

STAFF PAPER



# Rethinking the catalogue

Paper delivered to the Innovative Ideas Forum, National Library of Australia, 19th April 2007

**Alison Dellit**

**Kent Fitch**

## Stepping into the future

The last decade has been a time of profound changes in our society's relationship with information. What parents once spent on an encyclopaedia set, they now use to pay for broadband and wireless connections. Instead of buying bus timetables, we bookmark public transport sites. The high speed highway of the internet provides access to an astonishing array of argument-ending trivia - the complete biography of any actor who ever appeared on *Buffy*, and the local time across the world from Rome to Burundi. Things we never knew we wanted to know are at our finger tips, and consequently, what we expect to be available is vastly different.

The most powerful uses of the internet, however, are still being unlocked, as key texts are digitised and made freely available: from Shakespeare, to Herodotus, to Captain Cook's diary. Much modern scholarship is now available online, although most requires payment to get access. The fact that there are more gems of human knowledge that you can't find online than that you can doesn't invalidate the significance of what is there or the pace at which it is going up. The plentiful supply of information raises unknown problems (how do you find what you need amid the deluge?) and, most significantly, is changing what we mean by information and knowledge. Interaction, collaboration and engagement have become information issues, as people demand not only to find information resources, but to add to them, engage with them and build communities out of them.

What does all this mean for libraries? For decades, libraries were the places that ensured that knowledge was not the domain of a few, but accessible to the many. We preserved and selected and maintained collections. We have kept information safe, and provided pathways to it. But now we can't keep up - so much information to collect, as we select, and describe, and preserve a trickle, there is a flood we are missing. And people and corporations develop their own pathways through, which are easier to use than ours. However, we believe that library collections, and that of the National Library and Australian libraries in particular, enormously enrich the information landscape of this country. Libraries have a responsibility to take advantage of new technology, new models and new patterns to make our collections easier to find, our systems fun to use and our digital spaces as much a centre of our communities as our physical spaces are.

In this paper we propose four basic strategies based around re-imagining library catalogues, the arterial systems of our libraries. We discuss some of the thinking

that the National Library has done, and some of the steps that we are now taking.

Be warned: what follows is not all "build on what we've got" and "embrace and extend"; as with most revolutionary endeavours, plenty of creative destruction is required.

## Four strategies for modern catalogues

*"In some ways we have end-to-end integrated library systems where the ends are in the wrong places. At one end, we have a catalogue interface which is unconnected to popular user discovery environments or workflows. It is often a somewhat flat experience with low gravitational pull in the crowded network information space. We expect people to discover the catalogue before they can discover what is in (part of) the collection. And this points to the issue at the other end: the 'fulfilment' options open out onto only a part of the universe of materials which is available to the user: that local catalogued collection."*

- Lorcan Dempsey, "The Library Catalogue in the New Discovery Environment: Some Thoughts", Ariadne, July 2006

The catalogue is the arterial system of the library. We use it, we expect our users to use it, and its existence and structure is the difference between a collection and just a big accumulation of data. The goal of these strategies is to re-equip the library catalogue to meet our aims of providing access to information - from ours and others collections - into the future.

1. **Rethinking cataloguing: describe better and cheaper.** The data that we collect about resources determines to a large extent what our users can discover through the catalogue. At the moment, resource description is time-consuming (and hence expensive) and requires interpretation from a trained professional to understand. A rethink of cataloguing standards and better systems can help fix this, but a shift in attitudes to resource description is also required.
2. **Creating an interactive space.** Knowledge is created through conversation. We need to make our online spaces as interactive as possible, facilitating engagement with our collections, better decision-making by our patrons, and the creation of new knowledge based on our collections.

3. **Unify information resources.** Neither a single library nor even a large national union catalogues have enough "gravitational pull" to attract searchers. The solution is to augment the catalogue with a very large and highly dynamic set of metadata managed by a large number of separate entities external to the library.
4. **Improve access.** Library systems are notoriously unfriendly, seemingly designed for the expert librarian user rather than the general public. We now know how to transform these systems and must plan for the next step: the embedding of library interfaces in external systems.

## Describe better and cheaper

*"I'm going to too many places to work out what I have to do, and I'm going to so many places I forget what I have to look up where."*

- National Library cataloguer, March 2007.

Dewey cataloguing frequently urge cataloguers to "learn to think like Melvil Dewey". This is not because librarians inherently admire Dewey, but because his system is a standard through which our collections are interpreted. Highly skilled cataloguers can glance at a 12-long string of numbers and immediately know a great deal about the resource being described. An experienced reference librarian could give a patron looking for a book the best Dewey sequence to start browsing from. Library of Congress subject headings work in a similar way, as experienced cataloguers learn the "language" of LCSH, and reference librarians learn how to suggest terms to users searching our catalogues. Librarians are not so much gatekeepers of our resources as they are translators, and Dewey and LCSH are the codes.

There are good reasons these standards have developed. Human beings think in vastly different ways to each other. No two people are likely to describe the same resource in the same way, nor will they try to search for it with the same keywords.<sup>1</sup> A thesaurus with clear rules helps override individual preference, and group like material with like material, and maps the relationships between topics and corresponding material. For years, we have known that this causes problems for users, who don't want to use a translator or don't "think like Melvil", and these problems are getting worse. The good news is that the development of

---

<sup>1</sup> Bates, Marcia J. "[Rethinking Subject Cataloging in the Online Environment.](http://www.gseis.ucla.edu/faculty/bates/rethinkingcataloging.html)" Library Resources & Technical Services 33 (October 1989): 400-412, online at <http://www.gseis.ucla.edu/faculty/bates/rethinkingcataloging.html>

technologies like tagging may offer us a way to both have an underlying thesaurus and offer multiple points of entry for a search. But this isn't just a problem for catalogue users; it is not sustainable in the long term to expect all cataloguers to be experts in Dewey and LCSH either.

Cataloguing is hard. The more vast and complex the system of authorities, the harder it gets to develop proficiency in using it. As cataloguing has grown, so too have the number and complexities of standards we expect cataloguers to know well. A cataloguer at the National Library would learn to refer to the Anglo-American Cataloguing Rules; the Dewey Decimal Classification system; a set of Library of Congress Subject Heading H-Lists; the Libraries Australia subject and name authorities; Library of Congress rule interpretations; MARC 21 format rules; at least two in-house cataloguing guides; a guide to the Voyager cataloguing module and various other guides such as the list of geographic codes. Take a walk around a cataloguing section sometime - desks groan under the weight of texts, and cataloguers' internet browsers are cluttered with bookmarks - and several of those go to aggregate sites such as "Cataloguer's Desktop" and "Classification Web".

It is unsurprising that new cataloguers are often told "it takes two years to make a good cataloguer" - and that time is on top of formal study. A project undertaken by the National Library's Bibliographic Standards and Strategy branch to investigate subject cataloguing workflows bears out the "two year" figure. Amongst the 15 cataloguers used as a focus group, there is a clear accuracy and speed difference between those in the first two years in the job, and those with several years' experience.

It is not just the information world that is changing, so are the patterns of work. In 1959, average job retention in Australia was 15 years<sup>2</sup>. In 2006, it was four years. According to the Australian Bureau of Statistics<sup>3</sup>, one in four of those aged between 20 and 24 change jobs in any given year. Whereas jobs once lasted a lifetime, now they rarely do, and expectations have changed as a result. All of us now expect variety in our working life, and a career progression. For the complexity of the work, cataloguing is not well paid, and doesn't get necessarily get better paid as you get better at it. How many of those starting cataloguing now intend to be cataloguing in two years, five years? We are already seeing the impact of these changes: how many institutions still keep new workers on copy cataloguing for an extended period before moving them on to original

---

<sup>2</sup> McCrindle Research : Jobs & Generations, 2006 <http://www.mccrindle.com.au/fastfacts.htm>

<sup>3</sup> ABS Labour Mobility, Cat 6209.0

[http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/6209.0Feb%202006%20\(Reissue\)?OpenDocument](http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/6209.0Feb%202006%20(Reissue)?OpenDocument)

cataloguing?

Another driver for change is the growing amount of digital material which increases the pressure to describe and analyse material quicker. In 2004-2005 at the National Library, the amount of electronic format material being voluntarily deposited increased by 100%, and we expect this trend to continue. If we want to keep up, we need to find ways of cataloguing more efficiently. If we continue to rely on human knowledge of the detail of many, often rapidly changing, highly complex standards, then we will start to see a decline in the quality of our data. We need to find new ways of facilitating data collection that will preserve accuracy and increase speed. This paper is not arguing in favour of ditching cataloguing standards, particularly in the area of subject description. To do so without any way to substitute for the data would simply mean that users searching our systems will only find items in which their search terms appear in the title, author or other descriptive data (although we accept that widespread digitisation will offer other solutions in the future). The challenge that we face in libraries is to work out how to use standards in an efficient and effective manner, without requiring years of study and experience in order to get it right.

### **Better standards**

Ultimately, this will involve a re-examination of cataloguing standards to determine what is necessary to keep, what we can discard, and what we need to add. This discussion does need to incorporate a "blue sky" approach, in which we are clear on what would be the best solution for our current needs, if we weren't tied to our existing set of standards. Before we can get to a new solution, of course, we need to acknowledge the realities of working with legacy data, standards and systems. The discussion about fixing standards is not well developed internationally, and is not the focus of this paper: we certainly don't claim to have answers. There are three points we want to make.

Firstly, the National Library is participating in the Resource Description and Access process to rework AACR2. The RDA process is based on an understanding that fundamental changes are needed to that standard - hence the name change - to make it simpler, easier to use and more relevant to digitised material. In deciding what to include and what to throw out, the steering committee has used an FRBR access model - so what facilitates access stays, what doesn't, goes. We are yet to see the results of this process, and there has been discussion about the progress, but we need to invest in this process to make it as effective as possible.

Secondly, we need to re-examine how we use transmission or exchange standards,

like MARC, Dublin Core or MODS. MARC in particular has become so entwined with other standards, particularly AACR/RDA (and now does contain elements of a content and data standard), that it can be hard to tell where one ends and the other begins. An assessment of these standards needs to examine whether they need to be a content standard, or just a transmission standard. If we don't store our records as MARC, should we enter them in that format? Can we have simple form for entering the data, and then use our computer systems to turn that into MARC or Dublin Core as needed? This discussion is only just beginning, and these are some very preliminary thoughts, but the future involves assessing not just what we are used to, but what is useful for us.

Finally, we do need to consider how much we need to catalogue, and how to provide access to, and control of, material that we do not catalogue. Google Books and Google Scholar provide access to lots of resources without having ever catalogued them, and this is an approach we need to consider. For example, as part of the Newspaper Digitisation Project, the National Library and our partners are intending to digitise between 8 and 10 million newspapers articles a year. We do not have the resources to catalogue each of these articles individually, so we have had to find other ways to provide access. In a world where we can capture the whole of the Australian domain of the internet, how much internet-based material should we selectively catalogue? Most libraries, including the National Library, do not catalogue all our material already, of course. The need to find other ways to control and describe material will only grow. Moving into the future, we will need standards that are flexible enough to provide for a range of approaches.

### **Better cataloguing systems**

The good news is that we don't have to wait until we have better standards to improve cataloguing. Just as library catalogue systems aren't very user-friendly, neither are cataloguing tools. It seems sometimes that as cataloguers we base our professional pride on being able to find what we want in a forest of unfriendly systems. If we can teach the systems to do the work, however, teach the systems to be the interpreters, rather than the librarians, then catalogue records can be accurately and efficiently created by people without years of formal training.

As part of the work of the Bibliographic Standards and Strategy branch, Alison carried out an observation test of 13 cataloguers assigning Library of Congress Subject Headings and Dewey Decimal Classification numbers. The results indicated that much of the time taken involved navigating different systems to

ensure a correctly structured heading or number, rather than determining what the item is about, or how to best make it accessible.

For example, one cataloguer, who is an excellent librarian, but still in her first year of cataloguing, had an item to catalogue that had the title *Hotrods*. She was in little doubt what the item was about, but finding the appropriate LCSH heading was another matter. Looking for similar items did not produce clear results, in part because some imported records had incorrect terms assigned. As the item had the term spelled in a single word, and the Library of Congress regards it as two, she didn't find the right term using the limited search capability on Classification Web. Attempting to follow the "broader" and "narrower" terms in Classification Web was difficult because the listing went on for pages. In the end, she resorted to constructing headings using free-floating subdivisions, which meant checking more rules in several different systems. Even when she had a correct heading, the fact that she had to retype it in meant that she needed to keep checking she hadn't made a mistake. In the end, the process took more than 20 minutes (without the DDC number), and, as it happens, she hadn't found the most appropriate heading.

Now I know that many cataloguers reading this would be thinking "she could have done x" or "she should have done y", which is undoubtedly true. Years of perfecting cataloguing skills clearly makes a difference to speed and accuracy of assigning headings. But there is no need for it to take so long.

Imagine if she had been able to go to one place, to search for headings that allowed her to see:

- Whether they validated against the LC Subject Authorities
- What other items had been catalogued with those terms
- What other headings had been used on items catalogued with those terms
- Alternative spellings for what she had typed in.

Imagine even that this tool could automatically drop correctly formatted headings in to the appropriate fields in the cataloguing module. Imagine if we used the data from our Web Dewey subscriptions to suggest statistically and editorially mapped DDC numbers for each classification. We could easily have saved this cataloguer 15 or more minutes in this process, and while it might not be so great a difference in other cases, it would speed up almost all cataloguing work, and make it less frustrating. This is the sort of tool we are hoping to develop at the National Library. To be clear, it does not take away from the core professional skills of a cataloguer - determining "aboutness" and thinking about access points. It just simplifies processes that the cataloguer is carrying out anyway, and makes the job of memorising LCSH rules the computer's, not the human's.

This approach could be broadened to other systems. We could build cataloguing modules that ask for a title, not 245\$a and \$b field, or ask how many pages something has, not for a formatted 300 field, and where it was published, not for a GAC code. We could build in spell checkers, MARC validators and added punctuation. Many of these features are already in existence, but haven't really been implemented - in part, because most cataloguers are still skilled enough in speaking MARC that they aren't necessary. And there are implications: not all cataloguers will be experts in standards. Someone might catalogue for 2 years and not know what a 100 field was. Those who do know may not be cataloguing themselves, but instead involved in quality control, training and standards development. But we can ensure high accuracy levels even with transient staff, faster output and hopefully re-focus our profession on what it does best. We understand how to describe, classify and organise, how to design and enforce thesauri and how to provide access to a wide range of material. There is great depth of intellectual skill involved in collection development that we will never be able to automate, and these are the skills, not the in-depth familiarity with particular vocabulary or classification schemas, that the future needs from us.

## Creating an interactive space

*"[My library should] improve its website more - I like the catalogue, but if it could reference some sort of rating system, it would be even better - I was looking at a new author today who has many books, and I had to go to an internet computer, check on Amazon, and see what books were most highly recommended and then go back to the catalogue to see if they were available"*

- 15-year old high school student responding to OCLC survey, 2005

Human knowledge is not created in isolation, but through engagement and interaction. By providing access to scholarly journals, books and other items, libraries have always facilitated the development of new knowledge. In an age of electronic communication, and easy digital manipulation, this process has been accelerated. Interaction is the basis for Web 2.0, and the knowledge it is unleashing is impressive. In many professions, including the library profession, the blogosphere facilitates a professional discussion that is more informal and fast moving than the discussion through more scholarly journals. Wikis facilitate the cooperative authoring of documents - Wikipedia is the most well-known public wiki, and despite its flaws is a tremendous achievement of collective knowledge. Like many institutions, the National Library now uses wikis to develop

documents, and facilitate communication between staff over specific questions.

It's not just scholarly knowledge that is being created. Creative media is developed collaboratively through shared music, artwork, pictures and software. Sites such as Flickr and YouTube have become not only places where people show off their own original products, but also where they can modify and build on existing work to create something new. Participants in the open source software movement are old hands at this, with vast interactive spaces facilitating the work of literally tens of thousands of people in developing, refining and testing thousands of software projects. Sites such as TripAdvisor, TV.com, and in the US, RateMyProfessor.com and hundreds of others, facilitate not just ratings and rankings, but also discussion, about how to make choices about holidays, television programs or who to study under.

So what does all this mean for libraries? For a start, it means our patrons are likely not only want to "get" but the "create", and to build upon what others have created. They want features such as reviews, as the above respondent says, that enable them to learn more about our resources and also to start discussions about those resources. Amazon's provision of reviews has become not only a place to learn more about their books, but also to participate in some frivolous, some serious, discussions about them. User added content will not immediately replace subject designation and description but it can certainly enhance it, and enrich our catalogues with much more than we can afford to record about an item.

One of the most powerful aspects of crowdsourcing is tagging. The STEVE project<sup>4</sup>, coming out of American art museums, gives an indication of how powerful tagging technology can be in providing more access points for users. By encouraging users to simply type in whatever terms they identify with a work, we develop not only plain English descriptions, but also pathways we had never thought of. The success of LibraryThing – which now has more than 16 million tags applied to its bibliographic records – indicates that the approach is viable for books as well as more visual material. Tagging not only adds descriptive terms to individual items, but may also allow us to track similar words as well. By knowing, for example, that items tagged with swimming are often also tagged bathing, the system can guess that these terms are related, and people searching for one may be interested in the other. LibraryThing also allows users to draw connections between tags – to explain that Chick Lit is the same thing as Chic Lit, or myth and mythology are functionally equivalent terms.

---

<sup>4</sup> Steve – The Art Museum Social Tagging Project: <http://www.steve.museum/>

We should not only allow users to add data locally, but also think about what other sources of data we can use, for example, linking in with Library Thing to use their impressive collection of tags. What if a user could move seamlessly between Libraries Australia and LibraryThing, comparing their personal collection (or borrowing history) to others through LibraryThing, then moving to borrow recommended other books direct from their local library? LibraryThing founder, Tim Spalding, has floated the idea of a "Tag Consortium" which would allow the collective maintenance of a large tag pool for bibliographic resources. LibraryThing currently maintains a huge tag base which would be invaluable seeding material for a library catalogue.

We already create subject guides on topics that are the subject of frequent questions. We should put more energy into enhancing these, and enabling users to create their own. Our patrons should not only be able to search for individual items on "Australian federation" but also dynamic lists of resources, overseen by, but not exclusively added to by, reference librarians. If we can make our systems more intuitive, we should be able to redirect reference resources from answering questions about the system, or teaching classes in Boolean search techniques, into added-value items such as resource guides and encouraging user participation.

This discussion leads to a central point: not all our users will come to the physical library building. That is particularly true of a national institution, but also of other libraries. Our digital spaces need to be as interactive, welcoming and pleasurable to hang out in as our physical spaces. This needs to go well beyond the catalogue, but it should also impact on the catalogue. Subject guides, and tags, are a way of providing basic reference help in a generic way - so we can tag items as suitable for "beginners" or "domain experts", for example, making it easier for someone looking for a guide on fence building, or information on the Great Fire of London, to work out what is appropriate. Subject guides in particular would make the catalogue a preferred destination for many users (preferred over Google, that is) because it would offer them an authoritative introduction to the topic and links to follow for more information. Some of these links would take searchers to online resources such as Wikipedia and specialist web pages, but others would be to resources available from their library.

If users can identify themselves with a library or libraries, subject guides can be dynamically tailored in interesting ways:

- locally available resources can be highlighted
- local content can be incorporated

As an example of the second case, university lecturers may augment or author a subject guide and reading list for their students and give them the URL of this guide on the catalogue. Having this information in an open system is of benefit for the lecturer and the wider community because the lecturer benefits from the wider resources of the catalogue, can easily piggy-back on other subject guides and because the community benefits from the information being generally available and indexed by search engines, rather than hidden in proprietary e-learning systems behind firewalls.

Although subject guides could be commonly initiated by reference librarians, the system will evolve into a wiki, benefiting from the large base of contributors and backed by the bibliographic resources and ranking, clustering and tagging abilities of the catalogue.

This process is an example of what is becoming known as 'crowdsourcing' - understanding that if we can generate enough of our users to input into our spaces, then those who are constructive will vastly outweigh those who aren't. Of course, that means reaching a critical mass, and how we ensure that we encourage our users to participate is an issue that will require some thought.

Interaction is not just about providing people with more information to choose what items to get. It is also about allowing users to add to our collections, through annotating and adding to resources, and providing spaces for discussion about items in our collection, or research being created by our users.

Another key advantage of an interactive space is allowing our users to add and correct our metadata. We have limited knowledge of many of the items in our collections, particularly those that have never been published, and our community could add a great deal more. From photographs of unknown people, or unknown places, through to satirical cartoons with an uncertain target, or pamphlets we cannot date accurately. Allowing people to annotate resources also gives us the ability to incorporate that Merv Smith believes that this is his Great Aunt Agatha, who worked in a rural mission in the 1800s, or that these are pictures of the Anangu people during an annual ceremony.

Obviously, we will have to resolve how we check and rely on information, but even as annotated notes, such information enriches our collections. An example, a month ago during a Ask Now shift, Alison took a online query from Australian swimmer Ilsa Konrads, who had found some pictures of the 1960 National Swimming Championships in which several swimmers were mislabelled. During the conversation, she not only gave the correct names of the swimmers, but also talked briefly about her own memories of that event. Imagine if, instead of finding

an online service to log on to, she had been able to add her comments briefly under the picture - not only would it have been faster for her, but it would have provided a significant enrichment to the item.

The National Library has recognised the need to incorporate this data into our online spaces in the fifth outcome in our directions statement<sup>5</sup>. The Library has made some progress in meeting these directions, but we have also assessed that we have a way to go. For example, the Flickr partnership with PictureAustralia, where photos upload to Flickr are incorporated into PictureAustralia, has been a solid success, netting the National Library new items for our collection, and we have been able to use spaces within Flickr to have discussions about the project with users. We are currently working on developing an online collaborative space built around Australian folklore. This is an area the Library is keen to develop, and share and learn from other libraries experiences.

## Unify information resources

*“Users are often unaware that there are multiple discovery tools for the resources the library has to offer: the library catalogue, abstracting and indexing databases, the e-Scholarship Repository, various collections of digital library objects, archival collections, etc. As a result, they are frequently frustrated by their lack of success in finding what they seek. The few sophisticated researchers who are aware of the differences are justifiably unhappy with the need to search one “silo” at a time.*

*Users who are accustomed to Google expect to enter one search and retrieve information pulled together from across the information space and presented in a single ranked list. They want more than the ability to search multiple catalogues or multiple A&I databases simultaneously. They expect to search the full range of tools cited above or subsets the user wishes to select.”*

- University of California Libraries Bibliographic Services Task Force,  
"Rethinking How We Provide Bibliographic Services for the  
University of California: Final Report, December 2005"

---

<sup>5</sup> National Library of Australia Directions for 2006-2008: <http://www.nla.gov.au/library/directions.html>

Lorcan Dempsey has described the need for libraries to increase their "gravitational pull" by becoming places worth visiting to get information<sup>6</sup>. Whereas once seekers of information had little choice but to start at a library, they now begin at Google, Google Books, Google Scholar, Amazon, Wikipedia or other web sites with large "gravitational pull". They do so because it is more likely that they will find more information well presented and easily accessed at these sites, and find instant access to full text (not just metadata) and useful "value-added" information such as citations, reviews and related material. The material found may not always be the "best" there is, but numerous studies have demonstrated that for most people, "good enough and fast" is better than struggling to find and access the "best".

Libraries have the responsibility of making their "best" material not only easy to access but of integrating it with the material outside their collections.

Libraries are caught in a dilemma: they cannot reduce their niche to servicing "scholars" whilst not supporting the features scholars find vitally important, such as full text and citation linking; they cannot reduce their niche to servicing "less sophisticated" users without supporting the features they value, such as full text, related items, reviews and annotations.

The options for libraries are:

1. Abandon the scholarly and general bibliographic search and discovery space to Google, Amazon, LibraryThing etc. Provide information to aggregators and search engines as MARC records or crawlable web pages which will allow searchers to discover library resources.
2. Form viably-sized aggregations with other libraries and together partner with Google, Amazon, LibraryThing etc to enhance library resource discovery systems with the information available from these partners .

The main problem with the "abandon" option is that aggregators and search engines are not disinterested actors. Rather, they are commercial entities with shareholders, profit targets and performance bonuses. Their interests may often be partially aligned with the searching public, but they are not disinterested. They will inevitably arrange their systems and promote results which result in their corporate interests being advanced and these decisions will not always be the same that a purely disinterested broker would make.

---

<sup>6</sup> Lorcan Dempsey, "The Library Catalogue in the New Discovery Environment: Some Thoughts", *Ariadne*, July 2006 <http://www.ariadne.ac.uk/issue48/dempsey/>

A "viable sized aggregation" is one of sufficient "gravitational pull" to be of interest for searchers. It must make it particularly easy to discover and obtain local information resources and available electronic resources from anywhere whilst also exposing relevant information from all sources. The searcher has to reason along these lines:

"I could search Google, or I could search my library: If I search my library, not only will I see everything I'd see if I separately searched Google, Google Scholar, Google Books, Amazon, Wikipedia, Library Thing, Internet Archive's Text Archive and .. BUT ALSO I'll find easily obtainable local resources I wouldn't have found if I'd searched those other aggregations (and it will have extra "added value" [strategy 2] and be presented better [strategy 4]) "

The steps to making this happen are:

- Libraries must apply a "single business model" or "one stop shop" to their own collections. Searchers don't care to know that they must learn and use one user interface to search the map collection, another to search the manuscripts, another to search the photographs, another to search the journals and so on.
- Libraries must form mutually beneficial partnerships with large aggregators which will allow libraries to supplement their own search and discovery interfaces with the aggregator's data. That is, they must not just "aggregate supply" (in Dempsey's terms<sup>7</sup>) of library resources, but of all information resources.

For example, when a searcher submits a search to the library web site, the library's systems can search their own catalogue for matches, ranking and clustering them, and also:

- shows relevant journal articles (using Google Scholar), some of which will be available in the library either physically or electronically
- searches the full text of books (using Google Books, Amazon, IA Text Archive), some of which will be available in the library or as free text on the web
- searches tags (using LibraryThing) finding books, some of which will be available in the library or as free text

---

<sup>7</sup> Lorcan Dempsey's "[Libraries, logistics and the long tail](http://orweblog.oclc.org/archives/000949.html)" <http://orweblog.oclc.org/archives/000949.html>

- finds relevant Wikipedia articles and web pages

When showing the details of a found work, the library in partnership can not only show related material from library collections (using library-managed cataloguing and circulation data) but can also:

- show related books (using data from Google Books, Amazon, LibraryThing)
- show sample full text (from Google Books, Amazon, IA Text Archive)
- search the full text of the book showing hits (from Google Books, Amazon, IA Text Archive)
- show citations of and by (books and journal articles) (from Google Books, Google Scholar, Amazon)
- show reviews (from Amazon, LibraryThing)
- show tags (from LibraryThing, Amazon)

Such partnerships deliver these benefits:

- the library attains sufficient gravitational mass to be worth searching
- Google, Amazon LibraryThing are handed the searchers seeking to do what they can't do at the library: purchase books, purchase access, contribute reviews and tags; these referrals are the commercial lifeblood of the aggregators
- the searchers interact with a disinterested, trusted mediator which shows them free and commercial content and allows them to choose what they want

School libraries are conspicuously absent from the NBD. Yet school-aged children are large users of library services and, as the largest population segment typically able to borrow directly from multiple libraries, would benefit significantly from searching their school and local libraries as a subset of the NBD and from access to online resources available from the partnerships described above and demonstrated in a recent NBD prototype mock-up<sup>8</sup>.

## Improve access

*"I wish I had known that the solution for needing to teach our users how to search our catalogue was to create a system that didn't need to be taught and*

---

<sup>8</sup> Dellit and Fitch, NBD Prototype mock-up of integrated discovery. <http://l101.nla.gov.au/mock3.html>

*that we would spend years asking vendors for systems that solved our problems but did little to serve our users. I wish I had known that we would come to pay the price of our folly by seeing our users flock to commercial companies like Google and Amazon."*

- Roy Tennant<sup>9</sup>

*"Information literacy is also harmful because it encourages librarians to teach ways to deal with the complexity of information retrieval, rather than to try to reduce that complexity. That effect is probably not intentional or even conscious, but it is insidious. It is not uncommon for librarians to speak, for example, of the complexity of searching for journal articles as if that were a fact of nature. The only solution, from the information-literacy point of view, is to teach students the names of databases, the subjects and titles they include, and their unique search protocols - although all of those facts change constantly, ensuring that the information soon becomes obsolete, if it is not forgotten first. Almost any student could suggest a better alternative: that the library create systems that eliminate the need for instruction."*

- Stanley Wilder<sup>10</sup>

## **Ranking, Grouping and Clustering**

Users now expect results to be relevance ranked as a matter of course. Before the advent of Google there was widespread despair that only internet search engines based on humanly compiled directories (such as the initial Yahoo directory) would be useful and that this approach would not scale to the size of the burgeoning internet. Google changed that by including citation (or in-bound link) count to the previous ranking mechanism of search term density and positioning in the target documents.

An unranked search result which produces more than 10 results is annoying; an unranked search result which produces more than 100 results is virtually useless.

Unlike ranking based on the ordering of results on some common and fixed attribute (such as author, title or publication date), relevance ranking orders results based on the system's best guess of how relevant the results are based on the user's query. In general, the only source of information the system has is the user's query: what they entered in the search box and perhaps other input fields. In some cases the user may "self select" themselves into other categories by using a

---

<sup>9</sup> Roy Tennant, "What I Wish I Had Known", Library Journal, November 15, 2005, <http://www.libraryjournal.com/article/CA6282632.html>

<sup>10</sup> Stanley Wilder, Information Literacy Makes All the Wrong Assumptions, <http://www.owlnet.rice.edu/~comp300/documents/InformationLiteracy.pdf>

certain interface (Google Australia, simple search) or offer implied characteristics by being associated with a user profile (based on logon, cookies, IP address etc).

Effective relevance ranking is a domain specific problem. For a company accountant wishing to view outstanding invoices, the ranking may be based on factors such as the monetary value of the invoice and how long it is overdue, and perhaps the payment history or credit rating of the debtor. For a general web searching system, taking into consideration inbound-link counts and link text has been demonstrated as being an excellent approach.

Bibliographic systems are rich in relevance ranking possibilities. For example, given a search phrase such as "word1 word2", it would be reasonable to guess that:

- titles exactly equal to "word1 word2" where most relevant
- titles containing "word1 word2" as a phrase where very relevant
- titles containing "word2 word1" as a phrase where almost as relevant
- titles containing "word1" and "word2" separated by some other words where almost as relevant, perhaps with relevance dropping according to some function of the number of intervening words and perhaps slightly less relevant if "word2" appeared before "word1"
- "main" title (245\$a) is more relevant than series title, sub title, added entry title, translated title...
- shorter titles containing both words are more relevant than longer titles
- as above, but for authors rather than title; maybe authors are slightly less relevant (or maybe not)
- as above, but for subjects, but with subjects almost certainly less relevant
- as above, but for derived values from encoding MARC fields (such as the 008 audience, 008 various form codes, 048 \$a, decode Dewey and LC values etc)
- as above, but for notes, abstracts, table of contents, ISBN, ...
- if "word1" was a very unusual word and "word2" was a very common word, then if two records contained both but with the first record containing "word1" prominently (perhaps in a title) and the second record containing "word2" in a note, then the first record is more relevant than the second

It also seems reasonable to assume that other data could strongly influence ranking and presentation for most searches:

- popularity: the number of libraries holding the material, its circulation profile, its Amazon sales rank...

- availability: ease of getting the material from local libraries, 'foreign' libraries, book stores, second hand market, online via the Internet Archive's Text Archive, Google Books etc
- "merit": Amazon customer star rating, Amazon citation count, CiteSeer citation count, bibliographic system users rankings
- tags: how the work has been tagged in the bibliographic system, Amazon, LibraryThing, ...

From the above it is clear that relevance ranking is neither simple nor obvious but is a deeply "domain specific" problem. It must take into account at least 3 factors:

- the search term
- properties of the resources being searched (should collections rank higher than monographs? should widely held and highly ranked resources rank higher?)
- properties of the searcher (what is known about them, their home libraries)

Crude approaches to address ranking simplistically with general techniques unaware of the specific problem domain are unlikely to succeed as the ranking should be determined by a complex interplay of factors which will often include explicit and unstated user-specific criteria (such as item availability and reading age).

The good news for libraries and their users is that libraries are able to implement relevance ranking algorithms which take all of these factors into account, something which is much more difficult for a general search engine such as Google.

This represents a great opportunity for libraries to serve their users, not by competing with Google for general information search market but by providing an invaluable service which allows their users to discover the best information to fill their needs.

But ranking is not nearly enough. Consider the query "Ancient Egypt". Is the user wanting fiction or non fiction? Picture books or theses? Material on pyramid construction or hieroglyphics? Movies or text books?

Similarly, for the query "Patrick White". Books by or about? Literary criticism or biography? Contemporary journal articles or historical comment?

And what about "Java"? The programming language, the island, or coffee related?

Again, the rich metadata available in bibliographic records gives library systems a huge advantage with the consequent ability to give their users much better results than they'd get from a general search engines. But this is an advantage that is almost always ignored or squandered.

MARC records contain subject classifications, Dewey/LC classifications, material types, encoded classifications in many tags. All of these can be used to "group" or "cluster" results.

One approach, that used by the Lucene NBD prototype<sup>11</sup>, is as follows:

- Examine each of the top n ranked search results (n is ~1000)
- Extract
  - publication date
  - 6xx subjects (both as complete LCSH subject hierarchies and subject "facets")
  - material "form" (book, film, sound, periodical...)
  - "genres" (from 008 and various specialised MARC tags)
  - Dewey and LC classification names
  - precomputed "Conspectus" assignment
  - author names (from 1xx, added entry, 245\$c)
  - audience (from 008, some subject names)
- Count the occurrences of each in these top n results for each of these categories. Discard all but the top m ranked occurrences (m is ~ 30), except for publication date (all values are kept)
- For each category, find the total number of occurrences in the database of each m terms.
- For each category, display the top m terms ordered by occurrence in the hit set, highlighting terms statistically significantly overrepresented in the search result. Each term is hyperlinked and clicking it will reissue the search, adding the clicked term to the existing search.

This approach generates interesting groupings, allowing users to narrow their search using the information in the database. It gives them quick access to the library classification jargon without having to learn it.

For example, a search on "Ancient Egypt" can quickly be narrowed to material about "design and construction", suitable for children (juvenile audience),

---

<sup>11</sup> Lucene NBD prototype: <http://ll01.nla.gov.au> (ll01 is "LL ZERO ONE")

published in the last 6 years and held by a local library, all with a few clicks of the mouse.

The Lucene NBD prototype is an early demonstrator of ranking and clustering which indicates the need for more research and experimentation.

Another relatively cheaply implemented feature is the grouping of manifestations at the expression or work level. Although a full FRBR exercise is expensive, both LibraryThing and OCLC maintain databases of titles identified by ISBN that are somehow equivalent at the manifestation, expression or work level. By incorporating this information into the catalogue we can reduce duplicate/near duplicate results in the search list and hence make holdings more useful. It is very annoying trying to guess which of the dozens of "versions" of the "Da Vinci Code" (many of which are spuriously created artefacts of cataloguing variations) is available at which target libraries, but FRBR-like processing allows related copies to be grouped and considered as one. However, there are complications as, for example, whereas the paperback and hardback versions are likely to be substitutable, the Braille, Spanish and English versions are not.

Finally, library systems typically ignore or under-utilise the information stored in their authority structures. Rarely are search terms matched against the authority data to generate "see also", "see instead" suggestions or to automatically expand or cluster search results.

### **Exposing libraries to the flow**

Another aspect of effective discovery of library resources is what Dempsey describes as "leveraged discovery"<sup>12</sup>, making library resources visible from non-library systems. Dempsey describes how formerly, when resources were scarce and attention was abundant, users had no choice but to adapt to the library's ways of doing things. But now, with abundant resources competing for newly scarce attention, it is the library which must go to the user, fit into the users' information discovery and access flows.

As Dempsey observes, attempts at inserting libraries into "the flow" are currently not sophisticated (organising for resource pages to be indexed by Google, scripting hacks using Firefox's user scripting and bookmarklets using tools such as LibX<sup>13</sup>). However, the appearance of library specific links in Google Books and Google

---

<sup>12</sup> Lorcan Dempsey, "Discovery and disclosure": <http://orweblog.oclc.org/archives/001084.html>

<sup>13</sup> [LibX - A Firefox Extension for Libraries. http://www.libx.org/](http://www.libx.org/)

Scholar, creating "worm holes" back to library fulfilment systems, are good examples of "leveraged discovery".

### **Catalogue integration - matching and merging and deep linking**

The current NBD uses a match and merge approach which attempts to combine MARC records on ingest. This is a problematic approach because:

- Matching variant records is an extremely difficult task. The variants are often caused by minor spelling or formatting variations or minor variations in the application of cataloguing rules and name authorities.
- Matching and subsequently merging too eagerly creates a disaster: irrevocably merging together records which cannot be "unmerged" (just as an egg cannot be unscrambled). Hence, great caution needs to be exercised when merging, leading to very high thresholds being set for matching and hence lots of near (and true) matches being ignored or requiring expensive manual resolution.
- It makes some types of updates difficult or impossible. When several MARC records are merged together, the provenance of most data is lost. So, for example, if a source record from Library X has an incorrect subject Z which only they've added, and then they update their local catalogue to remove that subject, the NBD merge procedures do not know that subject Z should be removed from the merged record.
- Library local/unique data is often either lost (dropped or merged out of existence) in the process or incorrectly applied to all libraries with the same holding. Local system data was previously not of much use to an NBD whose purpose was merely to identify a resource held at some library, and the variations of local system data and the lack of context in which to display it made it more trouble than it was worth. However, we are now considering three extensions to the NBD which can profitably utilize local system data:
  - Libraries wanting to use LA as their OPAC. The NLA has recently proposed this for its own OPAC. However, to make this work various local system data must be available to the LA interface, such as:
    - local collection information ("this title is linked to 17 others (hyperlinked) as part of the XYZ collection")

- local access restrictions
  - local access URLs (eg, for licensed electronic resources)
  - multiple item (copy) information
- improved "deep linking" into contributors' OPACs. This frequently requires or benefits from contributor's local system numbers (their 001 or 035).
  - improved "getting" options, which requires a combination of the above two to represent information about local copies and allow a "getting" system to interact with a local system to determine, for example, item availability/shelf status/due date.
- Merging MARC records doesn't address the real problem of representing bibliographic resources in a meaningful context. The NBD will eventually move to an FRBR view of bibliographic resources to provide searchers with a more useful search result set and to display bibliographic resources in context. OpenWorldCat already supplies such a view and there is little contention in the library world that a FRBR view provides many advantages over even the best merged MARC record view.

The problem is that matching and merging is occurring at the wrong point in the process. The alternative to matching and merging MARC records on ingest is "soft" matching which creates a composite record from but which does not destroy the contributing MARC records. Composite MARC records may still be useful for copy-cataloguing purposes, but the main form of composite record should be the Manifestation level representation of a book in the FRBR model.

"Deep linking" refers to the ability to extract current information from a contributor's catalogue in real time, or generate a hyperlink to a resource in the contributor's OPAC. Deep linking is required to both take the searcher directly to the catalogue record of the resource (rather than to a search result screen) and for automatically extracting availability information in NBD summary results, so that the searcher can quickly see what resources are "on shelf", or currently on loan (with due date) or "not for general loan - course reserve" etc.

As the NBD collects information about journal holdings from contributing libraries, it should expose these holdings on Google Scholar on behalf of its contributors, assisting article discovery and getting for contributing library members.

## Better Getting

The resources of Australia's libraries are one of the country's greatest assets. Libraries contribute to the efficient running of the information market by aggregating the information demands of consumers (readers) and resources (the information readers want). Millions of Australians create a large market. Thousands of Australian libraries creates a huge resource. What is needed is a system to help supply meet demand as efficiently as possible.

The various current mediated inter-library loan systems are inefficient mechanisms and will become less used over time as their cost structures make them increasingly uncompetitive and inconvenient compared to booksellers and commercial electronic delivery. ILL accounts for a tiny fraction (less than 0.4%) of circulations from CAUL and public libraries in Australia and significantly lower than the 1.7% figure quoted by Dempsey for the US and described by him as suggesting:

*"we are not doing a very good job of aggregating supply (making it easy to find and obtain materials of interest wherever they are). The flow of materials from one library to another is very low when compared to the overall flow of materials within libraries."*<sup>14</sup>

Mediated ILL is just not an option available to the majority of Australian borrowers.

However, a mechanism which allows both parties (libraries and borrowers) to benefit from the wider circulation of materials can be constructed using an operational model loosely based on the very successful NetFlix approach.

For borrowers it means that they can get access to the items they want, quickly and easily, even if their local library doesn't have it on the shelf and even without visiting their local library. For libraries it means more borrowing and hence better utilization of their assets.

A simple view is that libraries lend most of their items "for free" to their patrons. However, there are obvious costs to both parties:

---

<sup>14</sup> Lorcan Dempsey, "Libraries, logistics and the long tail", February 15, 2006  
<http://orweblog.oclc.org/archives/000949.html>

- For libraries, the costs of lending: keeping track of what is borrowed and when it is due (check in/check out) and returning items to the shelf (aside from the costs of acquisitions and the fixed costs of running a library).
- For borrowers, the costs of travelling to the library to find, borrow and then return the item.

An unmediated home delivery system allows libraries to recover some of their costs as borrowers are willing to pay for the convenience of home or office delivery of the items they want to borrow. For some borrowers, this convenience is worth more than the cost of the service, so both libraries and borrowers benefit.

The major impediments to the implementation of such a system are:

- The development required to produce a system which efficiently meets the requirements of participating libraries and borrowers.
- Postal costs: there are no Australian equivalents of the extremely inexpensive "Library Rate" and "Media Rate" services offered by the US Postal Service. As a result, the unsubsidised 2-way postage of books in Australia is likely to add \$9 to the system cost.

Greater digitisation of resources, mentioned above as an aid to discovery, also confers obvious "getting" benefits on materials not the subject of copyright restrictions. But it also improves the efficiency and reduces the costs of systems such as the NLA's "Copies Direct" service which provides "fair use" extracts of books on demand.

## Conclusion

*"The future belongs not to those who merely navigate us through cyberspace, nor those who populate it with data. Rather it belongs to those who help us make sense of all the data that is available to us ."*

– John J. Regazzi<sup>15</sup>

We began by asserting that traditionally, libraries ensured that access to knowledge was available to the many rather than the few. But the exponential growth in the volume of information has left libraries struggling to cope with its

---

<sup>15</sup> John J. Regazzi, "The Battle for Mindshare: A battle beyond access and retrieval", 2004 Miles Conrad Memorial Lecture, 46th NFAIS Annual

classification, and their search systems capable of providing access to only a small fraction of what should be discoverable.

Consequently, there are “better” places to find information now than libraries: places that provide faster answers to questions, and easier interfaces. Scholarly researchers greatly value immediate access to full text and linked citations; Google Books, Google Scholar, CiteSeer/CiteBase even Amazon are starting to provide this access. More casual information seekers find Amazon, Google and Wikipedia gives them instant gratification and important value added information such as related items, “see also” lists and reviews.

While library collections currently contain millions of resources that cannot be accessed elsewhere, this is changing as rapid digitisation means more and more of our resources can be easily viewed, distributed and replicated. As things stand now, this world of vast information will be facilitated and brokered by the giant commercial aggregators such as Google and Microsoft.

We need to relate to this world, and stop thinking about ourselves as primarily collections of buildings housing physical collections. Librarians should provide and maintain trusted starting points for search which will combine results from multiple sources of information, which will locate information immediately available electronically as well as information locally available at the searcher's library, which will assist the searcher with hints and suggestions and which will do so in a commercially disinterested way.

Librarians need to describe our collections and particularly our unique resources better and more efficiently if our community is to discover them. To do this, we need better tools and simpler standards. We need to add value to our resources by engaging our communities in the conversation of knowledge creation. We need to be disseminating information about our collections in as many places as possible, making information as simple and easy to find as possible.

There is a growing body of librarians thinking and planning, and even implementing, systems that could take us to these places. Strengthening and developing these communities is a key task for libraries today. Most of all, we must remain focused on our role, which has not changed – ensuring easy and equitable access to information for all.