Chapter 1

Preface

Abstract

The changing landscape of library collections has led to a new take on services that libraries offer. While competing with on-demand access of content found on the open web, the library needs to position itself in a new way to capture the interest of its users by providing new services that are compelling, exciting, and self-service oriented. This chapter will identify the impetus for these changes and what technologies have been made available to better support library patrons.

Ever since I first began working for a library in 2004, I have been thinking about how libraries can make their collections and resources more easily accessible in this highly accessible digital world. In 2004 and 2005, I developed an application that leveraged federated search technology to allow researchers to search across the barriers of the individual databases that our library provided access to. However, federated search quickly proved to be inadequate for the needs of a library that subscribed to around 300 databases. Trying to query hundreds of databases and process the results was like trying to squeeze a square peg into a round hole. The problem was not that one product would be better than another, but simply that certain difficulties were inherent in federated search. This technology quickly proved to be one step on the path of the library's technology-based services. This project led us to another, which I led in 2006, to provide a modern search experience in accessing the library's collections. This project, now known as VuFind, was initially conceptualized to be what I called an XML framework. Because all the library's collections, including some of the vendor content, were stored in XML, I quickly realized the possibilities. We could easily translate MARC to MARCXML; our digital library stored metadata in METS XML; our institutional bibliography could be exported into XML; data from external repositories could be accessed in the OAI-PMH XML format. The grand vision was to develop a unified database via a native XML database product. By unifying all of the collections into a single platform, the library could host a single-search-box approach to discovering the library's collections. While the idea was grand, at the time the complexities of the technology, as well as the performance and scalability, simply did not meet our expectations—it was clear that this was not going to be the killer solution.

Definitions

Federated search: a software solution designed to solve the problem of searching multiple content databases. By taking the user's search query and broadcasting it to tens or hundreds of databases at the same time, the application can compile a sampling of real-time results into a single relevancy-ranked list.

Native XML database: a platform for storing and querying XML-based files. These solutions generally support querying the content via the XQuery standard and indexing the content based on element-level rules; they can easily fetch documents based on a unique ID.

Along Came Solr

Luckily, at this time—the winter of 2006—a very talented graduate student in the computer science department, Rushikesh Katikar, was enrolled in a work-study

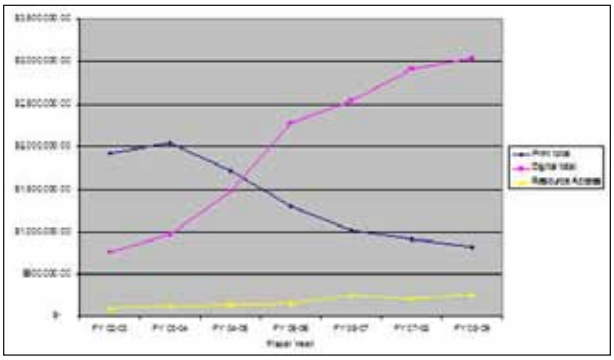

Figure 1
Print versus digital acquisitions spending at Villanova University Falvey Memorial Library, February 2003 to August 2009. Source: Joseph Lucia.

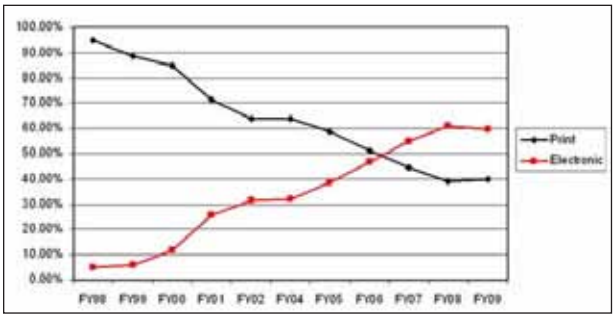

Figure 2
Wesleyan University Library acquisitions spending, 1998–2009. Source: Pat Tully, "1998–2009: How Libraries Have Changed," *From the University Librarian* (blog), Wesleyan University website, Nov. 19, 2009, http://ptully.blogs.wesleyan.edu/2009/11/13/over-the-past-10-years-how-libraries-have-changed.

program and placed in the library as a part-time software developer. Rushikesh worked side-by-side with me as we evaluated various native XML database products and analyzed their scalability and performance for querying these disparate collections. We evaluated both open source solutions and off-the-shelf commercial products. We first evaluated eXist, an open source database developed in Java. At the time, eXist was very early in its life cycle, and the current release was at a very early version number. We then evaluated a commercial solution, MarkLogic. This product is well regarded in the industry as a workhorse; however, its learning curve and the need for training forced us to push the product to the bottom of the evaluation list. The next product that was evaluated was Berkeley DB XML. Berkeley DB is also a highly regarded and heavily used database product, and its native XML counterpart felt like it would be a contender. However, after many rounds of performance testing and index tweaking, we were unsuccessful in gaining performance and scalability to meet our needs—searching a set of over one million records in under one second.

After quite a bit of frustration in tweaking configurations and the lack of scalability in the products evaluated against our collection size, we sought a different technology to support our needs. The first stop was native indexing products, whose core feature provided strengths that the native XML databases lacked. We began to play around with Apache Lucene, the industry standard for indexing. Lucence, a highly regarded and heavily used product for search, can be found in both commercial applications large and small and open source applications. When looking at how Lucene could be woven into our solution, we found Apache Solr—a Java application built on top of the Lucene product. Solr provided everything that we needed: a web-based API to develop front-end applications in any language or platform, the ability to index and store XML structured data, the ability to scale to large collection sizes, the ability to perform fast queries and facet on the search results, and a fairly low learning curve. This was the solution.

Changes in the Product Landscape

This was a very active time in the library community. Many new products were coming onto the market, and libraries were witnessing a fundamental shift from purchasing physical resources to subscribing to collections of electronic resources. We saw a growth in options around federated search products, and the next-generation-catalog product category had just been created. During my time of employment at Villanova University, the library witnessed a shift in the bulk of acquisitions spending from print materials to electronic between 2003 and 2009—a trend not uncommon at other universities in North America (see figure 1).

Wesleyan University had a similar experience between 1998 and 2009 (see figure 2).

Similar changes have happened in the publishing industry, as seen from information on the sales of O'Reilly books versus e-books (see figure 3).

Around this time, libraries began to see the value in a modern search experience that would meet user expectations, provide a highly customizable user interface, and broaden the scope of what can be discovered in the library. This is when the product line for next-generation catalogs was first established, and the race was on to implement. In a few years, thousands of libraries around the globe have adopted next-generation catalogs. This fairly rapid adoption piqued my interest, and I asked some questions:

- How will adoption of NGCs impact the services offered by libraries?
- Will adopting NGCs in an effort to provide a better solution for our patrons really make an impact in the research process and use of print materials in

Figure 3
E-book versus print orders on the O'Reilly website, January 2008 through June 2009. Source: Andrew Savikas, "Does Digital Cannibalize Print? Not Yet," *Radar* (blog), O'Reilly website, Aug. 7, 2009, http://radar.oreilly.com/2009/08/does-digital-cannibalize-print-not-likely.html.

the long run?
- Will this highly customized and modern search tool increase the library's value to its audience?

For seven years, I have worked on technology solutions to problems faced by libraries—a custom interface to a federated search product, a digital library platform, a subject guide platform, a question-and-answer system for online reference transactions, a next-generation catalog, and now a web-scale discovery solution. Throughout this period, these questions stayed with me. After seeing a bit of maturity in both age and number of installations hit the next-generation catalog product line, it was time to set out to find some answers.