July 2014

The Evolution of Library Discovery Systems in the Web Environment

Mark Dahl
Lewis and Clark College

The Evolution of Library Discovery Systems in the Web Environment

by Mark Dahl

Associate Director for Digital Initiatives and Collection Management, Aubrey R. Watzek Library, Lewis and Clark College

In December 2008, the Orbis Cascade Alliance, a consortium of academic libraries in Oregon and Washington, launched a new union catalog on OCLC's WorldCat.org platform. This change resulted in an updated Web interface, better keyword searching, and faceted results. However, we also lost some features that worked well in our old system. But the larger significance of this change might not be obvious. A shift has taken place, one that moves us into a new paradigm for the systems that support discovery of resources in libraries. The Summit catalog is now part of a great global organism known as WorldCat, and that organism is poised to be more dynamic and more ubiquitous than any of our old local catalogs could have ever been. How did we get here? I will attempt to answer that question through my personal account of library search and discovery as a librarian and technologist since the mid-1990s.

I entered library school in 1996. As the Web emerged, I developed a growing curiosity for it and delved into HTML coding, Web programming, and Web server administration. In those early days, the library community was just digesting the obvious advantages that the Web had over previous technologies like Gopher and Telnet: mouse click hyperlinking and richer graphics. The underlying discovery systems libraries used continued much as they had in the past with prettier Web-based interfaces on top.

By the late 1990s some transformative changes began to take shape in the online library world and on the Web. In the library world, full text databases and services like JSTOR arrived on the scene, putting large amounts of actual content, not just indexing, online. The general online fulltext database became the bread and butter of our online offerings at Central Oregon Community College, which we were positioning to support distance education. On the Web more broadly, e-commerce gained ground and people got used to shopping experiences that involved search, discovery, and fulfillment.

In 1998 Google was founded, and by the early 2000s it was the most popular search engine on the Internet. Google's clever PageRank algorithm harnessed the collective intelligence of the Web by using hyperlinks to help determine relevancy. It was a system that benefited enormously from the sheer scale of Google's computing power. More importantly, it got smarter as more people used it. Google proved that a Web scale enterprise could achieve things that small- and medium-sized players could not. In a similar way, dot-com crash survivors like eBay and Amazon established that in certain markets there was only room for a few large players on the Web.

While Google was growing its search business, libraries mostly ignored search and worked on the problem of organizing a growing array of full text resources. Libraries were acquiring access to electronic journals by the bucketful, but it was hard to find out if a given library had access to a particular journal. By 2001, I had moved to Watzek Library at Lewis and Clark College, and one of my first tasks was to develop a way to search our electronic and print journals by title. In response I created a database that mixed together data from our ILS and Serials Solutions and would later support an OpenURL resolver.

In the early to mid-2000s, library catalogs began to adopt more of the trappings of mainstream e-commerce sites by incorporating cover art, external links, and fancier Web design. They remained weak in search functionality. In 2005, major figures in the library technology community like Andrew Pace and Roy Tennant began asking rather loudly why OPAC search left so much to be desired when compared

5
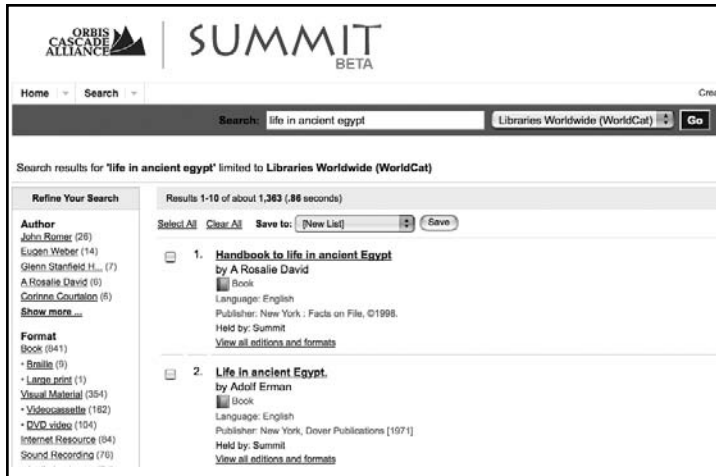
The new Summit catalog has relevancy ranking based in part on library holdings as well as next generation catalog features like facets.

with commercial Web search (Pace 2005; Tennant 2005).

Projects emerged that attempted to significantly improve search functionality in Web OPACs. They included North Carolina State University Library's catalog based on the Endeca search engine and Casey Bisson's WPopac (now Scriblio), an OPAC based on the modular WordPress blogging software.

In the early 2000s libraries also began to break important new ground with digital collections mounted on systems such as ContentDM and DSpace. These were the first Web-based discovery systems managed by libraries that harnessed the Web's global reach. Library catalogs largely contain references to books held by hundreds of libraries and are typically closed to search engines because of the redundancy of their data. By contrast, digital collections contain unique materials and are generally open to search engines, allowing people anywhere on the globe to find and use their content.

In late 2005 and early 2006 I co-authored a book, *Digital Libraries: Integrating Content and Systems,* with Kyle Banerjee and Mike Spalti. We started work on the book with a loosely-conceived thesis: that integra-

tion of disparate content and systems with Web technologies could create exceptional online services for libraries. We argued that library systems, including discovery systems, would be many dis-integrated units tied together by standards and clever Web programming. Modular digital library tools like OpenURL resolvers, electronic resource management software, and digital asset management software, the trend toward OPACs running atop ILSs, and federated searching systems that relied on new standards like SRU/W (search/retrieve via URL or Web service) all seemed to confirm this thesis.

But as we researched the book in late 2005, it became clear that this model did not explain it all. More and more, users were beginning to encounter library resources on the Web outside the "walled garden" context of library-managed discovery systems. People might discover books on Amazon or articles on Google Scholar and then acquire the content via a library's physical or virtual gateway. Moreover, Web 2.0 sites like Flickr, del.icio.us and YouTube allowed users to contribute and organize digital assets in a collective fashion. Like Google, these Web 2.0 sites got better as more people used them and aspired to a Web-wide audience.

In April 2006, I heard Lorcan Dempsey of OCLC give a presentation to the Orbis Cascade Alliance Council on "Moving to the Network Level: Libraries, Readers, and Applications." Dempsey discussed the shift from vertically integrating services within a single institution to "collaboratively sourcing" services in concert with external players. The Alliance's own union catalog, which aggregates supply and demand for books among 30+ academic libraries, served as a strong example of regional collaboration. Dempsey encouraged the group to broaden its thinking to resource sharing that would involve "multi-level" collaboration between individual libraries, regional

consortia and global players like OCLC, JSTOR, and Google. He challenged the group to think about "painful" activities being done at the local or regional level that could be more effectively done by higher level organizations and systems.

In some respects, the idea of outsourcing library systems to larger-scale players went against my instincts. I'd always enjoyed managing my own servers and writing my own Web applications. There was something inspiring about being able to load Linux on an old PC and run my very own Web presence from that little box humming away in the closet.

Nonetheless, I couldn't get the "moving to the network level" phrase out of my head. In late 2006 and 2007, I discovered that the idea related to the various Web applications that I began using at work and in my personal life. Gmail revolutionized my productivity at work. I benefited from its great search and organization features, powered by Google's huge infrastructure far away from my PC. At Watzek Library, we began using Basecamp and Google Docs for project management and collaboration. At a time when I supported a collection of digital images for teaching on MDID digital collections software, I was impressed with how much better Flickr managed digital assets. Meanwhile, buzz around the concept of cloud computing grew, especially with the publication of Nicholas Carr's *The Big Switch* in early 2008, which explains how computing power in far-away data centers is revolutionizing both personal computing and back-end IT infrastructure.

In 2008, our library began implementing two network level discovery services. In winter 2007/2008, the Alliance struck a deal with OCLC to create a union catalog solution based on the WorldCat.org platform. WorldCat Navigator is a consortial version of WorldCat Local that provides a catalog with the wide scope of WorldCat.org but with discovery and delivery features tailored to the needs of the Alliance.

Given the growing shift in my thinking, I saw several advantages in the Alliance move to WorldCat. The interface is more modern than the old Summit and offers conventions from the consumer Web such as narrowing searches by facets and creating user accounts for favorites. More compelling, however, is the broader concept of having a catalog that is a part of a larger organic whole. The WorldCat database is a dynamic, ever evolving thing, updated by a global community of catalogers. Unlike our local catalogs, where we download records and they remain mostly unchanged like a card in a card catalog, WorldCat operates like a Web 2.0 site: a community of people can cooperatively add metadata to improve digital objects, albeit in a much more regulated, library-world way. WorldCat's global, ever changing holdings information allows WorldCat.org to have an unparalleled relevance ranking of books, not unlike Google's PageRank concept. The WorldCat.org platform also supports user-contributed content like ratings and reviews, a service that will be progressively more useful as more libraries and users come on board.

Moreover, with WorldCat.org, OCLC takes a lesson from Google and Amazon and understands that Web scale matters. In order for library content to be noticed on the Web, it needs to be presented by a global player, not in a diluted fashion from thousands of separately managed library catalogs. Unlike local library catalogs, WorldCat.org provides a place to reference a book that is useful for anyone on the Web and maintains relationships with commercial search vendors so that its records will appear in search engine results. Furthermore, it provides a catalog with common conventions for searching and viewing records not unlike Google providing a certain consistency in its interface across the Web.

As Watzek Library threw its weight behind the Alliance WorldCat project, we got another innovative network-level initiative underway. Our visual resources curator, Margo Ballantyne, and a faculty member in Ceramic Arts saw an opportunity to create an online image collection of contemporary ceramics. The challenge would be collecting the images and metadata from artists dispersed throughout the world. The Digital Services Coordinator, Jeremy McWilliams, and I were avid users of Flickr and knew of its powerful Web-based tools for managing images. With little money behind the project for staff support, we came up with the idea of having artists contribute images and metadata and assign copyright through their own Flickr accounts. We would then assemble the images in a Flickr Group and present them as a coherent digital collection via a Web site driven in part by the Flickr API. We implemented this idea in the spring of 2008, albeit with some technical modifications to our initial vision (McWilliams 2008).

This site, http://accessceramics.org, is a live, growing collection of contemporary ceramics images that reside in individual Flickr accounts but are organized together into a digital collection with a defined set of metadata. In contrast to digital collections that are cataloged centrally, our metadata is entered by the contributors. We found some similarities to this model in the digital history projects launched by the Center for History and New Media such as hurricanearchive.org. We also found affirmation in our selection of Flickr when the Library of Congress launched a collection of images in the Flickr Commons in 2008.

These recent experiences have convinced me that a new model for library discovery systems may be emerging, one characterized by global discovery systems like Flickr, WorldCat.org, and new ones yet to surface in both the profit and non-profit sectors.

> The challenge of library technology and metadata professionals will move from managing a library's own set of isolated databases to managing their library's imprint on shared global discovery platforms.

These will be systems that benefit from the network effects allowed by Web scale: they will get better as people and organizations use them and contribute to them. The challenge of library technology and metadata professionals will shift from library management of isolated databases to managing their library's imprint on shared global discovery platforms. Libraries will still strive to provide specialized interfaces and metadata for their users, but the work will be done in this new global context.

If a library develops a special vocabulary for a subset of its collection, it will add the terms to a global database so that this vocabulary, however esoteric, can have a broader benefit. With our likely move to WorldCat Local here at Watzek, I'll encourage our cataloger to start adding genre headings for videos on WorldCat instead of doing the work in our local system. Rather than sweating out upgrades to library-managed OPAC software, we will enjoy WorldCat Local's "software as a service" model that assures it is being constantly improved and upgraded, just like Gmail. When we feel the need to customize, we'll use APIs to create interfaces tailored to our user communities. We'll also take the opportunity to mash up data from multiple sources on the network. For example, Watzek recently created a proof of concept mashup with the

WorldCat API and the Google Book Search API that creates a Google Books search with library holdings in the result set. This platform shift should benefit smaller libraries like Watzek, who will now have access to a search and discovery infrastructure that is as good as that used by the big players. Hopefully, these shared platforms will spark new innovations in collections and services by both small and large libraries.

The movement towards network-level discovery systems for libraries is emerging in an uneven manner typical of new technologies. I welcome the complexity, chaos, and change. As has been the case in the recent past, much of our job will be managing change for our user communities, both technically and via communication with our constituents. These network level systems should make it easier to do basic research and access common material. User expectations for more specialized materials and services should increase. Whereas we have historically concentrated most of our energy on commonly published material in familiar forms, in this global discovery environment we may find ourselves working at the extremes. We will be curating physically and digitally what we have that is unique and of interest globally, as well as assisting with new forms of intellectual output that don't neatly fit the book or periodical categories.

Wherever we end up, it should be a good ride. 🌿

References

Carr, Nicholas G. 2008. *The big switch: Rewiring the World, from Edison to Google.* New York: W. W. Norton & Co.

Dahl, Mark, Kyle Banerjee, and Michael Spalti. 2006. *Digital libraries: Integrating Content and Systems.* Oxford [UK]: Chandos Pub.

accessceramics.org is a collaboratively-sourced digital collection created with artist-contributed images and metadata. It uses Flickr as the underlying digital asset management system.

Dempsey, Lorcan. Moving to the network level: Libraries, readers and applications. 2006 [cited 12 February 2008]. Available from http://www.oclc.org/research/presentations/dempsey/cascade.ppt.

McWilliams, Jeremy. 2008. Developing an Academic Image Collection with Flickr. *Code4lib Journal* (3) (6 June 2008). Available from: http://journal.code4lib.org/articles/74

Pace, Andrew K. 2005. My Kingdom for an OPAC. *American Libraries* 36, (2) (February): 48-9. Available from: http://www.ala.org/ala/alonline/techspeaking/2005colunms/techFeb2005.cfm

Tennant, Roy. 2005. "Lipstick on a pig". *Library Journal* 130, (7) (April 15): 34. Available from: http://www.libraryjournal.com/article/CA516027.html