Catalog Collectivism: XC and the Future of Library Search

Collections without services are useless, and services without collections are empty. The future of library search lies between these two statements. It is about making search smarter and putting search within the context of the user.

Collections and services

From my point of view, libraries spend most of their time around four processes: collection, organization, preservation, and dissemination. Collection managers and bibliographers identify the needs of the library's patrons and amass content to fit those needs. Catalogers use controlled vocabularies and standardized methods to describe and bring together this content to form a coherent whole. The content, in whatever form, is saved for the long term and future generations by the preservationists and conservators. Reference librarians provide access to the collection by interpreting the needs of patrons and suggesting solutions to fit their "information needs".

None of these processes are independent of the others. None is above the others. Each is required in order to fulfill the goals of a library. This is not a chicken and egg problem. A library can not have great collections, provide no services against them, and call itself a library. Such collections are literally useless. Similarly, an institution or organization can not provide information services without collections and call itself a library. Such an institution is not a library put more like an intermediary -- an index as it were. We can all name the world's largest indexer. It has no content, per say, but it provides many services. It is a library? Collections without services are useless, and services without collections are empty.

Search plays a critical role between collections and services. Right along side with browse, search facilitates the discovery of content in collections. Search and browse are probably the two most fundamental services applied against collections. Again, without access via search and/or browse, the collections are useless.

Databases and indexes

When people think of search they often, and incorrectly, think of databases. Databases, specifically relational databases, are wonderful tools for organizing and maintaining data. Through the processes of normalization, databases enable people to quickly and accurately record and update data in discrete locations avoiding the need for duplication and massive find/replace operations. Ironically, databases are notoriously difficult to search because users need to know the structure of the database in order to query it; you need to know what fields you want to search before you can do a search.

Instead, when you think of search, think of indexes. Computer generated indexes are not very much different from back-of-the-book indexes. On both cases they are lists of words or phrases associated with a pointer to where the words or phrases can be found in context. In the case of back-of-the-book indexes the pointers are page numbers. In the case of traditional library catalogs the pointers are call numbers. In the case of journal indexes the pointers are citations. In the case of Internet indexes the pointers are URLs.

Indexes make search easy. Enter a word or phrase. Get back a list of pointers. In such an environment it is not necessary to give your query very much structure. That is done for you by the software. Adding "syntactical sugar" for phrases, field searches, truncation, etc. makes search results more accurate but increasingly the underlying software does that sort of thing for you.

Moreover, through the combined use of linguistics, pattern matching, statistical analysis, and the wisdom of crowds, it is not unrealistic to not only support result set sorting by author, title, and date but also by relevance. This relevancy ranking is literally calculated based on the number of times a word or phrase is found in a particular document, the length of the document, their location in the document, and the number of times the word or phrase is found in the entire corpus of the index. Thus, the word "human" never accounts for very much in PubMed because just about every record is contains the word "human".

Future of search

The future of search lies in: 1) the enhancement of the discovery process and 2) providing services against collection beyond simple identify. Putting the user's needs and characteristics at the center of the query process will greatly enhance the discovery process. By knowing more about the searcher -- placing the query in context with the searcher -- it will be possible to improve find significantly. For example, if you know the searcher is a freshman, then it is safe to assume their experience or knowledge is less than a senior's and therefore a different set of resources may be apropos for their needs. Search can take experience into account and present results accordingly. Suppose the searcher is an expert in anthropology but are searching for information on micro-economics. Given this it is unlikely the searcher will want advanced micro-economic data, at least not right away. Present the results accordingly. Assume the searcher has a history of doing many micro-economic searches. Either they are not finding what they desire or they are looking for more specific information. Return search results accordingly. Put another way, ask yourself questions about the searcher and modify the results. Who are they? What is their level of skill or education? Are they new to the subject or an expert? Who are their peers and what are they using? Use those resources as a guide. Do they want help? To what degree to they desire privacy? By knowing the answers to these sorts of questions search results can be tailored to meet individual needs; search can be put into the user's context.

Once the discover process is improved, it will be easier to move to the next step, providing enhance services against the found items. People do not just want know a library owns an item. They want to do something with the item. Get it. Read it. Buy it. Have it delivered to them. Compare it to other items. Annotate it and take notes against it. Review it. Add it to their personal collection. Use the ideas and facts it contains to find and trace other ideas and facts. Delete it from their collection. Share it with their friends. Cite it. Summarize it. Rank it. Index it along with the other items in their collection. In an academic setting, these services can be characterized as activities supporting learning, teaching, and research. In the future, it is these services that will distinguish libraries from commercial search engines. Libraries, by definition, serve a specific use population. They never exist unto themselves. They are there to support their particular constituents. By knowing and understanding their constituents in ways commercial services can not, libraries will continue to have a role when it seems as if everybody and their brother is getting into the library act.

XC

The University Libraries is proud to be a part of the XC project. Our responsibilities are clear: 1) dump/extract our bibliographic, holdings, and authority data from the ILS, 2) make this data accessible via OAI, 3) enable a patron authentication service, and 4) enable real-time item status reports. The folks of XC will then: 1) harvest/ingest our data, 2) normalize it into a central store (a database), 2) make it searchable (an index), and 4) give Notre Dame the resulting software. It will then be Notre Dame's responsibility to implement the software in a test environment, and 2) provide XC with our feedback.

The model XC is proposing is not very much different from the model proposed by others with the exception of their process. XC's process is more open and includes a wider community than other propositions. The result should be a set of community driven "standards" for creating, maintaining, and providing access to materials in a library catalog. Moreover, since it will be open and standard's based it ought to be modular and flexible, just the sort of environment necessary when the ultimate goal is to provide sets of enhanced services against library collections such as the ones outlined above.

Collections without services are useless. Services without collections are empty. Search bridges both.

Eric Lease Morgan
University Libraries of Notre Dame

October 29, 2007