January 2007

HELIN Federated Search Task Force Final Report

HELIN Federated Search Task Force

Follow this and additional works at: http://helindigitalcommons.org/task

HELIN Federated Search Task Force
Final Report
January 2007

Committee Members:

Susan McMullen (Chair) (RWU) - smcmullen@rwu.edu
Dusty Haller (CCRI) – dhaller@ccri.edu
Connie Cameron (PC) – ccameron@providence.edu
John Edwards (JWU) – john.edwards@jwu.edu
Tish Brennan (RIC)– tbrennan@ric.edu
Mimi Keefe (URI) – mjkeefe@uri.edu
Olga Verbeek (Salve) – verbeeko@salve.edu
Colleen Anderson (Bryant) – canderso@bryant.edu
TJ Sondermann (Wheaton) - tsonderm@wheatonma.edu

Background:

In Spring 2006, the Federated Search Task Force was formed as a working group of the HELIN Reference Committee. Our charge was to investigate and report on Federated Search Engines. A federated search engine allows users simultaneously to search multiple databases at one time.   The databases to be searched are pre-determined by librarians to correlate with their individual library database subscriptions.   It was generally agreed by consortium members that any product that required HELIN or an individual institution to configure its own server would not be feasible given staffing and budget constraints. Included in the task force charge was an investigation of the advantages and disadvantages of Federated Search Engines and an evaluation of the various subscription or vendor-hosted products currently being marketed.   Our findings from the evaluation process, as well as topics to consider when purchasing a federated search engine are presented below. Based on this report and their individual situation, each library will decide what is best for their users – does Federated Searching add value for your clientele and does it offer a good return on your investment?

Federated Search Demonstrations & Trials:

Our process was not to evaluate every product available, but to find the products that best meet the needs of the HELIN Consortium Libraries. With input from HELIN librarians we decided to evaluate federated search products from four vendors – MultiSearch developed by CSA, Central Search from Serials Solutions, WebFeat Express vended but not developed by Ebsco, and the Innovative products, MetaFind and Research Pro. Evaluation of these products was essentially a two step process: 1. demonstrations by vendors done in Spring and Summer 2006 for basic review and vendor information about the product   2. Performance trials (October 1 – November 15, 2006). For both steps we established evaluation criteria to aid us in comparing features and search results across all products. In developing these criteria, the task force hoped to construct tools that would enable a quick checklist of comparable features and a standardized performance test of all products being evaluated. The forms used to gather information from the vendor demonstrations and the combined responses for each product are found in Appendices 1 and 2 of this report.

After the information gathering phase of our project, we decided to trial CSA MultiSearch, Central Search, WebFeat Express, and Research Pro. When we first began, Innovative was vending both MetaFind and Research Pro. We decided to trial just the Research Pro product since it will eventually replace MetaFind. Although there was some discussion of trialing WebFeat rather than the WebFeat Express product being vended through Ebsco, the group decided against it since the cost was prohibitive to most HELIN institutions ($20,000 plus) and it would require more local technology commitments.

In order to ensure a standardized and consistent test of the search engines, we asked each vendor to configure their product to allow searching of the following 10 databases: *HELIN, JSTOR, ABI/INFORM, PubMed, ERIC, America-History and Life, Gale Literature Resource Center, Academic Search Premier, Providence Journal*, and *Google Scholar*. The databases were chosen in an attempt to include those with keyword vs. controlled vocabulary, multidisciplinary vs. disciplinary, full text vs. citation index and periodical vs. newspaper vs. mixed. Variety of database vendors was also important to see how well each federated search engine searched across different search

platforms. To further standardize this process we constructed a few search queries that could be replicated by all evaluators. The trial evaluation form (found in Appendix 2) and the access instructions, including suggested search queries, were sent out via the HELIN Listserv offering all HELIN librarians the opportunity to participate in this process.

The task force received approximately 30 evaluations, many from GSLIS students, for CSA MultiSearch, Central Search, and WebFeat Express. Because Research Pro was the last tool to become available and was not consistently working throughout the trial, we received only 15 evaluations for that product. For a complete trial analysis of each federated search engine, please see Appendices Three through Six.

After completing our formal trial of commercial products, the group also took a look at Google Scholar and how it might be used as a federated search tool. A comparison of Google Scholar and commercial federated search engines is included within this report.

Highlighted below are topics that the task force considered to be the most important when purchasing a federated search tool for your institution. We have also included a brief analysis of what we found from the individual products under each topic for consideration.

Topics to Consider:

Information Literacy & Search Functionality
Will a Federated Search Engine be a hindrance or an aid in promoting information literacy standards?

Pros: Although library instruction classes are designed to instruct students in finding the most appropriate tools for their research, many students engage in research without the benefit of library instruction. A federated search engine could be useful in helping students quickly find which databases have content on a research topic. Casting a broad net can lead students into the best database for their topic. By librarians pre-selecting the databases to search, they are more likely to find more appropriate resources than could be found with Google.

Cons: Will it keep upper level students from becoming familiar with the information seeking landscape in their discipline? In order to retrieve the best results for their research, students really need to get familiar with using the terms used in subject thesauri for specific databases; i.e. ERIC, PsycINFO. Federated Search engines can only offer a modified, superficial subject search by searching for keywords in a subject field. If that keyword happens to be housed in a different subject index or thesaurus then the student will pick up articles that are only tangential to their original research topic. If the student relies too heavily on the Federated Search Engine rather than the native database, they may never find the right search terms for finding the most relevant citations.

What we found:

Keyword searches in CSA MultiSearch, Central Search and WebFeat Express produced very good results that were comparable and consistent with results found in native databases. Because there is no direct access to each database's thesaurus, the subject search is actually a keyword search of the subject field.

CSA MultiSearch: This is the tool that was most traditional in its approach. It worked very similarly to other CSA Illumina databases. MultiSearch supports Boolean and wildcards and allows users to limit by date range, subject area, latest update, English only, and journal articles only. Fields that can be searched: author title, subject, keyword, and "anywhere."

The subject searching in MultiSearch did not always seem to work reliably. This could have been a consequence of a trial that was not easily established.

Central Search: The database offered a simpler initial interface for its basic search but its advanced search page did include a "search by database" and a "search by subject". Advanced search provided additional search boxes for combining terms and did offer basic Boolean. Fields that can be searched: author, title, subject, keyword, abstract, ISBN, ISSN, any. Restricting by date is allowed. Phrase searching with quotation marks produced more relevant results.

Central Search appeared to do an excellent job of subject searching, retrieving comparable results when using SU Descriptors in ERIC. But it should be noted that a subject search here is not the same as searching a native

database's thesaurus. It appears to pick out keywords contained in the subject term fields in each document.

WebFeat Express: Offered a quick and easy search form and retrieved relevant results with its html scraping technology. For subject searching, it appears WebFeat also searches the keywords contained in the subject term fields.

Research Pro: There were serious issues with its search mechanism. For example, putting the word AND between two concepts in the basic keyword search box produced very strange results. Subject searching did not work as well as others. This product is still under development, so it is a difficult to get a good read on how it will perform in the future.

Google Scholar: Although results were sometimes comparable with what we could find by searching the native database, there was no way of knowing what content or databases were actually being searched. Results were not returned by database in any particular order, so it is difficult to make this comparison. Phrase searching is supported (quotation marks required).

Subject Linking
Related to the topic of information literacy and search functionality is subject linking or clustering. Links to subject descriptors are a very useful feature of most native databases today.

What we found:

CSA MultiSearch – provides linked subject descriptors. They are clearly accessible at the right side of the citation. However, many noted that the linking seemed sporadic and only specific to certain databases. When available, the subject descriptors were very useful in helping to refine a search.

Central Search – Clustered subject headings help the user refine or narrow a search – however, clustered subject headings were not always available or relevant for every search.

WebFeat – no subject linking or clustering is provided.

Research Pro – ??

Google Scholar – no subject linking, but does offer a link to related articles, which is sometimes useful.

Customization & Resource Discovery
Does the federated search tool offer customization necessary for students to become familiar with databases for discipline specific research or course related assignments?

Pros: Federated Search engines can be an entry point for leading students to native databases that they may have been previously unaware of. For many undergraduates, federated search engines can be a way of resource discovery. Database clusters can be customized for particular subjects or courses. The search box can also be placed within course management software. Larger universities, i.e. Northwestern University, have noted an increase in database use since using implementing a federated search technology.

Federated Search tools give librarians the ability to create subject or course specific searches. A grouping of databases can be placed on a Research by Subject Web page so that students may quickly search across a wide spectrum of the databases in a particular field. However, this does not help the researcher who is doing cross-disciplinary research. Of course, a cross-disciplinary grouping can be made, but it is difficult to know which databases to include in such a grouping.

Cons: Can customization for our users be achieved better with the use of course management software or course guides available from library web pages that faculty point to from their syllabus. Does the course management software used by your institution lead students to resources more effectively than a customized federated search engine page? With course management software, librarians can work together with faculty to design a course site that includes a library module with links to appropriate databases for completing assignments. Federated search software does allow for customization within the courseware -- is it cost effective when you are only recommending 3 or 4 databases?

What we found: All of the products that we evaluated offered customization features. WebFeat Express offered fewer options for customization because it is a watered down version of the full WebFeat product.

Ease of use and Search Screen Appearance
Federated Search engines can be an alternative to Google but if the search engine interface presents a long list of databases to students to select from, it loses its simplicity. Students are so overwhelmed with information, we need to ensure that the search tool has a customizable interface and does not lead to more user frustration.

Librarians will have to decide the best way to configure their interface based on the needs of their clientele. The most simple search interface is a single search box that searches all of the databases available. Other choices would be to provide a long list of database for users to choose from, a short list of best bets, or allow users to search in a particular subject area.

What we found:

CSA MultiSearch: The search interface was very similar to that used by other CSA Illumina products -- perhaps too similar. Will patrons know that this particular tool is different from an individual database? Users could not readily see that they were searching multiple databases. Drop down menus and links were not intuitive with many features being overlooked by the casual or first-time user. This search interface seemed more librarian oriented rather than student oriented. However, many evaluators reported that the quick search screen is very simple and easy to use, especially for users who are accustomed to the Illumina interface.

Central Search: Offered a simple search box, however it did default to a title search. Central Search believes by defaulting to a title rather than a keyword search in basic search, users will retrieve more relevant results. Librarians are not so sure, but can set the default search to whatever they want it to be. Most evaluators found that the Central Search interface was easy to understand and maneuver. Although there was a long list of databases presented for the trial, it was not so intimidating that people would ignore their other options (for advanced search or subject search) entirely.

WebFeat Express: Many evaluators reported that the search screen had a clean simple design. Some felt that this tool offered too much information on the databases up front. The search button at the bottom of the screen forced users to select which databases to search before executing the search. Having a full-text limiter available from the initial search screen was a plus. However, with WebFeat Express, librarians will only get three search interfaces to choose from with little flexibility for customization.

Research Pro: This interface was more confusing to the user. Buttons did not seem to be placed strategically, i.e. the advanced search button was placed off to the side, not near the search box. Too much information and too many choices to make.

Google Scholar: Students are already familiar with the Google screen, so it will be easy to transition to Google Scholar. However, to get the full benefits from using Google Scholar, such as linking with the library's subscriptions and exporting to Refworks, students will need to know enough to set their Scholar preferences

Export features
As students become more familiar with data capturing and using bibliographic management tools, such as Refworks, they will be expecting this from all databases/search engines -- any Federated Search engine that does not offer this feature will be severely limiting to students.

What we found:

CSA MultiSearch: Good export features -- print, email, and Refworks.

Central Search: Cannot export results to email. Exporting full-text results is available only after linking into the native database and using that database's export functions. No export to Refworks at this time. Sales representative reports that this will be coming soon, perhaps in Spring 2007.

WebFeat Express: Good export features -- print, email, and Refworks.

Research Pro: Students can mark articles for export but cannot export to anything! Product is clearly not fully developed in this area.

Google Scholar: Importing into Refworks works very nicely from this free tool! Cannot email results.

Full-text Linking
Successful full text linking to the actual article or to the link in the actual database is essential when using a federated search engines. Some products provided this linking directly, others simply linked to the native database and used the article linking functionality employed by it.

What we Found:

CSA Multi Search: Linking to full text is fully functional, though some slowness was experienced. Full-text linking took the user into the native database.

Central Search: Full text article linking worked very well.

WebFeat: Can link directly to the full text with "view HTML" or "view PDF". It takes you directly to the full text in the native database.

Research Pro: Links directly into the native database for all functionality.

Google Scholar: Links you directly to the full text if your library has set up Google Scholar to recognize your subscriptions.

Speed & Results Display

Vendors seem to make a big deal about the technologies that were driving their products. We did not notice that one technology produced better results or fastest speed over the other. What seemed to really slow things down, was having to wait for all results to be loaded until all databases had been searched.

What we found:

CSA MultiSearch: Speed varied from painfully slow to acceptable. It seemed that the results were returned fairly quickly, but when maneuvering amongst those results, things slowed down and did not always work as expected.

CSA offered a clear merged results list with databases clearly indicated and linked descriptors when available. You can edit your original search, but there did not seem a way to limit from your result list. Results can be sorted by most recent first, relevance, title, author, author/title, and database.

Duplicates are identified, both records are displayed so the user can select the most useful one.

Central Search: Evaluators noted speeds that were from "speedy" to "slow taking almost a minute to return results." Response time is slowed down because it appears that the user has to wait for all databases to be searched before results are displayed. However, in actuality you can click on a finished database search while others are still searching. If you do click on a database while others are still searching, you will not retrieve subject clusters and the aggregated results list is not ever available unless you redo the search.

Duplicates are clearly marked and results can be sorted by date, title, author, and source. Especially useful was the ability to show or hide the abstracts. More results are displayed on a page, page is less cluttered, and the abstract is immediately available if needed.

WebFeat: Hands down, WebFeat Express was the speediest of the federated search engine. Actually rivaled Google Scholar.

Research Pro: Slow to return results with a keyword search. Evaluators liked best Research Pro's tabs at the side of the results list indicating the results for each database. Only the first 10 results from each database are returned. No way to refine a search. No sort options available at this time.

Google Scholar: Speedy but not sure what content is being searched. Results cannot be sorted. Duplicate citations are grouped together as a note on the first citation.

Customer Service:

CSA MultiSearch: The trial was very difficult to set-up and the vendor required that we do a lot of leg work in providing passwords for access etc. It was a disappointing trial process. However, their product developer assured us that none of these access problems with the proxy server would be an issue if we were to purchase the product. CSA provided the best vendor demonstration from a product development standpoint. However, there seemed to be very little technical support for this product as of yet.

Time to Launch: Vendor estimates one week. Site administrator does most of the work.

Libraries currently using MultiSearch- Penn State, Fairfield University, James Madison, Eastern Michigan

Central Search: Good customer service was provided. Demonstration by Jeff was excellent.

Time to Launch: Naval War College indicated that it took about 8 weeks for Central Search to make all of the connections.

Libraries currently using Central Search - Naval War College, Marist College, NOBLE

WebFeat Express: Although the demonstration was disappointing because the sales representative was asked to demo a product for which she had not received training or support. However, she provided very good feedback and was very responsive throughout the trial. The trial worked beautifully and we received excellent customer service from WebFeat.

Time to Launch: Vendor indicates 1-2 days

Libraries currently using WebFeat Express – Lindenwood University, Great Basin College

Research Pro: Customer Service was not as good as the others. The product was inferior and the cost was higher.

Time to Launch: Vendor indicates that it will take 90 days to set up this product for an individual library. This product is not ready to launch and is still being developed.

Libraries currently using Research Pro -

Google Scholar: No customer service but quick to set-up and FREE.

The Google Scholar Factor:

If the purpose of supplying a federated search tool to searchers consists in allowing greater access to a wide array of high quality, highly relevant academic material including both books and articles, Google Scholar holds up very well as a direct competitor to any of the federated search tools trialed and did significantly better with some searches. Directly comparing the search <"at risk students" and computers> among all five tools, the "Search All Databases" functions brought back widely differing pools of results. Equally relevant were the first 25 hits from Google Scholar, MultiSearch, Central Search and WebFeat. Research Pro was not considered in this direct evaluation as it was difficult to maintain consistent enough access to the tool. When material retrieved from ERIC alone was considered, both MultiSearch and Central Search produced identical result lists, with WebFeat only differing in a minor way (49 vs. 51 hits). By back searching by author, it was determined that, of the first 25 of 51 ERIC hits, Google Scholar included all 25 in its result set. In addition, Google Scholar tapped into pre-print items and the full-text volumes

contained in Google Books. On the other hand, of the 12 hits from Academic Search retrieved by MultiSearch and Central Search, only half were contained in the Google Scholar result set. Those excluded were clearly more journalistic in nature.

If the reason for supplying a federated search tool were to give access to a multidisciplinary array of general serial content, then Academic Search fills that function very well as is. If cross-database searching, especially within the broad discipline areas – humanities, social sciences, sciences – are the goal, most of our current vendors support this function. At Rhode Island College, for example, virtually all significant social science indices are within the CSA universe and easily cross-searchable.

Conclusion:

Although federated search engines have evolved considerably in the past couple of years, many HELIN evaluators felt that none of these products seemed fully developed.   How they will develop in the future still remains unknown. Central Search and WebFeat have the clearest vision for their products and their future development. There is some concern that WebFeat Express is merely sold by EBSCO and they are not vested in that product's development. Though customer support supplied directly by WebFeat was good during the trial, what support will EBSCO be willing to supply for a product that is not their own. Innovative customer support was inadequate. They did not seem to have a clear direction about the development of Research Pro. As can be seen from the trial, Research Pro was clearly lacking in many of the features offered by the others. Though the charge of this task force was not to make a recommendation of one federated search engine over another, Serial Solution's Central Search, CSA's MultiSearch, and Webfeat Express vended by EBSCO provided comparable functionality in most categories. Though costs were significantly different depending on individual library needs and budgetary resources, any of these three products could serve them well enough. The advantages and disadvantages of each product can be found in Appendices Three through Six.

A number of other important aspects need to be considered. For example, how do applications such as Google Scholar (and its competitors such as Microsoft's Live Academic search) affect future need for federated search engines? How much time will be needed -- and how much time will be available -- to educate your research communities in the use of federated search tools?   That being said, all of these products are vendor hosted and subscription based, so a decision to purchase doesn't lock you in forever.