Open Source, Crowd Source: harnessing the power of the people behind our libraries

Cindi Trainor, Eastern Kentucky University
“Open Source, Crowd Source: harnessing the power of the people behind our libraries”
Cindi Trainor, Eastern Kentucky University Libraries
Bridging Worlds 2008, October 17, 2008

Abstract

Libraries in large number are moving away from the traditional, vendor-sourced library catalogue software in favor of open source software that can be tailored to meet the community’s needs by the people who are most familiar with those needs: library staff. Open source products and some vendor products outside the traditional ILS market allow libraries to pool data created by users—tags, reviews, comments—thus allowing the smallest libraries to harvest richer data than that of their own communities. Come hear how libraries can combine open source software with user-generated content to create a richer discovery experience.

The state of automation

Recent years have seen dramatic consolidation of companies offering integrated library systems (ILS). New sales of traditional catalog products have waned of late due to market saturation and due to customers’ desire for modern web features such as faceted navigation and reviews, features available and in use by today’s popular websites (Breeding, 2008). Traditionally, integrated library systems have focused on acquisition, cataloging, and circulation functions; the user interface to the traditional integrated library system served as more of an inventory tool than a discovery tool, and its development was driven by the librarians that used it. Prone to arcane search structures and inflexible web layouts, the legacy OPAC still requires much library instruction and library staff/patron intervention to use effectively. Conversely, information searching by contemporary library users frequently consists of three or four word keyword searches, does not use Boolean operators, and often “is a hesitant, iterative, often random process of discovery” (Schneider, 2007b). Today’s users are not afraid of trial and error and will frequently try several searches and browse around before choosing an item. Library users also do not start at the library but may find an interesting item via Google or in Amazon, then think to check the local library’s catalog. OCLC reported in 2005 that 84% of survey respondents reported that they start an information search with an internet search engine; only 1% start at a library website (p 1-17). For this 1% of users, the traditional, inventory-based function of the library catalog is useful, but traditional catalog interfaces have been difficult to use, even when an item is known; for example, they are typically not very forgiving of typographical and other small errors, where Google has had simple spell-checking and alternate-spelling suggestions for many years (Calishain, 2000).

Integrated library systems were originally built around the MARC record, which was conceived to replace paper catalog cards (Spicher, 1996). The punctuation of a MARC record in a traditional OPAC even resembled a typed catalog card for many years (Cornell, n.d.). To create an original marc record for an item can take more than an hour, especially if all the 2000+ tags and subfields are given full consideration (Levy, 1995). Even with time saved with copy- and cooperative cataloging, is it in a library’s best interest to devote so much intellectual overhead into creating description and assigning subject headings to items that will likely be discovered by key- or title word searches conducted outside the library system? MARC still constitutes good descriptive data, but libraries are “openly questioning the cost/value of local metadata enrichment” (Schneider, 2007b).

Web 2.0, Library 2.0

The term “web 2.0” was originally coined by Tim O’Reilly to distinguish new and exciting websites from those of the “dot com bubble” era (2005, p 1). O’Reilly’s vision of the next iteration of the web saw the web as platform—sites would evolve from destinations from which users gleaned information into a computing platform, providing access to any number of applications previously existing only as software installed on local computers or networks. Examples of web applications run from online word processors and spreadsheets to photo-sharing sites to internet radio and video and everything in between. As one twitter user recently lamented upon seeing a website dedicated
to sufferers of pre-menstrual syndrome: "there is a 2.0 application for everything!" The idea of web as platform has morphed into the idea of "continuous" or "ubiquitous" computing or computing "in the cloud." The declining prices of laptops and other internet-ready devices such as the iPod Touch and the Chumby, paired with the proliferation of wireless network access points in libraries, hotels, coffee shops and other places have contributed to the popularity and utility of cloud computing. Much has been written on the evolution of the World Wide Web from small set of static pages of information to a social platform connecting people globally through their content.

O'Reilly asserted that Web 2.0 embraces "the power of the web to harness collective intelligence" (p 2). With the participatory capabilities of today's web, anyone with an internet-connected device can create and publish content or can contribute to a shared body of content. Compare Wikipedia and Encyclopaedia Britannica, for example: Britannica comprises 65,699 articles, the online equivalent of its 32 volumes, which were last published in 2007 (Britannica, n.d.). On August 11, 2008, the English version of Wikipedia reached 2.5 million articles (Wikimedia News, 2008), roughly equivalent to 1,218 printed volumes. The idea that a site like Wikipedia can harness the knowledge of literally anyone with a connection to the Internet is a powerful one, particularly considering that there are more than 250 other Wikipedia sites in as many languages (List of Wikipedias).

When the Web 2.0 concept first took off in late 2005 and early 2006, librarians were quick to envision "Library 2.0," a vision of a next generation of library services and tools that were as different from a traditional library as a "version 1.0" webpage was to its modern counterpart. The thought of espousing a new set of library ideals and tools that were social, that engaged users and that emphasized the idea that input from and conversation with users caused quite a stir. The initial debate was summarized succinctly by Walt Crawford (2006). Libraries everywhere began to experiment with new technologies that were emerging and evolving as fast as they could be pressed into service. It's important to note that the first library forays into experimenting with social software and social data entailed libraries and library staff using freely-available web 2.0 tools in their own library environment. Countless libraries created blogs at blogger.com and wordpress.com; libraries by the dozens created profiles on social networking sites MySpace and Facebook; librarians worldwide began to share images and connect with each other via flickr; and library workers worldwide created and participated in "Learning 2.0" programs (Hanly, 2007). It became expected for library users to have the opportunity to leave comments on a library's blog or flickr stream or to add their local library as a friend on MySpace or Facebook, though a large proportion of library social networking activity consisted and still consists of librarians and library staff making connections with each other on these popular sites.

Another aspect to O'Reilly's vision pertinent to the emergence of Library 2.0 is the idea that Web 2.0 meant the end of the traditional software release cycle: when software is made available on the web, boxed, numbered releases become meaningless (p 4). The idea of "permanent beta" grew up around this--features are rolled out as they are built, enabling the entire user community to test them and report bugs and desired updates at once and collectively. Sites are in constant flux, but this has meant constant growth in addition to not a little confusion and user frustration when site dysfunctions are not magicked away by programmers right away. The state of "permanent beta" has fed two key ideas in the evolution of library thinking: that it is ok to fail and ok to experiment and play (Stephens, 2006). If a new, experimental service does not work the way the library envisioned, the failure of original idea is not swept under the rug; instead, the library has the opportunity to seek user feedback and change the service accordingly. Web 2.0, and concomitantly, Library 2.0, have given library workers the opportunity to be transparent about our mistakes and to espouse an unmistakably more human and friendly environment in our libraries.

The Social Web

Today's websites allow users to interact with each other and to create and share content. Websites like Facebook connect users with each other and enable them to keep in touch and chat with each other online. Users post photos to sites like flickr, where they can add tags, comments, and notes on their own and each others' photos. Members of LibraryThing create a catalog of books that they have read or are interested in and are able to review, rate, and tag books as well as see what other LibraryThing members own the same books. Social websites bring together disparate users regardless of their geographic location, facilitating connections among people who have similar
interests (e.g., Ravelry.com), professional connections (e.g., LinkedIn.com), or who merely want to share information and have online conversations (e.g., FriendFeed.com).

The collective social information gathered by Amazon.com about its users' browsing and purchasing activities led them to develop and implement their recommendations engine (Linden, Smith, & York, 2003). Originally intended as a mechanism to put additional options in front of users' eyes—an effort to keep users at Amazon in hopes they would buy more items—the Amazon recommendations engine is an excellent illustration of the notion that more data is better data. In other words, Amazon's global user base enables it to collect information about browsing and purchasing habits of millions of users: the likelihood that a recommendation is a good one increases as the number of users who purchase the similar books increases.

**Open source: what do you get for “Free”?**

One often-misunderstood notion about open source software (OSS) is that it will save an institution the large amount of money that it would otherwise have to pay a vendor for new software. As Karen G. Schneider wrote so succinctly, open source software tools can frequently be characterized as "Free as in 'free kitten,' not free as in 'free beer,'" meaning that while there is no initial cash outlay to purchase the software, the institution is then responsible for the care and feeding of the software over its lifetime with the institution (Schneider, 2007a; Cervone, 2003). Although this is true of many programs, some open source software in use in libraries is clearly of the "free beer" variety: use of the Firefox web browser in libraries costs little over its lifetime; I would even assert that the virus, security, and spyware problems associated with Internet Explorer make it a more costly option for libraries, even though it is freely distributed (but by no means open).

Other myths persistently surround open source software. Schneider presents and writes about these myths in terms of "FUD": Fear, Uncertainty, and Doubt. Fears expressed about OSS include the assertion that such products are only good for developing countries (2008a) or that they "aren't ready for prime time" (2008b), that they are actually junky software packages put together by a hacker-wannabe in his (invariably his) garage (2008d), or that an open source product always requires a local installation, necessitating the care and feeding of a local server and its operating system (2008c).

One of the biggest advantages of the open source model is all users of the product have the potential to contribute to its development, so long as required technical expertise exists on staff. This equates to a team of developers that aren't tied to the "enhancement" model, wherein a long list is created of all the incremental changes desired by all customers, with the most popular being written and implemented each year. With an open product, the features that can be developed or tweaked are limited only be the ability of the developers using the software. An institution using open source software can set its own priorities independent of that of the rest of the user community, can contribute features and modules written in-house to the larger project, and can take advantage of enhancements written by other institutions.

The open source and web 2.0 movements in libraries have sparked sometimes not-so-quiet revolution among librarians. When I first started working in a university, the academic library was a place to which patrons came, and the librarian was the gatekeeper who unlocked users' access to information. Librarians were the wily keepers of the keys--we knew how to use the paper edition of the SSCI to hunt down how many times an article was cited; we knew how to pull the citation for that particular government document out of the Monthly Catalog; we knew which wing to send you to for PR6029.F3 A15. If a tool required us to learn how to use it before being able to pull information from it, it was our job, as librarians, to learn this new interface. It's my belief that librarians who began the profession after the advent of web searching were the first to really question this process--why is a tool so difficult to use? If a user or librarian has an idea to improve the functionality or utility of an information resource, why should users and librarians have to craft workarounds or interface tricks? Vendors should simply make their products easier to use. Our attitude has shifted in the last eight years or so from "Let me show you how" to "YES, the OPAC sucks!" to "we can make this better." Web 2.0 tools have allowed us to solicit feedback from users easily; the open source movement allows us to put those suggestions into action readily, with library-based development that is responsive to the needs of our own communities.
The first “next-generation” catalog interface was AquaBrowser, offering a single search box, faceted navigation, and a more modern search results display. AquaBrowser was available as early as September 2004, with the first U.S. installation in February 2005 at the Lexington, Kentucky, Public Library (The Library Corporation, 2004 & 2005). The next big splash in OPAC interfaces came when North Carolina State University in Raleigh, North Carolina, debuted its new catalog in January 2006 (NCSU, 2006). The new platform utilized software from Endeca to bring together the data in their Unicorn catalog with the ease-of-use of modern websites, providing an interface with a single search box and easier navigation through faceted results. The Endeca debut was the shot heard round the ILS world, sparking Innovative to announce its next-generation interface, Encore1, in May 2006, followed closely by the announcement of Ex-Libris’ Primo (Innovative, 2006 & Ex-Libris, 2006). The first vendor-offered next-generation interfaces to library catalog data had several things in common: tags or tag clouds, a single search box, faceted navigation, and a more modern presentation of search results sets. A cynic might observe that these product announcements were driven by a perceived change in market demands rather than in a change in philosophy driven by user-centered service; a hopeful cynic might assert that the market demands were driven by librarians in turn driven by a user-centered service philosophy.

The Social Web comes to library software

Library vendors have taken the first steps toward making their software social with the advent of personalization features in their products. The full-release version of Encore promises the ability for patrons to add tags and item reviews (Innovative, 2008). Primo allows users to create and browse tags and to bookmark an “e-shelf” of items (Ex-Libris, 2008). It will be interesting to see how these products evolve, but for now, the true spirit of the social web is more readily seen in products from OCLC and BiblioCommons. A good analogy of these two products’ social capabilities lies with the online photo-sharing website, flickr.com. On flickr, users are able to bookmark, comment upon, and tag individual photos and videos; in WorldCat.org, and in the catalogs of libraries using WorldCat Local and BiblioCommons, users can make lists of, review/comment or rate, and tag individual items held by the library (OCLC, n.d. & Oder, 2008). BiblioCommons takes user-generated content several steps further and provides users opportunity to create summaries, leave quotations, post notices, and give their opinions about the age suitability of any item (Oder, 2008). Both systems provide ability to see content created by other users, but only BiblioCommons lets users to mark each other as “Trusted Sources.” Marking other users as “contacts” or “friends” is a capability common many social software sites, and perhaps forms the soul of social software, if not the heart.

The major ILS vendors have largely not chosen to adopt an open philosophy when it comes to development of these new interfaces, and social features have been slow to emerge within them. A possible exception is OCLC, with their readiness to publish widgets to embed on webpages and the recent debut of their API, which is free to “most cataloging members” (Hadro, 2008b). Contrast this reluctance toward openness with the recent debut of the second-generation “Social OPAC,” the brainchild of John Blyberg, developer at Darien Library, a top-ranked U.S. public library in Darien, Connecticut. SOPAC comprises three parts: Locum Library, the SOPAC module, and Insurge. All three packages are available for installation, download, and continued development by anyone (Blyberg, 2008c). The three parts expose some of the complexity of the process of retrieving, redisplaying, and coupling library catalog data with social data.

The Locum PHP software library serves two functions, to communicate through a “connector” with a library’s proprietary ILS and to serve as the search interface to library holdings. The latter function is accomplished in conjunction with two other open pieces of software: the relational database software, MySQL, and the full-text search engine, Sphinx. Currently, a Locum/ILS connector exists only for Innovative Interfaces’ Millennium ILS, but a similar PHP connector could be written to interact with any ILS, depending on the knowledge of a library’s developers and the degree to which bibliographic, patron, and transaction information can be gleaned from the system.

SOPAC itself is a Drupal module that integrates fully into the open source content management system in use by many libraries. SOPAC provides features common to next-generation OPAC interfaces, such as faceted navigation, cover images, and saved searches, but it goes one step...
further in enabling the library to customize its catalog design: SOPAC ensures that the catalog interface looks like the library’s website design, allowing libraries to make their catalog look exactly the way that they want it (Blyberg, 2008b). The heart of this module is in its social capabilities: users can tag, rate, and review any item in the catalog. A key difference in the development of SOPAC has been Blyberg’s foresight to include the social data in the Sphinx index, alongside library holdings information; incorporating reviews and tags into the search process truly harnesses the content contributed by the library’s users (Blyberg, 2008c).

The true uniqueness of SOPAC comes with its Insurge social repository. The Insurge software library stores social data created by library users and associates it (via Locum) to appropriate bibliographic information. Libraries can choose to run a client installation of Insurge or can contribute their social data to a growing repository of data that will be created and shared by all SOPAC libraries, if they so choose. Blyberg also envisions the capability for libraries to choose which social data is brought into their local SOPACs, allowing libraries to choose similar institutions or institutions whose collections or social data might complement their own (2008b).

Because the SOPAC module and libraries are a part of Drupal, the extendability and utility for libraries is limitless. We begin to envision a truly seamless experience for library users: an open suite of Drupal modules could provide and bring together under one design umbrella the capabilities of catalog search, federated search, electronic resources access and management, OpenURL linking, user-generated content, and general website information such as hours, staff directories, events, and news. Darien Library has taken the first huge step toward this seamless experience with the unveiling of its new (Drupal-driven) website and the first implementation of SOPAC2 (Berry, 2008). The imperative for libraries to hire software developers is highlighted by the fact that the three pieces of SOPAC are built in PHP and require LAMP (Linux/Apache/MySQL/PHP) development and production environments.

Crowd Source: the dataset begins to grow

With the exception of WorldCat, all these systems suffer from a paucity of data. The likelihood that the parties interested in tagging, rating, and reviewing across the spectrum of subjects collected by a single library is extremely small. Take, for example, the first iteration SOPAC, implemented by Blyberg at Ann Arbor District Library in Michigan, January 2007 (Blyberg, 2007). Just over a year later, Blyberg declared SOPAC’s use of user-contributed tagging at AADL a failure: ‘For the past nine months, the top ten tags have included ‘fantasy,’ ‘manga,’ ‘anime,’ ‘time travel,’ ‘shonen,’ ‘shonen jump,’ and ‘shape-changing.’ As a one-time resident of Ann Arbor, I can assure you that these are not topics that dominated the collective hive mind” (Blyberg, 2008a). Systems that collect and reuse only tags created at the local level will always suffer this same fate: any given individual installation will likely never contain user tags or reviews for a wide array of subjects, but only for the subjects of interest to active taggers and reviewers. Second-generation SOPAC is constricted in this way due to its age; as libraries install and contribute to Insurge, the spectrum of reviews and tags will expand.

I assert that WorldCat does not suffer this fate because of its global user base, and because OCLC have announced that tags may be factored into search results, once a(n undefined) critical mass of tags is reached (Hadro, 2008c). A very unscientific search of WorldCat for a wide range of subjects and recently-published items reveals that many have ratings and reviews, with fewer items containing tags. This seems reasonable, as reviews have been available since September 2005 (Tennant), but tags only since August 2008 (Hadro, 2008a).

Contrast this with the vast store of data collected by LibraryThing. LibraryThing.com began in 2005 as a small website allowing users to create web-based catalogs of books that they own, have read, or are interested in. Over time, social features were added to LibraryThing, enabling the site’s members to connect with each other and see what books other members owned. LibraryThing members can add tags, reviews, and ratings and also connect with each other via groups centering around a common theme or interest. Social data is utilized at LibraryThing to make recommendations for books that members may enjoy. The social data is also used to make anti-recommendations, books that a member is least likely to enjoy, via its Unsuggester tool (Rethlefsen, 2007).
What is next?

It is my hope that the Tag Consortium alluded to by LibraryThing's Tim Spalding comes to fruition (Rethlefsen, 2007). A partnership between SOPAC and LibraryThing or OCLC and LibraryThing would be a powerful step toward a true rival to Amazon.com's website built and living in the library world. This statement, of course, is made interesting by Amazon's recent announcement to acquire Abe Books, which holds a 40% stake in LibraryThing (Hadro, 2008a). A product combining the open development potential of SOPAC, the 10 million tags of LibraryThing, the worldwide union holdings listing of WorldCat and the popularity of Google would secure libraries' future as a vital part of today's web community by linking local library holdings and user-generated data with that of library users and book readers worldwide.
References


---------------------

Footnotes

1 In the interest of full disclosure, I must tell you that my library is a development partner with Innovative Interfaces in their Encore product, a third-party catalog overlay that brings a prettier, more amazon-like interface to our Voyager catalog. You can search Encore from my library's website, library.eku.edu by clicking "encore beta search" on the Books tab.