Chapter Five

# Library Discovery
## From Ponds to Streams

*Kenneth J. Varnum*

Resource discovery in libraries has undergone a remarkable evolution over the past five years, tracking (but lagging behind) what has happened on the public Internet. As a handful of companies (Google, Bing, Yahoo!, Yandex, Baidu, etc.) have emerged to provide near-universal access to public information on the Internet, there has been a rising desire within the library world for similar access to licensed content. Libraries, and libraries' perceptions of the patrons' needs, have led to the creation and acquisition of "web-scale" discovery services. These new services seek to amalgamate all the content a library might provide access to—the catalog, online journals, abstracting and indexing databases, institutional repositories, open access sites, and more—into a single index. Much like the big companies on the public Internet, these new services build their indexes of content and provide access.

## PRIMORDIAL DISCOVERY PONDS

The evolution of online library research has undergone a significant transformation as well. Not long ago, a library's online resource portfolio could be thought of as collection of ponds. While they might coexist in a broad ecosystem, they were only

A portion of this chapter is based on "Perspective on Discovery," which was published on RSS4Lib on September 26, 2013, www.rss4lib.com/2013/09/perspective-on-discovery/

loosely connected closer to the surface (if at all). To conduct effective research, a library user would need to navigate to each pond in turn, dip his toes in (or dive in), but then would need to travel to another pond when investigations in the first were completed. Many scholars would find their particular favorite ponds and keep going back to them, regardless of how appropriate that particular collection of information was to the specific research problem at hand. For many users—particularly those who are deeply knowledgeable about their areas of study—this was an excellent tactic and served them well. These researchers often know the specific databases for their field. However, for researchers who are not experts in a field, or for those same experts who might be exploring the overlapping margins of two related disciplines, finding the right database, or series of databases, was onerous and often problematic. This feeling of inefficiency was strengthened by the rise of the Internet search engines that gave the appearance of total coverage.

To meet the desire for breadth of coverage, libraries turned to federated search technologies.[1] In a federated search, a single user query is directed to multiple databases simultaneously. A limited number of results from each database is retrieved and merged into a single result set. A number of vendors offered federated search products, including Ex Libris (Metalib), Millennium Access Plus (Innovative), and WebFeat's eponymous offering.

## THE INFORMATION SEAS

Federated search moved the library patron from the pond to the ocean—or rather, allowed the searcher to more efficiently query multiple databases at once. In slightly more time than it took to get to and conduct a search in a single database, these tools searched multiple databases simultaneously and returned consolidated results from all of them. However, as significant an advance as federated searching was, it was beset by a host of challenges. A federated search is inherently slower than the slowest target database. This is because federated searching relies on a series of sequential processes. Once the search term is sent out to each target database, the search tool waits until results come back from all providers (or the response is "timed out" or otherwise fails to come back after some preset threshold of time). Because each database provides its own independent (usually proprietary) relevance ranking, the collected results then need to be reranked and presented in a consistent way. Until all this processing is done, the user sees either a "processing" message or an evolving results list where more relevant items supplant the ones already displayed on the screen.

Initially, this process was seen as a significant advance. However, as Internet search tools became faster and more all-encompassing, the user experience of federated search quickly began to pale. Library researchers became accustomed to the lightning-fast response times they would find when searching other online resources; the perception of a twenty- or thirty-second delay in a federated search product became intolerable. Patrons rarely were concerned about the fractured nature of back-end technologies that were being searched and integrated; if Google could do it, people reasoned, why can't the library?

## OCEANS OF DATA

Despite these major challenges, federated search tools solved a need of many library patrons and libraries: they were a means of trawling the ocean of information available to them. And once they had seen the ocean, few wanted to return to the ponds. If federated search was not the solution, then perhaps the situation could be improved by taking a page out of Google's playbook and building a single index. These emerging tools combine the full text and indexing for hundreds of millions of scholarly objects into one huge index, which then can be searched in "Google time": quickly and efficiently, with a single relevance-ranking mechanism in effect, regardless of the source of the data.

Most discovery services provide a view of the entire collection of searchable information that is connected with the entitlements of the user who is doing the searching. The library provides its licensed content holdings to the discovery service, which then can mediate user queries against the corpus of material that is available to that user. Because the index includes a greater breadth of content than is available at any participating library, it is possible for a user query to be conducted against all content, not just content licensed for local access. Most libraries choose to limit the default search to readily available content but to allow the user to expand to the whole universe of materials.

Index-based discovery products for library content have a relatively brief history. Scholars Portal, a consortial effort in Ontario that brought together the physical and licensed collections of Ontario's twenty-one university libraries, pioneered the single-index approach for licensed library content in 2002.[2] Commercial products soon emerged from other sources, developed along the same conceptual lines. The first to reach the market was Serials Solutions' Summon service in 2009. Analogous products from EBSCO and Ex Libris were launched thereafter, joining OCLC's FirstSearch in this market area. While each tool has a different technological approach, the end effect for the library user is approximately similar: one search box that includes a broad selection—approaching totality—of content that approximates the library's owned and licensed digital content.[3]

In the four years since discovery services became commercially available, they have seen a rapid uptake by research libraries and, to a smaller degree, by public libraries. The concentration of discovery services in larger public libraries and academic libraries is due to two factors. The first, and far from trivial, is cost. These services tend to be expensive. The second factor is the breadth of content acquired by the library. Larger libraries tend to have more disparate data sources to search, with corresponding effort required by a researcher to be truly inclusive when looking for information. Smaller libraries often have smaller portfolios of licensed content and, often, acquire multiple databases from a single provider that already offers platform-wide searching. The need for integration is less keenly felt in these cases.

## MOVING TO STREAMS

This brings us to the current stage, where many (mostly larger) libraries offer their patrons a Google-like experience: a single entry point to everything the library has to offer. Now that we have finally found the holy grail, we are beginning to understand that it comes with challenges of its own. In many use cases, library searchers indeed want—and need—access to the breadth of the library's holdings. They are doing truly exploratory searches and desire access to everything, or at least a representative sample. The increasing focus on interdisciplinary research highlights the benefits of searching the entire ocean of library content in one go. At the same time, many use cases indicate the advantage that a narrower scope provides, a simultaneously all-encompassing search of materials within subject disciplines, rather than across them.

Over the past decade—a time period that encompasses libraries' efforts adapt to discovery tools—libraries have been finding ways to integrate their resources into their broader communities. This concept was defined by Lorcan Dempsey in 2005 in a well-known post, "In the Flow." He described a world in which the library should strive to make its services and resources available to researchers "in the user environment and not expect the user to find their way to the library environment," and in which the "integration of library resources [into other systems] should not be seen as an end in itself but as a means to better integration with the user environment, with workflow."[4] In the years since Dempsey wrote this, libraries have followed the course he described: embedding librarians in academic departments and research groups, making library-curated data available through open and automated mechanisms, integrating research tools into their websites, and more.

Dempsey's recommendation was that libraries place themselves in the flow of the research process. While being in the flow is essential, with current discovery technologies, libraries can now do more than make sure they are *in* the flow; we are now well positioned both to *create* streams of information that researches will dip into, and to provide functional and valuable access points to these streams. In many ways, this new capability to divert the oceans of information available through discovery systems into streams of context-appropriate resources for researchers, individually or collectively, will unlock the true value of resource discovery systems.

Streams, or flows, of information are not as useful if libraries do not tailor them to the specific categories of library users who are will benefit most from accessing them. It is time for discovery to focus on the user interaction and experience, now that vast bodies of information are available to be searched.

I have long felt that discovery tools are most effective when they are presented in the context of the library whose patrons are doing the research. The discovery services' native interfaces are well designed, user tested, and largely accessible to users with varying degrees of print or physical ability, but they do not easily allow for mapping resources into categories that make sense at the local level. In an academic library setting, for example, it may well make sense to offer focused searches based on users' course enrollments. The University of Michigan library built an experimental tool several years ago to do just this.

In our experimental project, nicknamed "Project Lefty" (the goal was to get the user the right content, at the right level, in the right subject areas—and as we all know, three rights make a left), we developed a front end to our Summon system that scoped searches based on a user's course enrollment.[5] The system would take the user's search query and pass it on to the discovery tool, along with a set of resources (largely online journals, but the pool of resources could also include e-book collections or other full-text sources) that should be searched. A group of subject-specialist librarians organized journal titles relevant to their subject areas into one of three categories: novice, expert, or both. Journals in the novice category were those that were appropriate to lower-level undergraduate courses; generally, these were the more accessible peer-reviewed and popular journals in the subject area. Expert journals were the more narrowly focused, deeply scholarly publications. Some journals, of course, are broadly relevant to a subject area and were included in the "both" category. In terms of courses, we categorized 100-, 200-, and 300-level courses as "novice" and 400-level (and higher) as "expert."

Because we could ask campus users to log in to the system—and could therefore access basic course enrollment information about them—we could know which courses each student was enrolled in an present an option to focus a user search query on just those resources, identified by librarians, that were most likely to be appropriate for search query. For example, a student who was enrolled in a 100-level geology class, a 300-level economics class, and a 500-level psychology class could opt to apply her course filter to a search for *depression* and see, in turn, introductory materials on depression (the geographic feature), materials on depression (the economic condition), or deeply scholarly materials on depression (the psychological condition).

Conversely, such a system could allow a researcher to remove one's own native discipline from a results set. For example, a psychology scholar interested in the effects of the Great Depression on the mental health of the populace could search for *depression* but exclude economics content from the search results.

Discovery services offer a plenitude of keyword vocabularies, controlled and free-text, connected with the articles they index. Subject terms provided through controlled vocabularies or uncontrolled author association are rarely easy to distinguish, , making pure keyword filtering a challenge. Having a focused title-based categorization is a more effective mechanism for scoping a search and allows libraries to bridge the gap between subject-specific databases and broad, source-agnostic discovery tools.

At a more generalized level, even if a library does not wish to provide purely individualized search results, an intermediate level is possible—a course-specific search interface that scopes the user's search to the materials determined to be most relevant. The same basic data categorization process would take place (though perhaps at a more granular level), through which subject specialists could assign course-level indicators to a range of online resources, taking into account course syllabi and the librarians' subject and content expertise. Course-specific searches would then cover this suite of resources. This would be of particular benefit to novice researchers who are getting familiar with the concepts of resource review and selection, by giving them simple, powerful interfaces into sets of resources that are germane to their needs.

## AN UNTAPPED RESERVOIR

Much of the work around resource discovery up until now has been focused on the content being searched. The construction of large-scale single-index repositories of scholarly information has been a huge technical challenge. There have also been difficult negotiations with the publishers of full-text materials and the abstracting and indexing services that provide additional value and access points to the scholarly world through their contributions. The untapped input is the researcher. The next phase of discovery will focus on the consumer of the search results: the scholars, researchers, and library patrons in general. The researcher's identity carries with it a range of associated information that, if tapped on an individual basis, can streamline the research process. The researcher does not have to be only a consumer of the discovery service; the researcher can also be an input, transparently, to the search query.

This data is particularly accessible and relevant to academic libraries because they are generally well positioned to access the complete researcher ecosystem: information about students through campus directories and data from the registrar (courses enrolled in), information about faculty from institutional repositories (results of research already conducted) and the registrar (courses taught), along with campus directory information for everyone, and more. While equivalent data are possible to obtain in the public library context, the data sources are not as obvious and are much more likely to be sparsely provided through an opt-in system.

## PRIVACY

Whether a system that treats "environmental" data about the researcher is opt-in or opt-out, it is important to remember that the system must be sensitive to the user's need for privacy and confidentiality. Libraries have the technical ability to access user-specific data to build personalized discovery environments, but should tap that reservoir only if there are ways for users to opt out on a permanent or case-by-case basis. Some researchers may not want to conduct a particular search through logging in; they may wish to remain anonymous throughout the process, and may not want to see tailored results ever. Other researchers may generally want to have access to tailored search results, but may want to step out of the personalized stream back into the ocean to see what they might be missing. Both needs are real and must be designed into whatever system is offered.

There is a strong argument to be made that these customizations are better done at the library level than by the vendor.[7] Libraries—and academic institutions—generally have more clearly articulated policies and procedures around information privacy than do vendors. It would be a difficult sell to pass along a user's course registration information to an external vendor, connected with a login, when the same information could be used on the library side of the relationship to tailor search results without exposing an individual user's identity to any vendor's system.

## CONCLUDING THOUGHTS

The power of discovery, in my way of thinking, is not just in harnessing the local and the global—which is something in and of itself—but in providing tailored, focused access to that breadth. The value of discovery is much more than accessing the torrent of the Mississippi River as it dumps into the Gulf of Mexico. It is being able to tap into all the right tributaries out of the thousands that feed into the sea. Through careful use of user data and customizations (almost inevitably on the local side), libraries will be able to better serve their patrons in their quest for the right breadth and depth of information to meet their needs. The library of the near future will partake of communal data repositories, but do so in their own way.

## NOTES

1. I will use *federated search* broadly, as a synonym for *metasearch*, ignoring the technical differences between the two. *Federated search* generally means searching across multiple databases at different network and physical locations, while *metasearch* generally means searching across multiple databases within a single network and physical location.
2. For more information, see "Scholars Portal: About," Ontario Council of University Libraries, http://spotdocs.scholarsportal.info/display/sp/About.

3. For more details on the discovery service landscape, see Marshall Breeding's library technology report, *Discovery Product Functionality* (ALA Editions, 2014).
4. Lorcan Dempsey, "In the Flow," June 24, 2005, *Lorcan Dempsey's Weblog*, http://orweblog.oclc.org/archives/000688.html
5. See my article "Project Lefty: More Bang for the Search Query," www.infotoday.com/cilmag/apr10/Varnum.shtml, for a full description.
6. See Cody Hanson's 2013 presentation on "Why Web-Scale Means the End of Build vs. Buy" for more on what local interfaces can offer a library: www.slideshare.net/BaltimoreNISO/hanson-niso-nov-20-vc.