Web Scale Discovery, a white paper

Web scale discovery is a service that provides a single search of a range of preharvested and indexed content. This service connects resources that would currently have to be searched separately, providing relevancy-ranked results in a faceted interface allowing further refinement of search results. As the library continues to accumulate a diversity of information resources via a range of providers and at the same time is mindful of the need to provide academic value while making sure resources are being used, improving the search experience is important.

The Problem

A library user seeking information on the topic of “effective communication in nursing” would have the following options to conduct a reasonably complete literature search:

1. Find books, e-books and media in the library’s online catalog
2. Search Academic Search Complete for journal articles from over 8500 publications
3. Search ProQuest Central for journal articles from over 14,000 publications
4. Search CINAHL for journal articles from over 600 publications
5. Search Communication & Mass Media Complete for journal articles from over 200 publications.

However, we know that many people will initially search Google for their topic, which yields nearly 92,000 results, few (if any) from the scholarly literature. Training our users to complete an effective search of our dispersed resources presents a challenge. As we acquire additional collections of ebooks, transition the reference collection to online format, and create other digital collections it becomes increasingly important to provide comprehensive access.

Literature Review

Vaughan (2011) defines and describes web scale discovery and provides an overall rationale for its use, saying that the growth in the complexity and variety of our physical and licensed resources make a simple search next to impossible with current technology.

As a result, the library, or systems supported and maintained by the library, is often not the first stop for research – or worse, not a stop at all. Users have defected, and research continues to illustrate this fact. People do not just use information that is easy to find; they even use information that they know to be of poor quality and less reliable—so long as it requires little effort to find—rather than using information they know to be of high quality and reliable, though harder to find.

Luther & Kelly (2011) give a cogent overview of web scale discovery and make the important point that while these tools greatly assist the undergraduate user’s navigation of unfamiliar resources, it also meets the needs of researchers, particularly those embarked upon interdisciplinary pursuits where they may be unfamiliar with some resources. In general, they say, these tools leverage a library’s investment in its resources by increasing overall use and revealing content that may be hidden through a piecemeal search.

Breeding (2010) calls upon libraries to adopt web scale discovery in order to help users “concentrate on selecting and evaluating the resources returned rather than struggling through the search tools that the
library provides.” Though enthusiastic about the promise of such systems, Breeding also presents a critique of discovery, saying that there are discontinuities between discovery products and the sets of materials that they index, that some publishers are hesitant to make their materials available through interfaces not under their immediate control, and that end users get in return for ease of access a set of results that may not be as exhaustive as produced using previous methods (2012).

Burke (2010) describes a “value gap” which is a discontinuity between the increasing size and complexity of a library’s holdings and navigational barriers that prevent effective use of these resources. Among the barriers she cites to full use of the library are:

1. No clear and compelling starting place for research – instead, a confusing set of catalogs, databases, and resource lists,
2. Difficulty identifying appropriate resources – database names are often poor clues to their contents,
3. General lack of awareness of resources. (3)

Burke says that “disintermediation,” or the removal of intermediaries (e.g. libraries) in the information supply chain is the new reality driven by today’s end user (1).

Cox & Jantti (2012) explore the correlation between the use of library resources and student achievement. Through analysis of EZProxy logs they are able to visualize a clear positive relationship, thus adding a significant demonstration of the value of these resources to student success. Thus, a web-scale discovery system that improves library resource utilization would also have a direct effect on student success.

A thorough post hoc review of web scale discovery is not yet present in the literature except in the form of empirical reports of increased use of resources. The jury is still out on whether implementation of web scale discovery translates into any qualitative improvements in student work. This may indicate a limit, common to examinations of many library services, to what we can actually learn of the end effects of what we provide.

The Web Scale Discovery Experience

Gross & Sheridan (2010) describes their library’s experience using Serials Solutions Summon. They focus on three questions:

1. Did the students find the discovery searching platform easy to use?
2. Did the new interface provide smooth navigation?
3. Did students obtain satisfactory results across a range of typical search topics?

Through focus groups and usability studies, the research conducted indicated that webpage redesign that focused users on the discovery search box did succeed in stimulating use of the tool. They reported that navigation to resources was improved. They said that the last objective was the most problematic to resolve because it involved how the users interpreted and made use of their results. Overall, they said that the final question reveals what is perhaps the signal issue with web scale discovery: that the easy search and copious results do not themselves improve the user’s awareness of where the information is coming from or its relative appropriateness to their work.
Way (2010) investigates the effect of using a web scale discovery tool (Summon) on the use of library resources over the course of an academic semester through comparison of COUNTER reports. These measures were cross-checked with Google Analytics so that access via the Summon interface could be tracked. His findings indicate a dramatic increase in full-text database use after implementing web scale discovery. Likewise, aggregator database use increased, along with journal collections (e.g. JSTOR). Way reports that their implementation of discovery did drive increased usage of licensed resources as well as library monographs. He concluded by saying more research is needed to assess, qualitatively, whether discovery improves search results.

**Pros & Cons**

<table>
<thead>
<tr>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simplified search needing less explanation</td>
<td>Not a panacea. Users still need information literacy education to understand information types, sources, and appropriateness.</td>
</tr>
<tr>
<td>Search all content, one place. Improved access for both on-site and off-site users. Inclusion of future digital repository content</td>
<td>Vendor bias toward groups of publications (Discovery owned by publishers)</td>
</tr>
<tr>
<td>Single relevance ranking</td>
<td>Common search interface may lack capabilities found in subject-specific databases</td>
</tr>
<tr>
<td>Fast</td>
<td>Single search may promote user unawareness of information diversity.</td>
</tr>
<tr>
<td>No local hardware or software, vendor maintained</td>
<td>May exclude some existing resources</td>
</tr>
<tr>
<td>Increased traffic to existing licensed and purchased resources</td>
<td>Cost</td>
</tr>
<tr>
<td>Increased user awareness of library resources</td>
<td></td>
</tr>
</tbody>
</table>

**Discovery Providers**


**Ebsco Discovery Service** – [http://www.ebscohost.com/discovery/about](http://www.ebscohost.com/discovery/about)


**Worldcat Local** - [http://www.oclc.org/worldcatlocal/overview/default.htm](http://www.oclc.org/worldcatlocal/overview/default.htm)

Summon and Ebsco Discovery Service are provided by database/aggregator publishers and serve, in part and despite their sales descriptions, as vehicles propelling users towards their products. Primo Central is a product of Ex Libris, our current ILS vendor. Worldcat Local has, perhaps, the least commercial skin in the game as an OCLC product. At the same time, Worldcat Local probably suffers because of its distance from publisher content by offering less to users.

**Conclusions**

The purpose of this paper has been to provide an explanation of web scale discovery and some issues surrounding its use. There is a clear case made for its use in simplifying the information seeking process and providing a search interface that roughly parallels the Google experience. The services examined improve on the Google search by providing results that can be filtered, and providing a search result
superior to that from Google (for undergraduate research) because the results are from owned or licensed library content.

Web-scale discovery systems are not perfect and not meant to be a replacement for reference services, library instruction, or the features of subject-specific or higher level databases. All of these currently offered library services remain important. The discovery system helps remote users, distance education users, and those who for whatever reason do not avail themselves of personal interaction at the library with an evolved and more capable ability to obtain relevant search results.

Kurt W. Wagner
Assistant Director, Library Information Systems
David and Lorraine Cheng Library
William Paterson University

References


As with every information resource, it is important to understand the benefits and limitations of each. Google Scholar in many ways resembles the proprietary databases, yet has important differences. Clearly, the attributes of Google that make it a desirable search tool carry over to the “scholar” search, including:

- “Cited by” search results
- Quick link to full text when you search for an article title
- Links directly to Google Books
- Searches full text for your keywords

Some limitations of Google Scholar include:

- No peer-reviewed articles limit
- No limit to journal article
- Inclusion of some unpublished items
- Student uploads to university websites sometimes are returned as search results

Google Scholar does not use the same processes to ensure accuracy in author names, title and publication information as do the vendor-supplied databases. Common mistakes include:

- Missing or wrong authors
- Missing publication name
- Duplicate records for the same item
- Misspelled article information

Depending on the topic searched, availability of full text results will vary widely.

- No limit to full-text results
- Full text links or links to local database not 100% accurate

Google’s single search box gives the appearance of ease, but it makes it much harder to build a complex search. You can build a complex search in Google Scholar, but it will look like an algebra problem.

- Parentheses group words for a concept
- Quotation marks create phrases
- Operators limit how a word is found

For example, you could have a search that looks like this:

**author:** Tomlinson "differentiated instruction" (elementary OR "middle school")

In their forthcoming article in College & Research Libraries, Asher, Duke and Wilson (2013) discuss comparative search effectiveness between web-scale discovery tools (Ebsco Discovery Service and Summon), Google Scholar, and conventional library databases. They cite the strengths (and limitations) of each, finding no clear “winner.” Their conclusion centers on a discussion of how the different search interfaces “exerts a form of epistemological power” over the result that is usually not entirely perceptible to the user. They emphasize the power of thoughtful configuration of a discovery interface’s search
settings so as to leverage searches toward the databases deemed appropriate by the library. They say that this ability to configure is the chief benefit of these discovery interfaces.

**REFERENCES**

