## Next-generation library catalogs, or “Are we there yet?”

Next-generation library catalogs are really indexes, not catalogs, and increasingly the popular name for such things is “discovery system”. Examples include VuFind, Primo combined with Primo Central, Blacklight, Summon, and to a lesser extent Koha, Evergreen, OLE, and XC. While this may be a well-accepted summary of the situation, I really do not think it goes far enough. Indexers address the problem of find, but in my opinion, find is not the problem to be solved. Just as much as people want to find information, they want to use it, to put it into context, and to understand it. With the advent of so much full text content, the problem of find is much easier to solve than it used to be. What is needed is a “next-generation” library catalog including tools and interfaces designed to make the use and understanding of information easier.

### Numbers of choices

There are currently a number of discovery systems from which a library can choose, and it is very important to note that they have more things in common than differences. VuFind, Primo combined with Primo Central, Summon, and Blacklight are all essentially indexer/search engine combinations. Even more, they all use same “free” and open source software -- Lucene -- at their core. All of them take some sort of bibliographic data (MARC, EAD, metadata describing journal articles, etc.), stuff it into a data structure (made up authors, titles, key words, and control numbers), index it in the way the information retrieval community has been advocating for at least the past twenty years, and finally, provide a way to query the index with either one-box-one-button or fielded interfaces. Everything else -- facets, cover art, reviews, favorites, etc. -- is window dressing.


The “Catholic Portal”

Koha, Evergreen, and OLE (Open Library Environment) are more traditional integrated library systems. They automate traditional library processes. Acquisitions. Cataloging. Serials Control. Circulation. Etc. They are database applications, not indexers, designed to manage an inventory. Search -- the “OPAC” -- is one of these processes.

### Find is not the problem

With the availability of wide-spread full text indexing, the need to organize content according to a classification system -- to catalog items -- has diminished. This need is not negated, but it is not as necessary as it used to be. In the past, without the availability of wide-spread full text indexing, classification systems provided two functions: 1) to organize the collection into a coherent whole with sub-parts, and 2) to surrogate physical items enumerated in a list.

Because wide-spread full text indexing abounds, the problem of find is not as acute as it used to be. In my opinion, it is time to move away from the problem of find and towards the problem of use. What does a person do with the information once they find and acquire it? Does it make sense? Is it valid? Does it have a relationship other things, and if so, then what is that relationship and how does it compare? If these relationships are explored, then what new knowledge might one uncover, or what existing problem might be solved? These are the questions of use. Find is a means to an end, not the end itself. Find is a library problem. Use the problem everybody else wants to solve.

### Text mining

Through the use of a process called text mining, it is possible to provide new services against individual items in a collection as well as to collections as a whole. Such services can make information more useful.

Broadly defined, text mining is an automated process for analyzing written works. Rooted in linguistics, it makes the assumption that language -- specifically written language -- adheres to sets of loosely defined norms, and these norms are manifested in combinations of words, phrases, sentences, lines of a poem, paragraphs, stanzas, chapters, works, corpora, etc. Additionally, linguistics (and therefore text mining) also assumes these manifestations embody human expressions, meanings, and truth. By systematically examining the manifestations of written language as if they were natural objects, the expressions, meanings, and truths of a work may be postulated. Such is the art and science of text mining.

The process of text mining begins with counting, specifically, counting the number of words (n) in a document. This results in a fact -- a given document is n words long. By comparing n across a given corpus of documents, new facts can be derived, such as one document is longer than another, shorter than another, or close to an average length. Once words have been counted they can be tallied. The result is a list of words and their associated frequencies. Some words occur often. Others occur infrequently. The examination of such a list tells a reader something about the given document. The comparison of frequency lists between documents tells the reader even more. By comparing the lengths of documents, the frequency of words, and their existence in an entire corpus a reader can learn of the statistical significance of

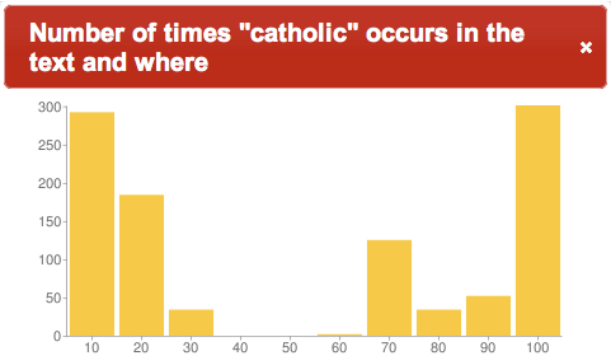
given words. Thus, the reader can begin to determine the "aboutness" of a given document. This rudimentary counting process forms the heart of most relevancy ranking algorithms of indexing applications and is called "term frequency inverse document frequency" or TFIDF.

The results of text mining processes are not to be taken as representations of truth, any more than the application of Library of Congress Subject Headings completely denote the aboutness of text. Text mining builds on the inherent patterns of language, but language is fluid and ambiguous. Therefore the results of text mining lend themselves to interpretation.

Assuming the availability of increasing numbers of full text information objects, a library's "discovery system" could easily incorporate text mining for the purposes of enhancing the traditional cataloging process as well as increasing the usefulness of found material. In my opinion, this is the essence of a true "next-generation" library catalog.

**Two examples**

An organization called the Catholic Research Resources Alliance (CRRA) brings together rare, uncommon, and infrequently held materials into a thing colloquially called the "Catholic Portal". The content for the Portal comes from a variety of metadata formats (MARC, EAD, and Dublin Core) harvested from participating member institutions. Besides supporting the Web 2.0 features we have all come to expect, it also provides item level indexing of finding aids, direct access to digitized materials, and concordancing services. The inclusion of concordance features makes the Portal more than the usual discovery system.



Where the word "catholic" is used

Through these interfaces, the reader can learn many things. For example, in a book called Letters Of An Irish Catholic Layman the word "catholic" is one of the most frequently used. Using the concordance, the reader can see that "Protestants and Roman Catholics are as wide as the poles asunder", and "good Catholics are not alarmed, as they should be, at the perverseness with which wicked men labor to inspire the minds of all, but especially of youth, with notions contrary to Catholic doctrine". This is no big surprise, but instead a confirmation. (No puns intended.) On the other hand, some of the statistically most significant two-word phrases are geographic identities ("upper canada", "new york", "lake erie", and "niagara falls") . This is interesting because such things are not denoted in the bibliographic metadata. Moreover, a histogram plotting where in the document "niagra falls" occurs can be juxtaposed with a similar histogram for the word "catholic". Why does the author talk about Catholics when they do not talk about upstate New York? Text mining makes it easier to bring these observations to light in a quick and easy-to-use manner.

Some work being done in the The Hesburgh Libraries at the University of Notre Dame is in the same vein. Specifically, the Libraries is scanning Catholic pamphlets, curating the resulting TIFF images, binding them together to make PDF documents, embedding the results of OCR (optical character recognition) into the PDFs, saving the PDFs on a Web server, linking to the PDFs from the catalog and discovery system, and finally, linking to text mining services from the catalog and discovery system. Consequently, once found, the reader will be able to download a digitized version of a pamphlet, print it, read it in the usual way, and analyze it for patterns and meanings in ways that may have been overlooked through the use of traditional analytic methods.

**Are we there yet?**

Are we there yet? Has the library profession solved the problem of "next-generation" library catalogs and discovery systems? In my opinion, the answer is, "No." To date the profession continues to automate its existing processes without truly taking advantage of computer technology.

On the other hand, our existing systems do not take advantage of the current environment. They do not exploit the wide array and inherent functionality of available full text literature. Think of the millions of books freely available from the Internet Archive, Google Books, the HathiTrust, and Project Gutenberg. Think of the thousands of open access journal titles. Think about all the government documents, technical reports, theses & dissertations, conference proceedings, blogs, wikis, mailing list archives, and even "tweets" freely available on the Web. Even without the content available through licensing, this content has the makings of a significant library of any type.

The problem of find as reached the point of diminishing returns. The problem of use is now the problem requiring a greater amount of the profession's attention.

---

Eric Lease Morgan
University of Notre Dame

May 30, 2011