Out-googling Google: Federated Searching and the Single Search Box

Verne W. Newton and Kathryn Silberger

Introduction

Students may not be voracious book readers, but that doesn't mean they aren't voracious consumers of information. Indeed, by the time they arrive on campus it borders on a habit conveniently spoon-fed by Google. Students don't approach libraries as an alternative source of information, let alone a warehouse of books. In most cases it is just a place with a lot of public computers bunched together in a much more pleasant environment than your average computer lab or dormitory room. At Marist, like most colleges, the library is, or at least was, a place where a clash of civilizations played out.

Marist College, located in Poughkeepsie, New York, is a medium size liberal arts college with about 5,000 students on its main campus. It has four extension sites and several online degree programs. The library has about 200,000 print volumes in the stacks, over 34,000 electronic journal titles, 91 databases on the A–Z list, and hundreds of faculty recommended web sites on the subject guide web pages. The Library staff includes nine full time librarians.

Several years ago librarians could try to convert students from their diet of internet offerings, which we tended to equate with intellectual junk food, to our more wholesome information resources. But that is no longer a defensible position. Google has improved qualitatively and quantitatively and provides some dishes that even scholars cannot—and should not—resist.

At Marist, instead of focusing on the shortcomings of student preferences, we looked more rigorously at our own. Database names, for instance, are often opaque, providing no clue as to why one should be searched rather than another. The interfaces were cluttered with too many boxes and too many explanations that were often highly idiosyncratic and frequently counter-intuitive. Having never searched through a series of annual volumes of Social Sciences Index, students don't see something superior, they see something that is cumbersome and confusing. They could either rely on us to escort them into this exotic environment, or they could politely turn down our offer and return to a world they had mastered. Overwhelmingly, they did the latter.

Verne W. Newton is Director of Library Services, James A. Cannavino Library, Marist College, Poughkeepsie, New York, email: Verne.Newton@marist.edu. Kathryn Silberger is Automation Resources Librarian, James A. Cannavino Library, Marist College, Poughkeepsie, New York, email: Kathryn.Silberger@marist.edu.

Why would they want to come into our world when theirs seemed so much better?

Web Site Redesign

One of the first things the new Marist library director did in 2003 was to coordinate the redesign of the library's web site. By then our A–Z database list had grown to 58 titles and the breakdown ran to several pages. Nonetheless, we still saw students searching *ABI/Inform* for everything from archaeology to zoology simply because it was the first database on the page.

To simplify navigation, not just for students, but for faculty as well, the redesigned web site provided each of the college's majors with a single uncluttered page providing access to all relevant subject resources in electronic format. At the same time we exchanged broad Library of Congress subject groupings for the nomenclature of the registrar, far more familiar to students and the college community. This resulted in 27 subject pages each with links to 1) research databases (mainly subscription-based); 2) faculty recommended websites; 3) top scholarly journals in the field with links to more journal listings; and 4) relevant electronic reference resources.

Log statistics revealed an immediate impact. Students showed a measured preference for using the subject pages as a gateway to our databases over the A–Z list. What we could not measure, though, was how often students were turning to an Internet Search engine, especially Google. Our interactions with students indicated a significant portion of our students still used Google, Yahoo, Alta Vista, and other search engines to find information. This supposition was confirmed in the OCLC's 2005 report, "College Students' Perceptions of Libraries and Information Resources": 89 percent of college students were going to a search engine to search for information and only two percent were starting that search with the library's web site.[1] It also confirmed the obvious: Google was the search engine of choice.

Moreover, we believed there was another truism. Students and faculty were discouraged from using proprietary databases not because what they were looking for was not there, but because it was there in overwhelming abundance. While librarians loved the increasing number of options and enhancements offered by the various database vendors, patrons were being overwhelmed by the complexity. In a sense the more database content we added, the more we were driving our patrons to Google.

This led the library director to ask, still in 2003: why can't we have a single Google-style search box that cuts across multiple databases simultaneously? At the time federated searching was more of a concept than reality, and the products on the market place were still in a beta testing phase, regardless of how they were being marketed. With adequate development it was clear that in the future federated search tools would offer a way to combine the quality information of subscription databases with a single simple search interface.

Soon we were exploring various federated search products. We attended seminars and conferences, met with vendors, talked to the larger libraries who were already deeply invested in finding or implementing products, road tested several ourselves, and at least once came close to writing a check. Yet none that we tested or examined quite met our requirements or budget.

Implementing the Federated Search

Finally, in 2005 Serials Solutions introduced Central Search. We were impressed with the product, in part because it was a hosted service and thus did not require hardware, local installation and customization programming, training, and other costs that most of the other products we examined did. Furthermore, Serials Solutions offered extensive local customization, and it could meet our chief requirement: a simple single search box for each of our subject pages, which would search only the databases relevant to that subject.

We worked closely with them over the summer towards the goal of launching Central Search in the fall 2005 semester. Much of the conceptual work was a continuation of the principle of simplification that guided our recent website redesign: the decision to use the college's majors for our subject pages and the distribution of databases accordingly. But we also wanted to end the clash of civilizations, the notion that we were going to force the students to choose between their Google world and our world of subscription data bases.

Including Google

Based on student confidence in the "Google brand," we added a search box slogan: "Search library databases and Google at the same time." Student testing confirmed that including Google was a significant public relations advantage, essential to out-Googling Google. Although this increased student willingness to use it, in practice students began choosing the results from the full text articles in the subscription databases rather than from Google. When asked if they would like to isolate out the Google results, they indicated they liked the results set better than what they normally found on Google.

In addition to Google, and the newer Google Scholar, each search would also simultaneously search the library's catalog and our approximately 25,000 electronic books. We continued to develop and test. In the field of web usability, there is a concept of "progressive disclosure," which tries to provide for initial simplicity, but includes access to more powerful features for those who want it.

1. Initially, show users **only a few** of the most important options.

2. Offer a **larger set** of specialized options upon request. Disclose these secondary features only if a user asks for them, meaning that most users can proceed with their tasks without worrying about this added complexity.[2]

In our design the default search was simple, single search box with the word "advanced" hot-linked in small letters for those looking for more power and control. This "advanced" option linked to a guided search interface with the ability to select specific database titles to be searched (see figure 1).

Fox Hunt—A Marist Name

Because of the high degree of customization involved in implementing Central Search, the library director was anxious to put the Marist brand on it and chose "Fox Hunt," in reference to the college's Red Fox mascot. The library's Web Student Assistant, Gary To, developed an appropriate "Fox Hunt" Flash graphic that helped bring attention to the whole enterprise (see figure 2).

Usage Patterns

In the fall semester of 2005, the Cannavino Library launched Fox Hunt. Because our database offerings were the same in both semesters we were able, at the end of the fall semester, to make reasonable comparisons between fall 2004 and fall 2005. We saw a 13 percent increase in the number of hits to our website but a 29 percent increase in database searches. Most dramatic, however, were the increases in full text articles viewed: we saw an overall increase of 350 percent. We were concerned that something in the interaction between the federated search engine and the vendor database might be skewing the numbers. However, we had great confidence in the number of PDFs viewed. A patron has to click on a link to open up a PDF, so we were certain that the number of PDFs was an accurate reflection. PDF usage increased by 63 percent.

The greatest increases in documents viewed occurred with the newspaper databases. For example, *Newspaper Source* experienced an increase in articles viewed of over 2200 percent. We can speculate that in 2004 usage may have been negatively affected by the alphabetic effect that buries it in the middle of any list. The name is accurately descriptive, but not very catchy. In a federated search environment, the alphabetic effect is mitigated, and the broad range of subjects covered and sorting by date places newspaper articles prominently in the search results. It is also very easy for a patron to click open, read the full text and advance to the next article. Because newspaper articles tend to be short, the effect of

Figure 1: The Fox Hunt simple single search box.

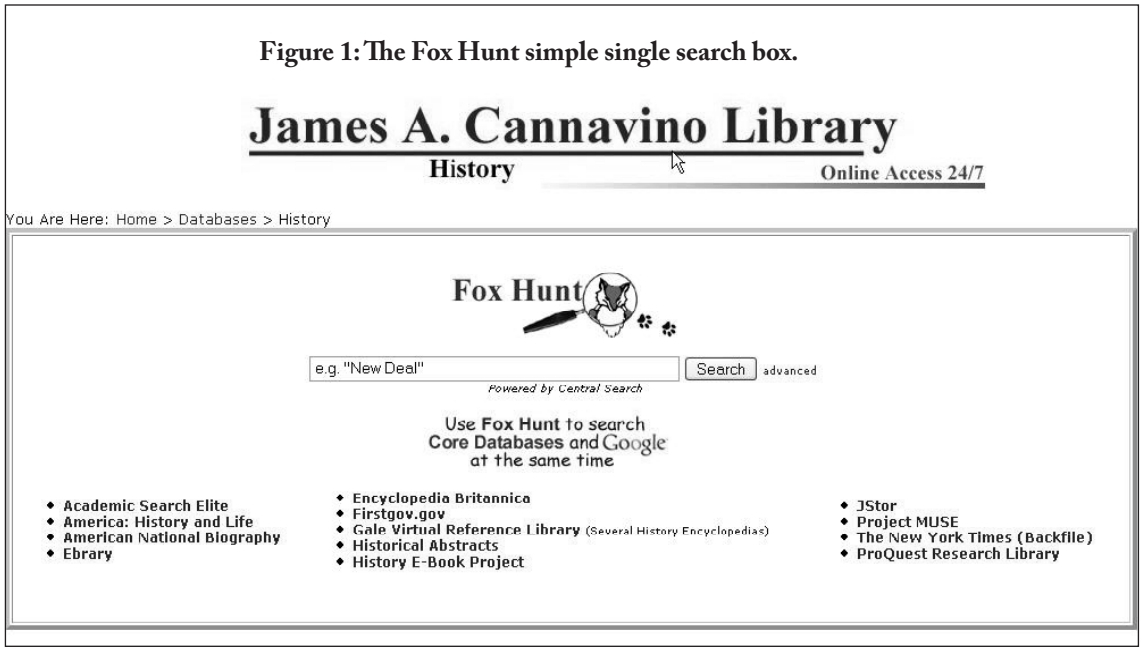
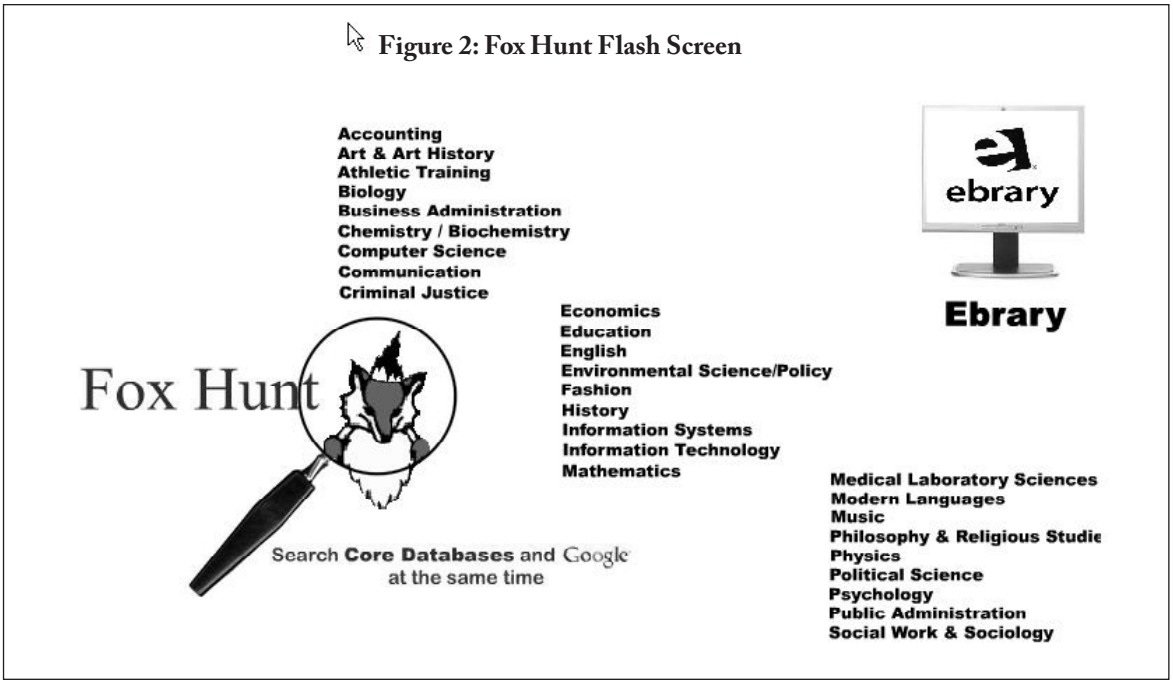


Figure 2: Fox Hunt Flash Screen



that behavior may have been magnified. We were happy when a clustering feature was added, because it tempered the effect of the date sort order.

Very large increases also occurred with databases that were segmented and required more perseverance to navigate effectively. *Business & Company Resource Center* (BCRC) is an example of a segmented database. It includes a directory of companies, a directory of associations, an SIC/NAICS database, investment research reports, company histories, a financials database and an article database. Patrons are required to navigate ten different tabs to see the items specific to the different segments. We can speculate that previously patrons tended to view the information on the first tab only. With federated searching, the most relevant items from that database were easily visible and accessible to the patron.

There was also an impact on databases that were not compatible with federated search technology. One major database in this category saw a 14 percent decrease in articles viewed.

The impact on scholarly databases was also positive. We saw increases greater than 50 percent in documents viewed in databases such as *Science Direct* and *Jstor*. Major-specific databases showed similar increases.

One of the questions we asked ourselves as we investigated various federated search products was if the money should be spent on more content or on making currently held content more accessible? The significant increases in full text usage made us feel confident that we had made the correct decision in pursing the federated search engine.

Students Discover Resource Discovery

Anecdotally we learned that many students had used Fox Hunt for resource discovery. Many students told us that they began a search with Fox Hunt and thus determined which of the databases would be most useful for their search. They would look at the side bar with the list of database names and the number of relevant articles in each. Then they would select the database with the greatest number of hits and search more extensively with the native interface. We were surprised by this. Initially we thought that resource discovery was something only librarians would appreciate, so we didn't introduce the idea in bibliographic instruction sessions. Students, instead, discovered it on their own. In selecting a database based on the number of articles rather than its alphabetical placement on a list, students found more relevant material.

A Virtual Federated Database

We also used Fox Hunt to create a locally defined "virtual database" of public websites to meet a very specific subject need. Marist has a Fashion Design major. A single search box to the relevant text-based databases is available on the Fashion Design Subject Guide page. Our students, however, needed a resource dedicated to locating images of clothing and fashions throughout

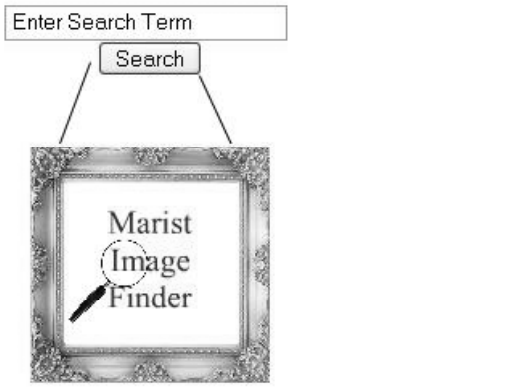


Figure 3: Marist Image Finder

history and across cultures. No commercial database focusing on images of clothing and design is even available, but a few public databases are rich in such images. These databases, however, are not picked up in a Google search because they store their images and records in a database rather than having them in html-coded pages. Google, and Google Images, only search html pages. They are not able to access the content on the web stored in a database and searched by a local search box.

To access this important content we used Fox Hunt to create a narrowly defined image search. We called it the Marist Image Finder. The New York Public Library's Picture Collection Online (http://digital.nypl.org/mmpco/index.cfm) contains thousands of fashion images spanning centuries of time, with detailed indexing. It was developed, in part, to serve the fashion industry of New York City.[3] The American Memory Images Collection (http://lcweb2.loc.gov/ammem/browse/ListSome.php?format=Photograph), a subset of the wonderful Library of Congress American Memory website, is so well indexed that it includes descriptions of the clothing worn by people in the photos. The images retrieved are included in various collections, none of which is devoted to Fashion Design nor does anything in the section name to suggest that images of clothing might be found there. Yet American Memory Images has proven to be a good source for images of different types and styles of clothing. We also included ArtCyclopedia (http://www.artcyclopedia.com/) and Google Images (http://images.google.com) in this search group. We have found that the search results are generally satisfactory and improve the service we are able to offer our Fashion Design students.

Including what Google can't index

Through our image search we have been able to take advantage of open access web resources not accessible through Google. Adding open access databases to Fox Hunt enabled us to create a virtual resource, and to incorporate high quality content into our core sources. Extending access to quality resources beyond the scope of Google, and creating virtual collections are other aspects of federated searching that offer great future potential.

Conclusion

As every librarian knows, the information marketplace is very volatile and dynamic. Our approach from the beginning has been highly experimental. This article reports a snapshot in time. Fox Hunt is being constantly improved, reviewed, and revised. There are flaws, biases, and imperfections. But when you think of what each search is doing, it is an astonishing undertaking. Fox Hunt may never be, and perhaps should never be, a finished product. We have witnessed an immediate impact in enhanced library services and increased the use of subscribed and open access resources. This certainly strengthens the case that the college should continue to expand its investment in proprietary databases.

Have we managed to out-Google Google? Well, we certainly have made our peace with them. The informal feedback we have received from students and faculty is very positive. They appreciate the simple clean look of the search box, and they like seeing the global picture of the best results. We have no way of measuring whether students are also using Google more heavily as well, but we certainly see far fewer Google screens on the library's public computers than a year ago. Clearly students have found Fox Hunt to be a resource worth utilizing.

Notes

1. Cathy DeRosa, Joanne Cantrell, Janet Hawk, and AlaneWilson. "College Students Perceptions of Libraries and Information Resources." Dublin, Ohio: OCLC. Retrieved October 25, 2006 from http://www.oclc.org/reports/pdfs/studentperceptions.pdf.

2. Jakob Nielsen. "Progressive Disclosure" *Alertbox*, December 4, 2006. Retrieved on December 28, 2006 from http://www.useit.com/alertbox/progressive-disclosure.html.

3. For a description of this remarkable collection see Anthony T. Troncal's "Worth Beyond Words: Romana Javitz and The New York Public Library's Picture Collection (http://www.nypl.org/research/chss/spe/art/photo/pchist/pchist2.html).

Bibliography

Carlson, Scott. "Has Google Won? A Librarian Says Students Have More Data Than They Know What to Do With." *The Chronicle of Higher Education*, February 7, 2003, A.28.

———. "The Net Generation Goes to College." *The Chronicle of Higher Education*, October 7, 2005, A34–A37

DeRosa, Cathy, Joanne Cantrell, Janet Hawk, and AlaneWilson. "College Students Perceptions of Libraries and Information Resources." Dublin, Ohio: OCLC. Retrieved October 25, 2006 from http://www.oclc.org/reports/pdfs/studentperceptions.pdf.

Ghapery, Jimmy. "There's an 800-pound Gorilla in Our Stacks." *College & Research Libraries News*, November 1, 2004, 582–84.

Hane, Paula J. "Web Search Trends." *Information Today*, January 1, 2003: 1, 37.

Marcum, Deanna B. and Gerald George. "Who Uses What? Report on a National Survey of Information Users in Colleges and Universities." *D-Lib Magazine*. October 2003. Retrieved on January 3, 2007 from http://www.dlib.org/dlib/october03/george/10george.html.

Nielsen, Jakob. "Progressive Disclosure." *Alertbox*, December 4, 2006. Retrieved on December 28, 2006 from http://www.useit.com/alertbox/progressive-disclosure.html.

Ojala, Marydee. "Google Views the Present, Future." *Information Today*, January 1, 2003: 1, 48+.

Troncal, Anthony T. "Worth Beyond Words: Romana Javitz and The New York Public Library's Picture Collection." Retrieved on January 3, 2007 from http://www.nypl.org/research/chss/spe/art/photo/pchist/pchist2.html.