Federated Search and Discovery Tools

Gayatri Vastrad **Jaya Bharathy** **P Dharani Kumar**

Abstract

The article describes how federated search helps the users to find resources across library collections. Federated search is essen-tially an umbrella search often utilizing, as far as possible, the search facilities already offered by each library resource. Federated searching, also known as meta searching, broadcast searching, cross searching, and a variety of other names, is the ability to search multiple information resources from a single interface and return an integrated set of results. These targets include the library's collection of books, software, databases and online public access catalogs.

Keywords: Federated Search, Discovery Tool, XML, Z39.50, OPAC

1. Introduction

Internet content is considerably more diverse and the volume certainly much larger than commonly understood. In the early days of the Internet, it was reasonably easy to find information or data files using a variety of software that were usually command driven. In order to overcome the lack of retrieval facilities, a number of search engines came which prove to be of great assistance in allowing the information seeker to quickly find the piece of data that they require. But, the search engines all have their own shortcomings.[1]

Today's libraries offer more information than ever before. From traditional books to categorized websites to online databases, students and teachers can access thousands of resources with just a few keystrokes. Federated Searching (also known as meta-searching or cross-database searching) is a technology that allows users to search many networked information sources from one interface. When students conduct a federated search, they simply enter a word or phrase into the computer and receive results from multiple targets. These targets include the library's collection of books, software, databases and online public access catalogs. Federated searching allows users to search across a number of information resources simultaneously.[2]

2. What is Federated Search Technology?

Terminology found in current library automation literature associated with single-search interface technology includes the following:

- Broadcast searching
- Consolidated searching
- Cross-database searching
- Distributed searching

- Integrated searching
- Metasearch searching
- Portal searching
- Federated searching

Within the world of computer technologists, these terms have meanings that differentiate one technology from another. When referenced for library automation systems, these terms imply the ability to search multiple databases through a system's interface. Federated searching in this sense primarily utilizes the Z39.50 protocol in order to search local and remote databases. Paul Miller, in his 1999 article about Z39.50, states, "[federated searching] is designed to enable communication between computer systems such as those used to manage library catalogues." Some vendors are working at adding XML (eXtensible Markup Language) to Z39.50 in order to take advantage of existing Web clients, protocols, and tools. Dorman (2003) explains, in easy-to-understand language, the work being done in this area. As it is a technology in its infancy, library media specialists and technology coordinators should see rapid improvements to federated searching as it exists today. [3]

Federated search is the simultaneous search of multiple online databases or web resources and is an emerging feature of automated, web-based library and information retrieval systems. Library federated search (or metasearch) engines have become increasingly popular for libraries to improve services. The benefits to the library user are obvious – not having to repeat searches across several different resources, by carrying out a simultaneous search and receiving blended results. Although aspects of this kind of shared searching has existed for some time (especially with Z39.50 catalogue searching), the explosion of online content and the rise of Google as the dominant web-based search tool has made the development of this kind of searching more important than ever.[4]

3. Purpose

In recent years, libraries of all types have been spending an increasing portion of their budgets on electronic resources. These resources provide access to a rich body of information via an array of disparate packages: multidisciplinary databases, subject-specific databases, journal packages from major publishers, individual electronic journal subscriptions, electronic books, and online reference works. To address the growing complexity of online research, many libraries have implemented federated search products. Federated search provides a single search interface that allows users to search multiple online resources simultaneously—subscription databases, library catalogs, and other electronic resources—with one query, that returns one list of results.

Federated search helps users find resources across library collections. Federated search applications excel at finding scientific, technical, and legal documents whether they live in free public sites or in subscription sites. This makes federated search a vital technology for students and professional researchers. For this reason, many libraries and corporate research departments provide federated search applications to their students and staff. The user does not need to know of the existence of particular databases, nor does he/she need to know how each individual search interface works. In the basic search mode, federated search provides a single search box offering a Google-like experience.[5]

4. Common Federated Search Features

Today's federated search products are more robust than those of three to four years ago. Some of the features now commonly available include:

- LIMITERS. In the beginning, these were restricted to keyword and a few other choices. Today, many more are available and include subject, keyword/descriptor, author, title, date (range), full text, peer reviewed, and format (book, article, image, and audio).
- SIMPLE AND ADVANCED SEARCH. Simple search is more like Google's one-box search, usually the default, while the advanced search allows for further options such as the setting of limiters.
- CLUSTERING. With this feature, results can now be grouped together by subject. So, Jaguar, the animal, would be grouped separately from Jaguar, the car.
- VISUAL SEARCH INTERFACE. This gives the user the added option of viewing results through a visual interface, similar to AquaBrowser.
- FACETED RESULTS. Conceptually similar to clustering, this groups results by source (database, OPAC), subject, or format/type (article, book, image, audio).
- RSS FEEDS/SEARCH ALERTS. These tools allow users to follow the same search over a period of time and keep track of any changes or updates.[6]

5. Federated search: A new technology

Federated search technologies can be divided into two major categories: cross search which searches distributed sources simultaneously and presents the results in a common results interface; and harvested search, which retrieves the contents of multiple distributed databases, normalizes the records, and stores them in a large union index against which the searches are then run. There are benefits and drawbacks to each approach.

I. Cross-search - Cross-search engines are those which search distributed targets on the fly, and return the results to a common search interface. Cross-search is also known as cross-database searching, parallel searching, and broadcast searching.

Connectors. Each target database requires a "connector". The connector tells the search engine how to request results from a given source, and how to interpret those results for display. There are three broad types of connectors:

(1) XML Gateways – increasingly available in large commercial databases, XML gateways are fairly stable and return results quickly. Major standards for XML Gateways include SRU/SRW, OpenSearch, and MetaSearch XML Gateway (MXG).

(2) Z39.50 – library catalogues and other library-specific technologies often expose themselves to federated search engines through a Z39.50 server. These search targets tend to be quite stable, and the connectors rarely require updating.

(3) Screen scraping – Used when the target database does not support any of the other search protocols. Relevant elements are parsed out of the HTML code underlying the native interface. Connectors that use screen scraping are very unstable, and have to be adjusted with every modification of the target database interface.

Presenting results in cross-search. Cross search engines receive results from a number of different sources. There are various ways of presenting these results:

- Fastest first. Because some databases return results more slowly than others, a fastest first configuration will make sure that your user sees some result within a reasonable timeframe. The drawback to fastest first is that the fastest database will not necessarily be the most relevant source.
- Relevancy ranking. Federated search has very limited information with which to perform its ranking. The engines must determine relevancy using only the words that appear in citation fields like document title, journal title, or abstract. Often, the search word does not even appear in the citation. Native sources have access to the full text of their articles, so they can rank much more precisely. Some federated search systems do not perform ranking at all, but simply return documents as ranked by the source.
- De-duping. For federated search engines, true de-duplication is virtually impossible. In order to de-dupe, the engine would have to download all search results and compare them. Because databases return results 10 or 20 records at a time, completing a true de-dupe operation on 50,000 hits would take hours. Cross search vendors usually just de-dupe the first result set returned by the search.
- Clustering. Clustering algorithms look for similar words and phrases in the citations returned by a search, and attempt to group documents that have many words and phrases in common. Clustering software was originally developed to operate against full text documents, but federated search engines try to create relevant clusters out of the smaller number of words available in citations. Relevance ranking and clustering are not mutually exclusive features, both the hit list and the clusters should be ordered according to relevance.
- Faceting. Faceted navigation organizes results according to common metadata attributes like author, publication date, journal, or other citation elements. It provides a powerful discovery tool, allowing users to drill-down into the results and focus a query more precisely. Faceted navigation works best against highly structured data, so is well suited for cataloguing records

and citations. Faceting in the federated search environment is subject to the usual limitations: the sparseness of data in a citation, and the small number of citations returned at one time.

II. Harvested/union indexes. The second major approach to federated search is to harvest all of the relevant sources of data, normalize them into a single metadata schema, and index all of them together in one large union index. This approach offers huge advantages in speed and in the logic that can be applied to the presentation and sorting of results. In most cases a harvested/union index solution will require the provider to download the metadata records at a minimum, and ideally the full text documents too. There is a fair amount of expense involved in maintaining the hardware, software, and network infrastructure to support the frequent harvesting of large record and document sets from many sources. Each data source requires a unique parsing routine to extract and normalize the data, and indexes must be constantly updated and optimized. More challenging than the technical obstacles are the legal aspects of data harvesting, particularly rights management. It would be very difficult for a single library to negotiate the right to harvest data from each of its licensed databases providers, particularly full-text data. Federated search solutions based on this model tend to be developed by commercial vendors or large library consortia.

Harvesting Protocols. There are several major harvesting standards currently in play:

- OAI-PMH – commonly implemented in digital repositories, Open Archives Initiative Protocol for Metadata Harvesting allows a service provider to send an http request to a data provider, who returns an XML-encoded data stream containing the metadata for a specific collection of objects or articles.
- METS – similar to OAI-PMH in purpose and function, METS supports XML-encoded metadata harvesting, but unlike OAI-PMH, METS can harvest both metadata and object.
- LOCKSS – LOCKSS collects content by crawling a web site and downloading each page it finds there. The data provider must implement a "manifest" page for each collection, explicitly granting permission for LOCKSS crawlers to visit. The owner of the LOCKSS box maintains "plugins" that instruct the LOCKSS software how to crawl and audit content from each provider.
- Custom formats – a vendor might prefer to supply metadata or content in an agreed-on custom XML or delimited text format.

Presenting results from harvested/union indexes:

- Speed – a single index will return results much more quickly than multiple disparate indexes, particularly if the index is hosted on a local server so that Internet latency is not a factor.
- De-duping – true de-duping is possible in this environment, because the index already contains all of the results sets. It can compare and de-dupe them very quickly, as it does not have to wait for the results to be returned from each source in sets of 20 hits at a time.
- Relevancy – results can be relevancy ranked with much more granularity and accuracy if they are stored in a single large index, particularly if that index has access to the full text of structured

documents. The organization maintaining the index will also have access to tweak the ranking algorithms and customize the way in which results are returned.

- Clustering and faceting – both of these features are much easier to implement against a single large data store, because all of the necessary information has already been parsed out, normalized, and indexed appropriately. In a distributed (cross-search) solution the data has to be normalized and indexed on the fly. A full text data store will improve the engine's ability to assess similarities, although citation information can also be used.[7]

6. Discovery Tools

A discovery tool is often referred to as a stand-alone OPAC, a discovery layer, a discovery layer interface, an OPAC replacement, or the next generation catalog (NGC). Unlike the front end of an integrated library system or ILS OPAC, a discovery tool is defined as a third party component whose purpose is to "provide search and discovery functionality and may include features such as relevance ranking, spell checking, tagging, enhanced content, search facets" (OLE Project, 2009). Discovery tools should not be confused with federated search products. The former "promise to provide a single interface to multiple resources based on using a centralized consolidated index to provide faster and better search results", while the latter search remotely, rely on connectors, and provide "only partial and limited solutions" (Hane, 2009). In addition, a federated search tool usually requires user logon and works in a protected environment, while a discovery layer is open to the public. A federated search tool is dedicated to finding articles across a number of subscribed databases and as such is not within the scope of this paper. Libraries are disappointed with commercial ILS OPACs. Developed as a part of an integrated library system, they have remained relatively static over the years and have not evolved in pace with the discovery and search tools now commonplace at commercial sites such as Amazon. Most of them cannot and will never be able to provide advanced functionalities in order to meet current expectations. It is more practical for vendors and developers to field new OPAC systems that run alongside the older ones than to attempt to alter the proprietary code of ILS OPACs. Most current ILS OPACs do not offer the features of these standalone, next generation catalogs.

Until recently, libraries could do nothing about their outdated OPAC. Proprietary ILS OPACs offered only limited customization. Today, libraries using some of the ILS OPACS can add patches and a limited number of functional improvements by acquiring both free and commercially available plug-ins or add-on modules, but this solution will not completely transform an old OPAC into a next generation catalog. Additionally, libraries may adopt a "Web OPAC wrapper" solution to embed their existing OPAC within another user interface layer (Murray, 2008). The current trend some libraries seem to favor is to simply abandon their current OPAC in favor of one of the new standalone, next-generation discovery tools.[8]

7. Advantages of Federated Searches

- With federated searches, not as many results come up with a specific search related to their topic compared to the abundance of unrelated results with Google. More is not always better. In addition to filling out forms and combining documents from multiple sources, another important benefit of federated search engines is that they search content in real time. Real time data is crucial for researchers who are searching for up-to-the-minute content or for content that changes frequently. As soon as the content owner updates their source, the information is available to the searcher on the very next query.

- It is difficult for most students to choose appropriate, relevant sites from hundreds of thousands of hits. Using a federated search engine can be a huge time saver for researchers. Instead of needing to search many sources, one at a time, the federated search engine performs the many searches on the user’s behalf.

- Targeted searches are usually filtered for quality. Federated search engines show their value best in environments in which the quality of results matters, such as libraries, corporate research environments, and the federal government. In the case of the federal government, the constituents of the government benefit greatly from such applications. A major difference between a federated search engine and a standard search engine like Google is that the client who contracts for the federated search service selects the sources to search. In almost every case, the sources will be authoritative. Google, on the other hand, has very minimal criteria for source selection.

- Federated searches qualify the authenticity of the information. For example, anyone can write a report on a topic and post it on the Internet. That does not mean that information was checked for accuracy. By using this new add-on feature to the school’s library’s automation system, students can better ensure the information they use for their research is accurate. With a federated search engine, the information has been checked and verified by educators and professionals.

- The federated search includes books and other materials that already may exist in the school library. Thus, the federated search engine acts as a helpful librarian does, directing users to excellent quality.[9]

8. Other issues of the federated search

- The federated search has some other issues as well. First, it cannot cover all online library resources. The goal of one-stop shopping cannot be achieved completely by any federated search. There are various reasons for this:

- Some databases do not work with any federated search at all, such as SciFinder Scholar. SciFinder Scholar does not use a web browser but rather requires its own internet client. Neither MetaLib nor WebFeat can cover SciFinder Scholar.

- If databases require a login, they will not work with the federated search.

- Some databases work with one federated search product but do not work with the other. MetaLib cannot search LexisNexis databases because LexisNexis does not allow Z39.50 or XML gateway access. WebFeat cannot search databases that do not have a search box on their front page because WebFeat counts on the search box on the native interface to search.

- Many libraries have databases on a pay-per-search basis, and libraries normally do not want them to be searched by a federated search for budgetary reasons.
- Some databases have a limited number of concurrent users, and if these databases are included in a federated search, the limited seat(s) is/are taken immediately whenever someone logs into the federated search, and no other users can use these databases. Libraries normally do not want to include databases with a very limited number of concurrent users in the federated search.
- It may not make sense to add to a federated search menu the very specialized databases that most general users would not be interested in, or the databases that require special software. One example is Inter-university Consortium for Political and Social Research (ICPSR) that requires statistics software such as SPSS to view data.[10]

9. Access Issues with Federated Search

Verification, authentication, and certification can be difficult for the federated search vendor. Since federated search engines don't hold the data locally, meaning the engines perform the search and send the results back, the federated search engine must be able to access multiple, password-protected databases behind the scenes, all at one time, and show users their results in one easy-to-read interface. The challenge for federated search vendors is to ensure that only licensed users can access databases in an appropriate manner, as specified by their license. This may require a library or a corporation to set up multiple areas where only certain licensed users can access a federated search.

The number of different cookies a subscription database uses makes the authentication process either a simple or complex procedure. All the user needs to provide is the ID, password, and files to be searched for each subscription database. The federated search engine will handle the rest of the authentication procedure. However, the initial setup process can take a number of hours to a number of days, depending on the complexity and number of subscriptions.

10. Interface Issues with Federated Search

The second issue is the search query and results interfaces. For several years now, libraries and corporate information centers have faced the "Google phenomenon." Many patrons believe that doing a Google search covers all the bases. Libraries now have an excellent opportunity to provide a simple, yet powerful interface that out-Googles Google. They can set up their interface based on subject and sources, or customize it to specific user needs. Libraries and corporations need to take note of Google's simple interface—users expect an interface as streamlined as Google's. Uncomplicated and intuitive interfaces without a high learning curve will see expanded usage. Most of the federated search vendors allow clients to create their own "look and feel" for the search interface and results pages. However, if you do not have the staff resources, they will often allow a more static look where little decision making on your part needs to be done.[11]

11. Conclusion

Federated search technology is an integral component of an Information Portal, which provides the interface to diverse information resources. Individual end users will benefit from federated search technology. This blends e-journals, subscription databases, electronic print collections, other digital repositories, and the Internet. Federated searching reduces the time it takes to search and usually displays results in a common format. Most complete federated search solutions support multiple search protocols. Typically they offer integrated OpenURL resolution, spell checking, saved searches, alerts, de-duping, and single click access to the native interface.

References

1. RATHINASABAPATHY, G. (2007) "Invisible web and knowledge discovery tools: a study
2. YOHE, Paula. (2005). Libraries online: competing with search engines. Media & Methods. 41(4)
3. BARBARA, Fiehn. (2004). Federated searching: a viable alternative to web surfing! 11(2) p29-31.
4. BAILEY, Penny. (2008). Recent developments in federated searching. 7(3) p37.
5. http://en.wikipedia.org/wiki/Federated_search
6. ALEXIS, Linoski. and Tine, Walczyk. (2008). Federated Search 101. Library Journal 133. P2-5
7. GIBSON, Ian, Goddard, Lisa and Gordon, Shannon. (2008). One box to search them all: Implementing federated search at an academic library. 27(1) p125-128.
8. YANG, Sharon Q. and Wagner, Kurt. (2010). Evaluating and Comparing Discovery Tools: How Close Are We towards Next Generation Catalog? 28(4) p691.
9. YOHE, Paula. (2005). Libraries Online: Competing with Search Engines. Media & Methods 41(4)
10. CHEN, Xiaotian. (2006). MetaLib, WebFeat, and Google: The strengths and weaknesses of federated search engines compared with Google. 30 (4) p422-423
11. Online (Weston, Conn.) 28 no2 16-19 Mr/Ap 2004 (Reprint of article by Donna Fryer www.SearchitRight.com)

About Authors

Ms. Gayatri Vastrad, Librarian, Wipro Technologies
E-mail: gayatri.vastrad1@wipro.com

Ms. Jaya Bhrarthy, Librarian, Wipro Technologies.
E-mail : jaya.bharathy@wipro.com

Mr. Dharani Kumar P, Assistant Professor, Department of Library and Information Science, Kuvempu University.
E-mail: dharanikumarp@indiatimes.com