ac.uk/examples (accessed 2 April 2013).
11. Schema.org http://schema.org/ (accessed 2 April 2013).
12. Vu-Find, an source library-centric discovery platform offers the ability to create sitemaps and web pages that can be easily crawled by search engines.
13. Grant, C. The Future of Library Systems: Library Services Platforms. *Information Standards Quarterly*, 2012, 24(4), 4–15.
14. Network effect – Wikipedia The Free Encyclopedia: http://en.wikipedia.org/wiki/Network_effect (accessed 2 April 2013).
15. KB+: http://www.jisc-collections.ac.uk/knowledgebaseplus/ (accessed 2 April 2013).
16. RISE – recommendations improve the search experience: http://www.open.ac.uk/blogs/RISE/ (accessed 2 April 2013).
17. Pollock, R, Data is a platform –not a commodity: http://www.shuttleworthfoundation.org/data-is-a-platform-not-a-commodity/' (accessed 2 April 2013).

Edmund Chamberlain, Systems Development Librarian / Head of Discovery & Access, Digital Services
Cambridge University Library
E-mail: emc59@cam.ac.uk

To cite this article:

Chamberlain, E, Where do we go with Union Catalogues?, *Insights*, 2013, 27(2), 180–184, http://dx.doi.org/10.1629/2048-7754.83