Data-driven library infrastructure: towards a new information ecology

Based on a paper presented at the 35th UKSG Conference, Glasgow, March 2012

Data usually lies hidden beneath the myriad systems which it powers; like electricity it is only visible through the things it makes possible. Yet, data has the potential to transform the services and systems that institutions rely on to deliver world-class teaching, learning and research.

In this article, the opportunities of adopting what might be termed a data-driven approach to library systems and services are explored. What this approach means for libraries and how it can help transform and improve the services and support libraries provide to their users are also considered. In exploring the potential of a data-driven approach to library infrastructure and the benefits it has for libraries (as well as publishers, system vendors and third parties), it is clear a new information ecosystem begins to emerge.

The library finds itself transformed, effort is redistributed and the library's services and systems become more agile and take on an entrepreneurial flair. This emerging information ecology is, arguably, the study of the future of the academic library.

BEN SHOWERS
Programme Manager
JISC

In describing an artist's relationship to her work, the author and copyright activist Cory Doctorow employs a metaphor that contrasts the way mammals look after their young with that of a dandelion. Doctorow explains that mammals make very few copies of themselves and invest significant resource and effort in those few copies. In contrast, the dandelion produces a mass of copies (its seeds) and leaves them to the fate of the wind. For dandelions, the losses can be significant, but, critically, every spring the pavements are full of dandelions! What at first appears carelessness on the part of the dandelion is revealed as an impressive survival strategy.[1]

Doctorow's point is that the internet changes everything: the expectations of readers and consumers have changed, and artists therefore need to radically change the way they approach distributing their works. A similar re-conceptualization needs to be effected in the information landscape that libraries find themselves a key part of.

In this article, I want to explore the opportunities of adopting what might be termed a data-driven approach to library systems and services. To understand what it means for libraries and how it can help transform and improve the services and support libraries provide to their users. In exploring the potential of a data-driven approach to library infrastructure[2] and the benefits it has for libraries (as well as publishers, system vendors and third parties), it is clear a new information ecosystem begins to emerge. The library finds itself transformed, effort is redistributed and the library's services and systems become more agile and take on an entrepreneurial flair.

> "The library finds itself transformed, effort is redistributed and the library's services and systems become more agile and take on an entrepreneurial flair."

The current library systems landscape

Data, in one form or another, is at the heart of the systems and services that the library delivers: from the management of electronic resources to the discovery of books and journals. Yet, while libraries exemplify a data-centric service model, with an understanding of (meta)data not commonly found within the wider institution, there are a number of critical obstacles that prevent libraries from realizing the potential of the data that underpins their services and systems (both internal and user facing):

- discrete systems developed for specific problems: within a library systems context, the data is king. Without accurate, authorative and timely data it is difficult to manage and deliver effective services. Yet, a focus on the systems often fails to address deeper underlying issues that proliferate around the data underpinning these systems. These issues usually result in a new system development or a patch to the existing one, ad infinitum
- thinking in verticals: the current approaches to library systems encourage what might be called vertical thinking – a failure to think more holistically. Within a library context this would be a failure to affect the wider institutional 'ecosystem' with its systems and the various data that underpins those services
- fragility of systems: there is a valid concern about the sustainability of library systems; they quickly become outdated, or lack functionality. There is a fragility to library systems – born from these first two problems. The data that underpins these systems tends to be more robust, and endures while the systems crumble and fall into disrepair.

Libraries find themselves needing to tread a 'careful line between securing a good return on investment and more imaginative leaps to ensure accessibility and relevance to their user communities'[3]. These obstacles increase the risk and complexity of any 'imaginative leaps' as well as creating a number of direct issues for libraries, which include:

- the proliferation of 'systems' for managing resources: electronic resource management systems (ERM), spreadsheets, library management systems (LMS) and so on
- connected to the above are the compromised workflows. These become inefficient, resource intensive and increasingly ineffective
- there is an inevitable knock-on effect for the users, with poor services and the failure to find resources leading to reduced expectations from users
- redundancy becomes a significant implication: If you are not providing the functionality students and researchers want, someone else will.

The underlying problem of library systems has an inextricable flow that starts with the back-office systems and eventually out to the user and the services they are using. These obstacles and their attendant implications point to a fundamental problem that the current approach to library systems and services perpetuates: the infrastructure is, to a large extent, built upon pre-web assumptions. In general it is difficult to contribute data to library systems (this applies as much to users as to the library itself); systems are delivered 'finished' and lack a continuous cycle of development and evolution; only certain parties can develop services – only certain parties have access to the data. It is questionable whether adopting new systems as a solution to this problem would succeed, without radically rethinking how library systems and services are conceived, developed and deployed.

What is data-driven infrastructure?

Adopting a data-driven approach to infrastructure is about re-conceiving the way you approach a problem. It encourages a focus on the data that underpins the solution you want to develop, where it comes from, how it affects and influences other systems, and how it can help affect and enhance the solution you are building.

It starts to disrupt long-held assumptions. There is no longer a separation between the source of data and the service – rather they are all part of one data circle; one flow, that, in the case of a library ultimately leads the user to their resources. Openness becomes a virtue, making sharing easier and helping reduce the rekeying of data over and over again. The adoption and implementation of standards becomes an essential component, and interoperability is an assumed norm.

These components of data-driven infrastructure might be expressed more formally, adopting a model used by Tim Berners-Lee in his articulation of linked data[4]. In this context we might talk about the 'three stars' of data-driven infrastructure[5]:

1) *Make data available for re-use*: enabling your data to be used and reused by anyone. Fundamentally this is about having the right licence for your data. It also means thinking beyond your own interface, and adopting application programming interfaces (APIs) and programmatic interfaces

2) *Open and reusable vocabularies*: shared and reusable vocabularies save on effort and enable sharing between systems and greater interoperability

3) *Join data up:* bringing different data sets together enables greater context and meaning as well as enriching your own descriptions. Multiple data sets joined together become vastly more powerful, and that power increases exponentially as more data is joined (think of DBPedia[6] and other linked-data initiatives).

There are increasingly compelling arguments for adopting a more data-centric approach to the development of technical infrastructure. Nowhere is this argument more compelling than in the data-heavy world of the library.

Towards a new information ecosystem

If you are a researcher writing a thesis on a particular subject, it makes sense that you would want to know about other types of resource your library has that may be of interest (archive documents, multimedia, etc.). That is why discovery of local resources is such a critical issue for libraries. But, you might argue that the researcher would also want to know about collections that are held just down the road in a gallery or museum (and be able to search and discover them from one source). Alternatively, it might be that the researcher is looking to search across a very specific aggregation of data, one that is niche and probably only they could assemble. These two use-cases (the broad and the very specific) have potential in an ecosystem where data is open and usable and where the barriers for innovation are greatly reduced.

This vision is exemplified in the 'Discovery' initiative that is a collaboration between JISC, Research Libraries UK (RLUK) and Mimas at the University of Manchester[7]. The initial focus of the work is to make cultural heritage metadata open and usable; creating a fertile ecosystem where innovative new services can develop, where niches can be colonised and new niches created. Discovery is also agnostic about the location of innovation: it can be undertaken by the library, by researchers, commercial vendors or third parties. The important part is not *who* can innovate, but that they *can* innovate.

Initiatives such as Discovery also take advantage of scale in opening up and sharing data: something can be done once and shared with everyone; actions do not have to be repeated locally hundreds of times. This is the aim of services like Knowledge Base+ (KB+)[8]: a shared academic knowledge base for electronic resources. This shared knowledge base is an example of how, through sharing data as a community, libraries can improve the data they rely on, ensuring it is timely, high quality and as openly available as possible. One of the aims of the KB+ project is to address the fragmentation of systems and workflows that attend the management of e-resources, and which currently make it a very inefficient and frustrating process for the library. By sharing, updating, improving and re-sharing the data, libraries create better data while reducing the effort required managing those resources.

> "By sharing, updating, improving and re-sharing the data, libraries create better data while reducing the effort required managing those resources."

However, libraries and the data they create and consume do not exist in a vacuum: The data demands that the wider 'flow' of bibliographic data is exploited. Both Discovery and KB+ are attempting to affect the wider metadata landscape by engaging with commercial vendors, publishers and third parties. KB+, for example, is ensuring that the data the service provides is both enhanced by third-party data (usually the source of library data), and that the data managed by the service also flows back into their library products (link resolvers, ERMs, etc.). The implications of this are that a library need not implement a new system to benefit from the KB+ data, but that data will enhance those systems the library already uses. Indeed, a library may have no contact with the KB+ service, but still benefit from its improved data.

Within this emerging data ecosystem there are smaller data ecosystems that can spring to life. The University of Cambridge Library has had an open API to the library's catalogue data for some time. In late 2011, a student developed an app for the library catalogue independently of the library (without its explicit knowledge or permission). It made the front cover of the University student paper[9], generated significant coverage for the library, and while I don't know what its usage is like, I suspect an app created by a student is as likely (or more likely) to be used by other students as anything else. It also demonstrates the potential that accessible data has for enabling innovation to bloom outside of its usual confines. But, such developments are not without significant implications for the institution.

Another interesting example is the Lemon Tree[10] game developed by the University of Huddersfield Library. The game takes the relatively mundane arena of library circulation and its attendant data, and turns this into a game for users. A student interacting with the library by borrowing books, for example, enables the game to collect data and uses this to update the game and enhance the student's interactions with the library. The game also has a social element so students can see how their friends are doing – and stimulates a sense of competition too. The circulation data is used to allow students to interact with the library in an entirely new way.

A flourishing information ecosystem, with a combination of the right skills and effort targeted at the right places, has huge implications for libraries and information providers in general. The nature of data is one that encourages an iterative, responsive approach to service design and development. Data is collected, acted upon and new data is subsequently generated (data: action: data, and so on). In this emerging ecosystem such an approach begins to look more like an entrepreneurial attitude, one that would not look out of place in a tech start-up business. It might be argued that libraries need to be able to adopt such an attitude to compete with companies like Google et al, who are now their direct competitors.

Underpinning this entrepreneurial approach to library innovation is the emergence of support infrastructures. There are platforms, such as JISC Elevator[11] (a crowdfunding site for innovation ideas in education), and a Library and Information Science Stack Exchange proposal that uses the computer programming Q&A website as a basis for library information and knowledge exchange[12]. These sites are about providing a technical infrastructure that supports institutions and libraries in moving quickly with ideas and developing innovative services that improve the user experience.

It is clear that data compels us to appreciate and take account of the wider information ecosystem that it is a part of. In order for services to adapt to new and emerging use-cases, and for users to develop their own niche use-cases, innovation must be allowed to flourish everywhere. With the complex data ecosystem that libraries are tied to, this is an important implication, and one that has the potential to transform the way libraries deliver services to their users.

> " … data compels us to appreciate and take account of the wider information ecosystem that it is a part of."

Conclusion

Academic libraries increasingly find themselves at the centre of a scholarly environment transformed by data: whether it is the 'big data' of the sciences or the progressively data-driven practices of humanities students and researchers[13]. Wider social changes are resulting in an increasingly open approach to academic research and the associated data, as well as moves by national libraries to open up their metadata, for example, the British Library's Free Data Services[14]. Accompanying this data transformation are the technologically-driven changes that libraries, whether public, private or academic, are confronting, and the impact these changes are having on the expectations of their users and the services developed in response.

But, as I have tried to show, libraries are not simply reactive observers of this phenomenon; instead they are at the forefront of understanding the potential of the data that underpins their services and systems to deliver innovative services to students and researchers.

Adopting a data-driven model for the development and deployment of library infrastructure has the potential to transform the way the library interacts with its users and enables the development of new services. Importantly, such a data-centric approach changes the very nature of how libraries conceive and tackle the problems they face, both now and in the future.

References and notes

1. Doctorow, C, Think Like a Dandelion:
   http://www.locusmag.com/Features/2008/05/cory-doctorow-think-like-dandelion.html (accessed 15 May 2012).
2. In this paper the term *infrastructure* will refer specifically to the systems that libraries use in either their management of resources – the management systems, and the systems that deliver services and content to the users – discovery layers and platforms, for example.
3. Library Management Systems: Investing Wisely in a Period of Disruptive Change:
   http://www.jisc.ac.uk/publications/briefingpapers/2008/librarymanagementbp.aspx (accessed 9 May 2012).
4. Linked Open Data star scheme by example:
   http://lab.linkeddata.deri.ie/2010/star-scheme-by-example/ (accessed 8 May 2012).
5. The ‘three stars of data-driven infrastructure’ is an idea developed by my colleague Rachel Bruce, an innovation director at JISC.
6. DBPedia:
   http://dbpedia.org/About (accessed 9th May 2012).
7. Discovery:
   http://discovery.ac.uk/ (accessed 11 May 2012).
8. Knowledge Base+:
   http://www.jisc-collections.ac.uk/knowledgebaseplus/ (accessed 10 May 2012).
9. Varsity: ‘App enables students to search UL catalogue’:
   http://www.varsity.co.uk/news/3702 (accessed 11 May 2012).
10. Lemon Tree Blog:
    https://library.hud.ac.uk/lemontree/about.php (accessed 11 May 2012).
11. JISC Elevator:
    http://elevator.jisc.ac.uk/ (accessed 11 May 2012).
12. Libraries and Information Science Stackexchange
    http://area51.stackexchange.com/proposals/12432/libraries-information-science (accessed 11 May 2012).
13. An interesting overview of the increasingly data-driven world of research can be found here: Krotoski, A K, Data-driven research: open data opportunities for growing knowledge, and ethical issues that arise, *Insights*, 2012, 25(1), 28–32, doi: 10.1629/2048-7754.25.1.28
14. Free Data Services:
    http://www.bl.uk/bibliographic/datafree.html (accessed 11 April 2012).

Ben Showers, Programme Manager
JISC, Brettenham House (South Entrance), 5 Lancaster Place, London WC2E 7EN, UK
E-mail: b.showers@jisc.ac.uk

To cite this article:

Showers, B, Data-driven library infrastructure: towards a new information ecology, *Insights*,2012, 25(2), 150–154, doi:10.1629/2048-7754.25.2.150

To link to this article:

http://dx.doi.org/10.1629/2048-7754.25.2.150