Collection development or data-driven content curation?

DOI:
10.1108/LM-05-2016-0044

Document Version
Accepted author manuscript

Collection development or data-driven content curation?

Introduction

The University of Manchester Library is big. With around 350 staff and over four million items, size is one of its defining characteristics. The Library is ambitious and innovative, and less risk-averse than some other HE libraries. A major review and restructure of the former 'Research & Learning Support' division in 2012/13 positioned librarians to up-skill and specialise to meet new challenges, organising staff into functional teams and doing away with the traditional subject librarian model, with its broad spread of activities – not least of which was book selection.

After a year of the new structure 'bedding in', a vacuum in collection development was identified. The disappearance of the subject librarians created a managerial challenge for the curation of our similarly large collections, to which pragmatic solutions were needed. The feasibility and efficacy of a 'research resources robot' was humorously hypothesised, and a major library strategy project was designed, the overall aim of which was to experiment with what can be done with data and how far automation can be employed in the services of collection development at a large, research-intensive university. This paper will describe managerial responses to some of the unexpected challenges of the new structure.

Observations

The functional team structure provided the Library with an excellent basis from which to build up new services in support of the University's strategic goals in both teaching and research. The Teaching & Learning team was able to rationalise the Library's training offer and improve standards of delivery across the board. Employment of an e-learning technologist was crucial in developing the *My Learning Essentials* skills programme, which won the Innovative Blended Learning category of the 2014 Blackboard Catalyst Awards, and is now featured in the American Libraries' Association Instruction Section's Peer-Reviewed Instructional Materials Online (PRIMO) project, which showcases exemplary online instruction materials. In Research Services, the Library's bibliometrician has been so successful in providing tailored reports on citation analysis to our various Schools and Research Directors, that he was awarded the distinction of 'outstanding contribution' by the University last year. The Open Access team has rapidly developed expertise and new systems in support of open science and the RCUK compliance targets, collaborating with Jisc and other institutions across the North West and sharing best practice on the OpeNWorks website.

By contrast, collection development had a much lower profile, with little dedicated resource and no strategic leadership. However, the approach at Manchester is to deliver the Library strategy through projects, with individual project managers being supported by a central Project Office. A large project was agreed to look at 'collection development and profiling'. The basic hypothesis was that to supplement academically led resource selection, data could be exploited, through the increasing functionality in library management systems for example, and with standardisation of usage statistics, to support or even automate decisions about purchase, renewal of subscriptions, and relegation of material. A second hypothesis was that descriptive profiles of our various collections could be dynamically linked to profiles of academic disciplines, out of which might emerge new guidelines for collection development. The project management structure would provide a framework for testing these basic hypotheses and working towards specific objectives, as well as offering an established evaluation mechanism for reviewing the outcomes.

Some 18 members of staff from across the Library teams volunteered to participate in this somewhat experimental project, which was scheduled to run for 20 months. The overall project approach was to focus on a number of individual problems/possibilities in different areas, suggested by early pilot work in the previous year, and then attempt to combine them into a coherent framework.

There were differentiated approaches within the individual areas. Outside the project, the approach to developing teaching & learning collections might be labelled 'reactive yet dynamic'. For example, a centralised 'recommended reading' mail address has been introduced for academics to use when contacting the Library. We have made full use of CLA-cleared[1] digitisation requests. We have increased the availability of patron-driven acquisition for both e-books and more recently print books, in a 'books on demand' initiative which has democratised the book selection process and empowered students. For research level resources we have applied more 'proactive and dynamic' approaches in the project, involving three kinds of profiling.

Profiling for purchase 1: approval plans

The project employed three different kinds of purchasing profile. The first was a straightforward 'approval plan' – although the name is not fully accurate, since there was no 'approval' step. It was considered appropriate for those areas showing indications of developing gaps, as faculty were not actively selecting material. A profile was lodged with supplier, based around a specification of a range of parameters, notably class mark ranges (for mathematics, in this instance), readership level,

[1] Since the time of writing the UML has joined the CLA as a development partner in a national scheme to pool digitised sections.

year of publication, and of course budget. To make things easy the plan is dependent on the supplier having an accurate file of the library's holdings. The supplier receives full details updated every month, so the onus for 'checking against holdings' is on the supplier's systems not an overworked library assistant. This has worked very smoothly for both catch-up orders and current publishing, and we have seen pleasing usage for these non-textbook monographs – but the topic of usage later will be returned to later. The obvious disadvantage is that this approach is fairly undifferentiated, although the supplier's profile-building tool does allow a high degree of specificity in certain aspects.

Profiling for purchase 2: keyword-driven methods

A second, more ambitious and innovative approach involved drawing up profiles for purchase based on 'research keyword' methodologies: that is, keywords descriptive of Manchester's published research outputs. Two methodologies were attempted.

2.1 Extract keywords programmatically from Scopus. It was anticipated that this would be most suitable for subjects well indexed by Scopus, such as Mathematics, rather than History of Art, for example. The hypothesis here was that published research outputs would reliably represent the research interests of our staff, and that keywords describing them could be used to drive purchase of further research outputs from other institutions. The plan was to extract keywords programmatically from indexed University of Manchester publications, sort, count, and rank them in a list. These would be used as search keywords, filtered against the supplier's subject bands to ensure relevance of terms, and a list of suggested titles for purchase generated.

Statistical methods must be used with caution, however, and in our inexperience some unexpected problems surprised us. The ranked list of keywords was taken to the head of research for mathematics for a relevance check, and he pointed out that a statistical skewing had been introduced by a very productive but less significant member of staff, which made the keywords less than fully representative of the School's research activities. An alternative method had to be found.

2.2 Harvest keywords manually from web pages. The second iteration involved more manual intervention, in the form of gathering key descriptors from a School's self-presentation on their research pages, which it was assumed had been through some process of editorial/marketing review. This approach was tried with Electrical and Electronic Engineering. The application would then be a simple supplier-side search plus a few parameters (as with the approval plan); with embedded orders being created by the supplier. This was a small but significant feature: in all we do, we are trying to transfer as much processing load to the supplier as possible. However, the keywords we identified by this method were also found to be less than fully representative; an explanation may be that some academics may not place a high priority on keeping their web pages up-to-date.

Profiling for purchase 3: blended approaches

A third approach was more successful. This was a blended approach, a 'semi-automated' collection development if you will, aiming to utilise structured data and automated processes in combination with liaison at a human level to apply academic expertise. The relevant 'academic engagement' librarian worked with a focus group of academics from a discipline area to specify a simple 'profile' based on the most relevant monographic series from the most reputable publishers. For the discipline of Classics & Ancient History the series identified were from Oxford University Press and Cambridge University Press. Interestingly, the same publishers were chosen for the very different discipline of Physics. Again, the aim was to transfer processing workloads from librarian to publisher: rather than manually attempting to compile a list of what titles have appeared in a monographic series (which can be challenging, depending on the quality of your cataloguing, and indeed your data structure, as these series typically change their titles over the years in small but significant ways), a request was sent to the publishers to provide a comprehensive list of titles in the series, with all-format ISBNs and prices. It was crucial to specify that the data be supplied in tabular format, as PDFs are not amenable to data processing or linkage. This procedure allowed us to do gap-filling catch-up orders, and – to a certain extent – create pre-orders for forthcoming titles.

Descriptive collection profiles

To test our second hypothesis, that descriptive profiles of our collections could be dynamically linked to profiles of academic discipline areas, those descriptive profiles first had to be created. This involved using two main tools: COPAC's Collaborative Collection Management (CCM) tool for our printed collections, and dashboards developed in-house using Alma analytics. (For a good description of CCM in use see case study 2 in Showers (ed.) 2015, *Library Analytics and Metrics: Using data to drive decisions and services*.) Six disciplines were chosen and delimited by Dewey range(s). The choices were driven primarily by the expertise of the staff involved, but included a spread of subjects with a concentration Humanities disciplines, as turnstile data indicates these are the heaviest users of the physical library collections. On a small scale, the tool proved useful in the work with monographic series, in identifying the ubiquity or the distinctiveness of the series under consideration. At the level of collection, however, the limits of the tool were soon exposed, particularly in dealing with non-ISBN material. Batches of records for the six discipline areas were uploaded to the CCM tool for benchmarking. Good workflows were identified and much data returned, but it gradually emerged that the skills to analyse and synthesise this and other data into descriptive profiles were lacking in the project team. Data analysis should have told an interesting story about each collection, revealing what percentage of this collection is unique, or almost unique to Manchester, for example. From this one could have asked questions about the reasons: is the

material very expensive, very specialised, or very bad (poor scholarship, out of date) perhaps? Data has been gathered, but a full interrogation of it will not be possible until a follow-on project.

When our analysis is complete we should have a set of descriptive profiles which are benchmarked (compared to other (RL)UK libraries). Beyond benchmarking, the profile needs to adequately describe the depth and extent of the collection itself. Jakubs notes the afterlife even today of the 1980s Conspectus (Jakubs 2015). Not unrelated to that model, I think, is the categorisation that the University of Leeds Library has come up with, of 'heritage', 'legacy', 'self-renewing' and 'finite' collections (University of Leeds 2013). We aim to apply this categorisation too, which can subsequently inform decisions about collection development, as well as stock management and weeding. A question also arises about the label 'world class' collections. At present there is no consensus on which kind of statistics (on uniqueness, size, currency or chronological reach, etc.) would contribute to a categorisation of a collection as being 'world class'. If there is a requirement that the collection be multi-lingual there is a tension given the monolingualism of the current generation of scholars in the UK. Since COPAC can only tell us about the UK landscape, it may not be possible to make a 'world class' claim on the basis of fully transparent, comparable data without a global analytics.

Discipline profiles

The second hypothesis suggested that such descriptive collection profiles would need to be linked to a further set of academic profiles in order to drive collection development. Academic activities of research, teaching and learning have of course always driven collection development, but whereas previously this happened through a combination of subject specialists and faculty involvement, we are now needing to do it more programmatically. We cannot decide how strategically important a collection is and how we want to treat it without information about our operating context, and the strategic priorities of the University. It was not in the scope of the project to do more than initial investigations, but even these proved useful. While data and statistics have a major role, it was also important to build dialogue and partnership with the University's Planning Support Office, in order to obtain the data in the first place. It was beneficial that the University's IT Services had a change of leadership at the very top, but crucial for the project in getting a better response to queries than historically had typically been the case was finding the right person locally, someone who had a remit to help and mediate requests to the relevant colleagues in the Data Warehouse. They were able to supply detailed statistics on numbers of Research-active staff, for example, indicating the relative size of departments. Statistics on all courses (academic plans) taught included module codes

(important for undergraduate reading lists, for example) and also the subject codes assigned to them, even for research degrees. This kind of structured data offers powerful potential for linking and combining with other information. We also discovered how important it is to find the key link person, and to nurture that relationship, learning to understand and use their terminology in order to make requests in such a way that the information you receive is what you actually needed.

Supporting technology

As well as exploiting the University's data warehouse, the project has made heavy use of the analytics function in our library management system to really engage with our data, and make it work for us. The University of Manchester Library was an early adopter of the ExLibris product Alma, which promised extensive analytical functionality, although some of this is only now being delivered. Some of this power was harnessed by a newish colleague, who – significantly perhaps – had a background in mathematics and computing, not in libraries. She was able to build interactive statistical dashboards which have been useful to many library functions, but absolutely crucial to this project. There was, however, a limitation to the effectiveness of these tools, which will be explored below.

Combining data sources: the BSC for subscribed resources

Alma Analytics was once source of data to be fed into a posited Balanced Scorecard (BSC) for subscribed resources. The BSC is an established business tool, for evaluating success of a business from four perspectives, generally Customer or user / Internal or staff / Innovation & learning/ Financial (Kaplan & Norton 1992). The Manchester project attempted to put the subscribed resource (be it journal or database) as the focus of a balanced scorecard, and included a specific University of Manchester perspective (is it a journal that Manchester academics cite or contribute to, for example; does it support a key research area like graphene?). This adds an important element of local weighting which is vital for an academic library aiming to work in supportive partnership with faculty. Each perspective for evaluation will be populated by a range of metrics: over 80 desirable metrics were gathered in the initial phase, but this proved impossibly complicated, so have been drastically reduced. They include elements such as: percentage price rise; cost per use; platform preferences of users; past history of dealing with the supplier; quality of metadata supplied; annual updates of material etc. As well as Alma Analytics, other sources will be external data feeds such as JUSP (journal usage stats portal) and reports from our citation analysis service.

The tool could potentially be used to support decision-making both at the time of considering a new subscription, and as an evaluative framework at renewal time. This complex work was not finished during the lifetime of the project, but a prototype was built. The first steps moved from a flat spreadsheet to a system nicknamed 'Robyn the Renewals Robot', which was an Access database

doing a fairly limited mash-up. The 'customer' perspective only has one metric at moment (cost per student) and there is a simple pass/fail test with a threshold of a specific percentage rise and a specific cost per use. More development work will be required in the follow-on project, but the tool is in use as a supporting element in subscription renewal negotiations. Anecdotal evidence is already suggesting that the message is reaching vendors that libraries are resistant to over-inflationary subscription price rises and we really have developed a cold, calculating method to support our renewal decisions, so we are less open to rhetorical persuasion. Some suppliers appear to have decided to stop the annual bartering about percentage points and adhere to our preferred maximum increase.

Project evaluation

The stated aim of the Collection Development and Profiling project was to develop a range of innovative and efficient collection development methodologies, to be data-driven but strategically aligned with the University's priorities for research, teaching & learning as well as our role as a National Research Library (NRL). The project set out to use innovative methods to ensure that we continue to build our collections for the future, and aimed to push the boundaries of automation but maintain a sense of academic leading, which by necessity would involve understanding our collections and our users better. In order to achieve this, the project sought to identify, source, link, and analyse the appropriate data about collections and users. Thoroughly data-driven in its inception, the project's strengths and weaknesses as it progressed all hinged on human beings' attitude to data.

- Some 'quick wins' were achieved through the approval plans and back runs described under Profiling for purchase 1. Both methods proved to be very efficient, thanks to the active co-operation of our supply partners. More innovative ideas required an iterative approach, however, and a method of profiling research activity using keywords, and using these to drive book purchasing, was modified but found to be insufficiently robust. A potential line of future development would be to move on to a more forward-looking method: leveraging the new Current Research Information System (CRIS) for collection development driven by planned, not published, research.
- In the blended approach a workflow for a hybrid methodology was developed, one which exploited structured data alongside traditional (but carefully focused) academic liaison. An appropriate constitution for a focus group was established, to assess and identify some core monographic series to support the research of the whole discipline. Much of the associated processing work was effectively out-sourced. This model is an excellent example of a data-

driven approach which maintains clear academic leading, one which has already proven straightforward to implement and can be applied very easily to other subject areas.

- The project tested and refined a number of individual techniques that can contribute to a collection development framework. It also considered some principles about what constitutes value in a resource and how to measure it, and sourced new data about the University's teaching and research activities. A wide range of analytical reports was created to help us understand our collections and their usage, and tantalising new potential developments were identified. What is needed to implement them in a coherent way is a new, data-driven, collection development **strategy**, aligned with new developments in collection management strategy.
- In looking at descriptive profiling, some useful applications of tools were documented and a large amount of data was gathered. Where the process stalled was at the analysis and synthesis of that data. Paradoxically perhaps this represents one of the most useful project outputs: we have demonstrable evidence both that the requisite data sources can be accessed, and that useful analytical tools exist or have been created by the project, but that most University of Manchester Library staff do not currently have the skills or understanding of data to make used of them.
- In a work package 'Analytics & Value for Money' a comprehensive suite of business intelligence reports and dashboards were built using Alma Analytics, which meet some of the management information needs of the Library with regard to the value, content and usage of our collections. In support of these, an Access database was created to manage requests coming in for analytical reports, and a catalogue describing all the reports available was distributed at various team meetings, where the Data Analysis Assistant demonstrated the dashboards. Despite considerable internal dissemination, almost no use has been made of these dashboards since that person moved to another role. One might speculate on the reasons for this being that other colleagues did not have a use for them, or could not imagine their application; anecdotal evidence suggests, however, that many colleagues viewed themselves as not competent to use them as they were not 'data analysts'.

Discussion

People's lack of confidence is one limiting factor in data-driven approaches; another (widely documented) problem is the quality of the data itself (Morrisey 2010). Taken together this means we must concede that there are limits to what can be done with data. Nonetheless, the project achieved some concrete outputs as well as useful insights for future work in this direction.

Lessons learned

1. Tools are no good if you don't have the skills to use them. There was a widespread lack of data fluency, at a very basic level (inability to use Excel to create graphs) and very few staff outside IT support who really understood relational databases. For a project which was conceived from the start as a data-driven enterprise, this proved very debilitating.
2. There was a widespread inability (unwillingness? timidity?) to move beyond data gathering to actual analysis of data. Colleagues who have never been required to think in an analytical way will not suddenly become intellectually curious people keen to dig into data in the lifetime of a work package.
3. The project aimed for grass-roots involvement, but by widening the group beyond the usual suspects the project team comprised a group who in some cases lacked awareness of their roles (even work package leads) which necessitated additional training and support from the Project office and the Project Manager.
4. Involvement with such an innovative and experimental project effectively represented developmental opportunities to colleagues; not all of them rose to the challenge with equal enthusiasm or ability.
5. A data-driven approach will always be constrained by the quality and availability of data and meta-data. This has been a constant theme across all work packages of the project.

Summary

The 'Collection development and profiling' project aimed to do pioneering work in support of Manchester's strategic restructure of a department, starting to engineer a fundamental shift in the basis for our collection development from a system based on subject librarians to a new, data-intensive, analytical approach which exploits a variety of data and relationships to support decision-making. Such a shift cannot happen overnight. The project started to build a framework for how this might be achieved more fully, and surfaced some unanticipated hindrances which can now be dealt with. For the 'Manchester model', data driven approaches represent a vitally important opportunity for collection development (not to mention other areas), but not all colleagues are equally enthusiastic about novel methodologies. Skills gaps and inappropriate, uncertain staffing structures have been significant barriers to the project – but these are not insurmountable. More training, mixed teams, and further cultural change will be required. Manchester's recently launched comprehensive 'Digital First' programme will certainly make a significant contribution here.

Next steps

Research keyword methodologies can be rethought, to look forward rather than back. The CRIS can accommodate subject taxonomies (for example) associated with research applications, which could in turn be output to drive collection development which would support University of Manchester research activity in an extremely timely and efficient manner, potentially ensuring that resources were available to researchers before they had thought to ask for them. Metadata validation will be essential here.

Data normalisation and understanding *usage* (particularly of electronic resources) in the environmental context will be the next challenge. Work is underway with colleagues in Citation Services to understand aspects of this better. The aim is to get beyond bald usage statistics and 'cost per use' calculations, and grapple with questions of what constitutes good usage. This must include an assessment of what is typical usage for (say) a journal of its type, in its discipline area, and an understanding of how particular kinds of usage can demonstrate the value of resources, whether through simple downloads, multiple page views, consistent loan patterns. Looking beyond Manchester, it is possible to take a much wider view. The UK's emerging national monograph strategy and Jisc's Monograph Solutions work may provide tools that offer nationally aggregated usage and collection data. Better infrastructure will be required, as well as widespread sectoral agreement on changing workflows (commitment to regular WorldCat updates, for example). Libraries could be benchmarking not in order to compete but to learn from each other, and collaborate in making a distributed national research collection available.

In Manchester staffing structures are still evolving but the potential now exists for a more co-ordinated and strategic approach, which makes best use of staff aptitudes and interests. Appointing staff with strong data analysis skills is a generally recognised priority, and will enable us to produce more satisfactory collection profiles, and develop discipline profiles to complement them. The mechanism by which we will link the two will be a mapping of academic subject taxonomies to library classification schemes. A new project has been defined to build on our successful proof of concept in mapping JACS (Joint Academic Coding System) to Dewey Decimal Classification. Establishing this link between academic activity and library collections will both drive collection development and support the creation of compelling narratives of how collections support academic endeavours.

Conclusions

Data fluency is a major issue for librarians now. Crucial to the success of this project has been the data fluency of particular individuals involved. They have helped solve problems and built bespoke

tools which should make it easy to access the data needed for collection development – but the tools have been underutilised by colleagues who lacked the confidence or skills, or didn't perceive it to be their role to use them. Data fluency doesn't just mean the skills of being able to work with statistics or use a relational database – although this is a key part of it. It is an easy awareness of when it is appropriate to use Excel rather than Word, or Access rather than Excel, as well as the skills to exploit those programmes fully. What is needed is an appreciation of how data can be used and manipulated, how it needs to be structured in order to facilitate computational processing; and an enquiring, imaginative mind that can ask questions of data, allow one question to prompt another, and imagine what stories the data might help us to tell about our collections and the people who use them. What is needed is a real curiosity about what data can do for collections and their users, and a willingness to engage with it – in short, a research librarian needs to have the mind of a researcher. In the short term at least we need mixed teams with different areas of expertise, just as is typical with Digital Humanities research projects.

In a university context the institutional priorities may change, and research interests shift with the movement of staff, so collection development needs to be flexible enough to change with them – as well as protecting those collections deemed to be of long-term strategic importance. Dynamic, data-driven approaches to content curation can make a real contribution to achieving Ranganathan's goal of 'every book its reader, every reader his/her a book'.

References

Jakubs, D.L., 2015. Trust me: the keys to success in cooperative collections ventures. *Library Management*, 36(8/9), pp.653–662. Available at: http://dx.doi.org/10.1108/LM-08-2015-0058.

Kaplan, R. & Norton, D., 1992. The Balanced Scorecard - Measures That Drive Performance. *Harvard Business Review*, 70(1), pp.1–71.

Morrisey, L., 2010. Data-Driven Decision Making in Electronic Collection Development. *Journal of Library Administration*, 50(3), pp.283–290. Available at: http://www.tandfonline.com/doi/abs/10.1080/01930821003635010.

University of Leeds, 2013. *Collections strategy for Leeds University Library.* Available at: https://library.leeds.ac.uk/downloads/file/212/collections_strategy.