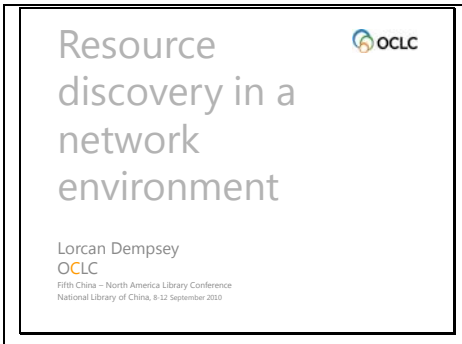


Resource discovery in a network environment[1]

Lorcan Dempsey, OCLC

Fifth China – North America Library Conference
National Library of China
8-12 September 2010

1. Prelude

I thought it would be interesting to begin with a view of the poet Li Bai from Worldcat Identities. Identities is a new way of mobilizing bibliographic data. In a typical search a user has to sift through lots of returned records trying to piece together information about an item or an author. Identities does this work up front. We mine the bibliographic data in Worldcat – a union catalog of libraries and other union catalogs – to surface what we know about the author. So, you can see a timeline of publications, an overview of their published oeuvre, a list of alternative versions of the name, a tag cloud based on subject data, and so on. In the future, this approach is likely to become more common as we make the data work harder to provide richer experiences for the user and better answers to questions.

Suggested citation:
Dempsey, Lorcan. 2010. "**Resource discovery in a network environment**." Presented at Fifth China – North America Library Conference, 8-12 September 2010, National Library of China, Beijing.
http://www.oclc.org/content/dam/research/publications/library/2010/lorcan-china-us10.pdf.

2. Overview

In this presentation I will talk about discovery of library resources in a network environment. I talk a little bit about the network environment first, then about library collections, and then bring the two together in talking about various discovery models for those collections. I would like to get across two things. First, library collections are not homogeneous and different parts have different discovery dynamics. Second, we need to think about discovery in terms of evolving network services. I bracket this discussion with very brief remarks about OCLC and Chinese libraries.

3. OCLC and China

OCLC is a not for profit library membership organization. From its origins in Columbus, Ohio, over forty years ago, it has grown to have a global footprint, providing services wherever libraries are active. OCLC builds shared capacity for libraries, allowing them to pool their resources and increase their impact through a common infrastructure. This infrastructure provides cataloging, resource sharing and discovery services, and we are now extending the benefits of the cooperative model into other areas of library operation through our webscale management systems initiative.

Chinese libraries have been working with OCLC for many years, and Andrew Wang, our Vice President for the Asia Pacific is known to many of you. OCLC's relationship with China goes back to 1986 when OCLC and the National Library of China formed a cooperative initiative to create MARC records in WorldCat of the *National Bibliography of the Republic Era, 1911-1949*.

We are pleased that Dr Anthony Ferguson, of the University of Hong Kong, is a member of the OCLC Board of Trustees, the central body in our governance structure.

Of the over 470 languages represented in Worldcat, Chinese is

the fifth most common and we expect the proportion of Chinese records to grow over time as more and more Chinese collections are represented in Worldcat. 218 libraries in China have used at least one OCLC information service. The National Library of China, Tsinghua University, and Shanghai Library, among others, contribute records to WC through Connexion.

In 2007, OCLC opened an office in Beijing and was also honored to host the previous conference in this series in Dublin, Ohio. In a major development in 2010, the National Library of China loaded two million Chinese records into WorldCat, resulting in an increase of ILL requests from other countries. OCLC expects to batchload the CALIS database into WorldCat soon.

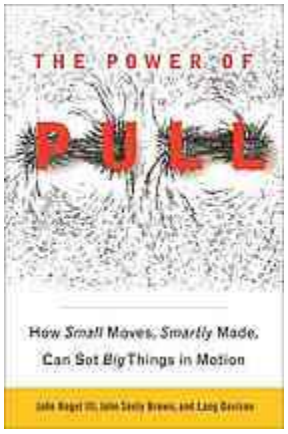
The network environment

4. The network environment

It is interesting to locate library developments in the broader context of general network information creation and use. Libraries do not stand apart from user workflows and patterns of interaction in the bigger network environment.

I have found two recent expressions useful when it comes to characterizing the impact of recent changes.

The first is the 'scalability of access' put forward by Hagel, Seely-Brown and Davison in their book, *The Power of Pull*. Access – whether to information resources, colleagues, expertise, entertainment – is no longer limited to what is immediately available in our physical environment. We can scale to the level of the network, and this influences expectations. So, we are used to being able to find a wide range of books on Amazon, to being able to prospect the whole web on Google or Baidu, to being able to connect to acquaintances through a variety of social networks. This expectation has extended to the materials historically available in libraries. Google Books is creating an expectation of web-like access to the book corpus. Services like JSTOR create an expectation of access to a broad range of journals.



The second is an interesting distinction made by Gavin Potter reported in *Wired Magazine*. He was a contestant in the

competition NetFlix ran to improve its algorithm.

> "The 20th century was about sorting out supply," Potter says. "The 21st is going to be about sorting out demand." The Internet makes everything available, but mere availability is meaningless if the products remain unknown to potential buyers.

Hagel, John, John Seely Brown, and Lang Davison. 2010. *The power of pull: how small moves, smartly made, can set big things in motion.* New York: Basic Books.

"The scalability of access"
Hagel et al.

"The 20th century was about sorting out supply," **Potter** says. "The 21st is going to be about sorting out demand. The Internet makes everything available, but mere availability is meaningless if the products remain unknown to potential buyers. "

Libraries spend a lot of time sorting out supply. The fragmentation of supply (across suppliers, databases, formats, business models, etc) has meant that we have created quite a complex staff, systems and service environment to cope. Furthermore, this has evolved piecemeal to manage evolving patterns of provision. There are separate workflows and supply industries for bought materials (think the integrated library system and catalog), for licensed materials (think knowledge base, a to z lists, metasearch, ERM), and for digital materials (think repository infrastructure). What is more this infrastructure is institution-scale - it is repeated in each library. There is significant workflow and systems redundancy across libraries. At the same time, large buildings have also been required to support this supply, as the model has perforce been to assemble materials close to the user.

This focus on supply has been because the transaction costs - in time, effort or money - for a university, for a student or faculty member, or for a member of the public, of interacting with the range of information sources is quite high. A major role of the library is to reduce those costs by integrating the sources of supply and bringing them close to the user.

However the transaction costs for the user have come down recently. Google has been a major part of this. But so has the general consolidation in a network environment: Amazon, Google Books, the discovery layers libraries are acquiring from vendors, and so on. This is related to the 'scalability of access' – access is scaling to large consolidated resources which remove the burden of integration from the user.

As supply consolidates, attention shifts to sorting out demand. Of course, libraries have always worked here, but not as much as they might have. What might this mean in our increasingly digital environment? Here are some overlapping examples:

WIRED

[This Psychologist Might Outsmart the Math Brains Competing for the Netflix Prize]

Wired, 16:03

By Jordan Ellenberg 02.25.08

- Ranking, relating, recommending. We are used to systems which provide hints and hooks for us, which guide us through large collections, which make suggestions. Worldcat Identities is an example of a service which helps here, by trying to map the data to user interests.
- Community is the new content. We expect services not only to know about resources on the web, but also to know about us. We are seeing services contextualised by their knowledge of people using those services and their relationships. Sites create value by facilitating the creation of community around 'social objects' (think of reading sites, Mendeley, BlipFoto, ...).
- Connective services. People encounter bibliographic resources in various research and learning contexts: reading lists, citation managers, personal collections, reading clubs, bibliographies, and so on. The connective tissue between these tools and library resources could be better.
- Indirect discovery. Users find materials in Amazon, in Google, in Google Scholar, in Google Book Search, and so on. How do we make connections between those services and the library? I discuss this further below.
- Embedding in other environments. It may be appropriate to tailor materials for the course management system, for the course resource pages, for reading lists and so on
- Institutional assets. Finally, one might note a major emerging area of engagement: consultation, curation, and other services around the institutional research and learning outputs that are becoming central to a wider range of activity. This is of course a big topic in itself.

It is interesting to consider the impact of these trends on a neighboring field: bookselling.

Think of the scalability of access. Amazon has become a central venue for book purchase as users scale up to wide availability. Think of sorting out demand. At webscale we are seeing the strong focus on customer relationship management through recommending and relating, reaching out to known customers and building context around their interests. At the same time, print outlets are moving towards niche, highly curated offerings reflecting the interests of a particular geographic or topical community. What gets squeezed are bookstores who had large collections in a print environment, but do not scale to Amazon's level, and who do not have the customer knowledge to manage demand very well.

Libraries are in an interesting position here. They are very much 'institution-scale' organizations, and do not have rich data-driven knowledge of their users.

Against this background, here are some summary themes.

Community is the new content: We are now very used to interacting with resources in a social context. The application of community to content, in terms of discussion, recommendation, reviews, ratings and so on, is evident in many of the services we use, and in some form in most of the major network servies we use (Amazon, iTunes, Netflix, ...). Indeed, this is now so much a part of our experience that sites without this experience can seem bleached somehow, like black and white TV in a color world.

Attention switch: In a print world resources were scarce and attention was abundant. There were a few places to go for particular types of resource – government information, technical reports, journals, and so on. At the same time, we recognized that we had to 'spend' time to interact with these. Now, however, resources are available from many resources and we are no longer limited to what is locally available. Resources are abundant and attention is scarce as people have more calls on their time. This puts a premium on convenience and ease of access.

Workflow switch: Much of our information creation and use is

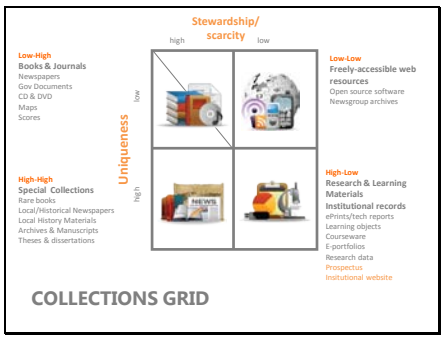
now carried out on the network. This may be assisted by the prefabricated workflow supported by a course management system, a lab notebook, or a pre-print archive, for example. Or it may be assisted by the bricolage of tools we use to find and organize information resources: citation management services, bookmarks, RSS readers, twitter clients, and so on. People have varying levels of sophistication of support within an overall trend towards adopting research and learning workflows on the network. What this means is that while users may once have built their workflows around the library, now, the library needs to consider how to build its services around the user workflow, to be available where its readers are doing their work. Think for example of Netflix which works hard to make itself available in as many ways as make sense for its users. We can get a DVD. We can also stream to a PC, an xBox, an app on the iPad, and so on.

Collections

5. Collections

When we think about resource discovery in a library context we think about library collections, individually or in aggregate. Accordingly, it is worthwhile to spend some time thinking about the shifting boundaries of library collections.

We have found the Collections Grid a useful tool to help with this.



COLLECTIONS GRID

The collections grid organizes resources according to two values: 'uniqueness' and 'stewardship/scarcity'. Things that are unique, or rare, tend to be in one collection only. Things that are not unique or rare tend to be in many collections. Things that are highly stewarded are things that attract library attention, have resources and time spent on them, have systems infrastructure devoted to them, and so on. Recently, we have observed that stewardship and scarcity go together; we have developed stewardship models around materials which are scarce.

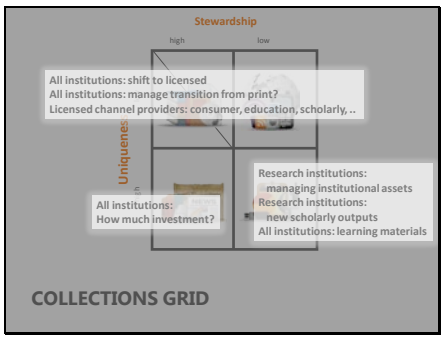
This gives us four quadrants.

Upper left: published materials: books, journals, DVDs, … These are the current core of library collections; they are supported by an extensive support industry and mature systems; and attract the major part of library staff and resources. An important distinction is between materials which are bought (books etc) and materials which are licensed (journals etc) a distinction that approximately corresponds to print/electronic.

Upper right: we put the open web here as it can be replicated, indexed, etc, across collections.

Bottom left: here is what we know as special collections and archives, rare and unique materials which are heavily stewarded. Theses and dissertations, local history materials, and other materials are here. These are attracting more attention as they are seen as distinctive and as they become digitized.

Bottom right: this is a major and growing category. As research, learning and administration are carried out in a digital environment, they generate materials (eprints, research data, learning objects, administrative records, etc) which need to be managed as institutional assets.

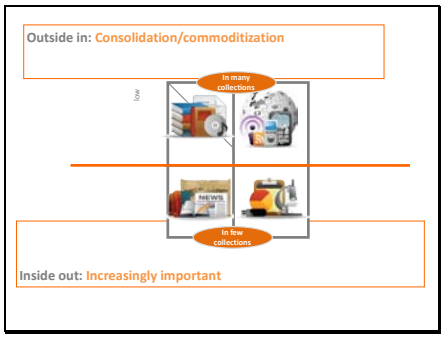


We can see different drivers in each of the sections, and some different directions.

The proportion of the collection that is print and/or bought will decline and managing legacy collections is becoming a big issue. Licensed collection will increase, and more books will be available this way. In the US the discussion around Google Book Search has been very important.

There is a wish to release special collections into research and learning activities through better description, digitization, and exhibitions. However, one wonders how much investment this area will receive compared to other areas needing attention.

Institutional materials are likely to be an area of growth, whether management infrastructure is sourced locally or in the cloud.



This analysis points us towards a distinction that is actually quite important in how we think about discovery.

Think of a distinction between **outside-in** resources, where the library is buying or licensing materials from external providers and making them accessible to a local audience (e.g. books and journals), and **inside-out** resources which may be unique to an institution (e.g. digitized images, research materials) where the audience is both local and external. Thinking about an external non-institutional audience, and how to reach it, poses some new questions for the library.

Think also about the relationship between the 'locally available' collection and the 'universal' collection.

* For bought materials (books, CDs, ...) the library provides access to the locally available collection - the materials acquired for local use - and then may provide access to a broader 'universal' collection through Worldcat or another resource.

* For licensed materials, access is first through the broader 'universal' level (in various databases) before checking for the subsect of locally available materials.

* For institutional digital materials, access is provided to local repositories but this will not typically be backed up by access to a 'universal' source for such materials (although, one can see attempts to do this, as, for example, where an institutional

repository expands a search to Scirus).

Of course, if one thinks about other discovery/disclosure channels (Google, for example), these collection types also behave differently, of which more later.

3 modes of discovery

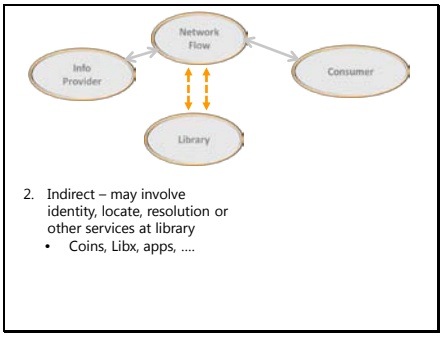
6. Three modes of discovery

I have spent some time talking about collections. Now I want to turn to discovery, and think about how these collection types play in various discovery modes or models. This connects back to my discussion of the network environment. I suggest that there are three productive ways of thinking about discovery here. I label these:

1. Direct
2. Indirect
3. Disclosure and syndication

1. Direct – outside in
   * Discovery layer

Direct discovery is the pattern to which we are most accustomed. A library acquires materials from providers, organizes them, and discloses this collection to local users through a discovery system. We have seen rapid development in this discovery system. At first it was a catalog. Then services were provided to provide access to the journal literature, leading to metasearch engines. Recently, there has been an emergence of a new service category, the 'discovery layer'. What does this mean? A discovery layer provides a single point of access to the full library collection across bought and licensed materials, sometimes pulling in metadata for local institutional digital materials also. Typically, a single search box is offered alongside a range of other navigation features. Products which support this approach include Worldcat Local, Summon, Primo Central, and the Ebsco Discovery Service, as well as a range of institutional, national or other initiatives.



However, we are very aware that increasingly 'discovery happens elsewhere': library users find resources in various network services outside the library environment. In this picture I refer to the 'network flow', to suggest the variety of services from which users construct their workflows, or learning flows, or … These include search engines, social networking services, social bookmarking or personal collection services, and so on. For my purposes here they also might include non-library services like the course management system.

If *discovery happens elsewhere*, then there are several important consequences for libraries. Most important is the recognition that a library's own, locally managed or provided discovery environments - the catalog, metasearch service or discovery layer - are only a part of the picture, that there are other areas of discovery which would benefit from attention.

Accordingly, libraries will also want to support **indirect discovery**. By this I mean that they will want to connect the discovery experience, whenever it happens outside of the library environment, to the possibility of fulfillment in the library.

This may happen in several ways. Importantly, it makes sense that libraries will want to *disclose* the existence of their resources into other discovery environments. Think of a library's unique resources for example, its digitized special collections, for example, or the institutional assets it manages in an institutional repository. As with other information providers on the web, the library will want to make sure that these are exposed in ways that optimize crawling, indexing and finding by search engines. Other approaches may be sensible, adding relevant links to Wikipedia pages for example, or selectively putting images from the collection on Flickr.

For non-unique resources a library may want to disclose the fact that it holds a particular item. Think here, for example, of making sure that Google Scholar knows how to resolve article metadata to your particular library (see the Library Links program). Or of being represented in one of the several union catalogs (including Worldcat) that Google uses to direct the 'find in a library' link on Google Book Search.

Another approach is to 'leverage' a discovery environment which

is outside of your control to bring people back to your environment. Here I am thinking of the use of tools like LibX, which may mobilize metadata found 'elsewhere' in a variety of ways to connect to a particular library resource. The developers report that LibX has been customised for over 700 different use environments.

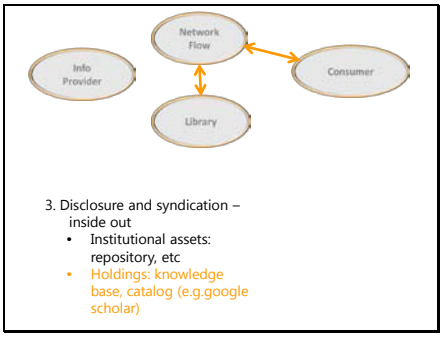
Other approaches could also be discussed. We don't yet have a routine way of supporting 'indirect discovery' or a shared inventory of use cases. This will be one of the more interesting development areas in coming years.

Here is an example of indirect discovery. I have configured Google Scholar to point me at the National Library of China resolver when I do a search.

Now, a discovery experience in Google Scholar is connected with a fulfillment experience in the library. Discovery happens elsewhere but the library can be available at the next stage.

A web index for books?

Here is another example, this time in Google Books. The 'find in a library' link will take you to one of several union catalogs depending on where in the world you are. Users in the US and many elsewhere are directed to Worldcat and thence onwards to a library of choice.



I have already discussed the third mode, **disclosure and syndication**. Here the library wants to make sure that its resources are appropriately represented in external discovery environments. The pattern varies by collection type. For unique resources, where the inside-out model is at play, the library will want to ensure that materials are effectively disclosed to search engines and that metadata is syndicated to relevant other resources (e.g. Oaister).

For licensed resources, the library may want to make sure than an external resource can find its resolver or is aware of its holdings. Similarly with the catalog and bought materials. The examples above with Google Books and Google Scholar show how disclosure and syndication support discover in other environments, and linking back to library resources.

7. OCLC

Before concluding with some comments about how Chinese libraries and OCLC might work together more, I will say a little about how OCLC services work within the pattern I describe above. A major role it plays is to try to work in an integrated way across these various environments and help libraries position themselves appropriately.

In Worldcat.org, OCLC has established a service in the network flow, and is building community around it. It provides a network level service which aggregates library presence. It also syndicates this presence to other network level services like Google, Facebook, Twitter and so on. It also works actively with partners on linking and data sharing agreements to support the indirect model by linking various discovery environments back to libraries. Wherever Worldcat is, there is a link to libraries.

In Worldcat Local, OCLC provides a direct discovery experience for library users, across the range of its collections. It is a 'discovery layer' service, however, importantly it connects the local library with other libraries who participate in the Worldcat network, for discovery and for resource sharing purposes. The

user can move seamlessly between the local collection and the collective collection represented in Worldcat.

Worldcat, and Worldcat Local, are available through Mobile apps, and a variety of widgets, allowing users to build it into their workflows. They can also create lists to manage resources and interact with Worldcat services in a variety of ways, creating a profile, their favorite library, and so on.

OCLC works with many providers, including JSTOR, to represent their intellectual content in Worldcat and to tie it to library users who are authorized to use it.

* Worldcat and Baidu?
* Use Worldcat as one channel to expose Chinese literature, scholarship, and cultural heritage to the rest of the world
  * Interlibrary loan to National Library went up when records loaded
* Participation in shared services?
  * VIAF, ...
* Analytics for the global collective collection

OCLC and China

8. OCLC and Chinese libraries

To conclude, here are some thoughts about cooperative possibilities.

I have spoken about Worldcat and Google. Does it make sense to develop similar arrangements with Baidu?

Worldcat is a great library switch – switching users from Worldcat to individual libraries, and switching users from other discovery environments to libraries. It is thus a valuable venue for disclosing the literature, scholarship and cultural heritage managed by libraries to other libraries, library users, and general network users around the world.

OCLC manages a variety of shared services, including VIAF, the Virtual International Authority File, which harmonizes authority files from national libraries around the world. VIAF and OCLC would welcome conversation about collaboration.

As the global book collection continues to change in our current environment it becomes more important to map and measure it so as to ensure its continuity, to understand its composition, and to find better ways of collaborating around it. OCLC is well-placed to answer questions about this resource based on its aggregate data and is interested in making connections between libraries in China and libraries in the US who have mutual interests around particular collecting areas.

9. Conclusion

Thank you for your attention. I look forward to working with you all to advance the outcomes of this event and I also look forward to meeting with colleagues again at the next China – North America Libraries conference.