

Discovery Platforms

By Richard W. Boss

Libraries have always given high priority to facilitating access to information by their patrons. The card catalog, with its multiple access points, provided access by author, title, series title, and subject headings. The online patron access catalog improved access by adding keyword and Boolean searching. The next step, less than a decade ago, was the introduction of the library portal, a single user interface for access to a wide variety of electronic resources both within and outside of the library.

A portal is more than just a gateway or a way out to resources. What distinguishes a library portal from a gateway is that it augments the user interface with federated searching, patron authentication, and link resolution—the last of which gets beyond the sources of the content to the content itself. [See the TechNote entitled “Library Portals” for a detailed description.]

Libraries are now moving beyond the library portal to a discovery platform, a bundle of technologies designed to give the user a searching platform that goes beyond portal technology by augmenting federated searching, patron authentication, and link resolution with fuzzy logic, faceted classification, data harvesting, the grouping of search results into “clouds” of related information, tagging by users, and the ability of patrons to add contents such as reviews and ratings.

The use of the word “platform,” instead of “application” is critical in distinguishing a discovery tool from a patron access catalog or library portal. A “platform” is a system that can be reprogrammed or customized by the purchaser and, in that way, adapted to needs that the original developers may not have anticipated. In contrast, an “application” is a system that cannot be reprogrammed by the purchaser. It does whatever the original developers intended it to do.

Library patrons who have used Google, Amazon, and NetFlix had experience with discovery platforms before the first library implemented a discovery platform. The first discovery platform developed for libraries and widely used by them was AquaBrowser, a product of a Dutch software company called Medialab (www.aquabrowser.com). When AquaBrowser was introduced in Europe in 2000 and in North America in 2002, it was used solely as a better way to access a library's patron access catalog. Over the next six years it evolved into a comprehensive discovery platform. By that time, several other discovery platforms for libraries were in various stages of development.

A discover platform is not a single technology. Not all libraries purchasing a discovery platform will utilize all of its technologies. For example, libraries that have already implemented federated searching and remote patron authentication as part of a library portal will not need to purchase these again.

Features of discovery platforms

Librarians who plan to investigate available discovery platforms should first become familiar with all of the potential features, and should determine which are essential for a particular library as not all features are available in each product.

Single point of discovery

A mature discovery platform offers a single point of discovery for all in-house and remote electronic resources in one place. As of the second quarter of 2009, some of the available library discovery platforms had not yet achieved that.

Fuzzy search

A fuzzy search is a process that locates bibliographic records or web pages that are likely to be relevant even when the search argument does not exactly match the desired information. It is done by means of a fuzzy matching program. Fuzzy matching includes a spell checker and spelling error corrector. The program can return hits with content that contains a specified base word along with prefixes and suffixes. That is based on the assumption that related words are likely to have the same core and differ at the beginning and/or end. Working like an online thesaurus, the program can also find synonyms and related terms. Fuzzy matching will usually return irrelevant hits as well as relevant ones. For that reason, relevancy ranking is an important related feature. A common component of this feature is “did you mean?” That allows the searcher to decide whether a match is potentially useful.

Harvesting

Harvesting is the pre-indexing of content that is pertinent to the users for whom the harvest is intended. The applicable standard is OAI/PMH, the Open Archives Initiative Protocol for Metadata Harvesting. The harvester is a client application that issues OAI-PMH requests to a repository that is managed by a data provider that exposes metadata to harvesters. Most data providers are not willing to allow harvesting outside their own organizations. For that reason, a library is most likely going to undertake harvesting only of its own repositories. As of the second quarter of 2009, not all library discovery platform developers had completed their work on harvesting.

Federated searching

Federated searching, an essential component of a library portal, is also essential to a discovery platform. It is the ability to simultaneously search multiple sources of information within and outside of a library using a single search statement. As database

aggregators do not allow harvesting, federated searching is the key to accessing their data.

Remote patron authentication

Remote patron authentication, an essential component of a library portal, is also essential to a discovery platform. It qualifies a searcher only once even though s/he is accessing multiple sources that require authentication.

Standards support

A discover platform developed for libraries will support the standards most important to libraries, among them MARC, Dublin Core, FRBR, Z39.50, and the aforementioned OAI-PMH.

As of the second quarter of 2009, the status of the RDA (Resource Description and Access) standard, was uncertain. Only AquaBrowser was supporting it.

Duplicate detection

One of the major issues in information retrieval is the sheer amount of information. It is, therefore, desirable to detect and remove duplicate data. That is not difficult when there is no difference between two results. It becomes more difficult when the only difference is in formatting or word order. It becomes very difficult when there is a high level of similarity. A number of metrics have been developed to automatically evaluate potential duplicates, but none of them are perfect.

Relevancy ranking

Relevancy ranking is the display of search results based on the level of confidence that the information is what is wanted. Term frequency is the primary way of determining whether the information is relevant. If one is searching for diabetes and the word "diabetes" appears multiple times in a document, it's reasonable to assume that the document will contain useful information. However, if the search term is a common one, or if it has multiple meanings, one could wind up with many irrelevant hits. A refinement used in relevancy ranking is the positioning of the search terms, reasoning that if the search term appears in the header as well as the text, it is more likely to be relevant. Another method is to determine which documents are most frequently linked to other documents. Relevancy is often expressed as a percentage. For example, an exact match in a subject field may be labeled as 100 percent while the presence of the search term in the full-text may be labeled 60 percent.

Normalization

Normalization is any process that makes search results conform to some regularity or rule. Examples are Unicode normalization, removing punctuation or accent marks from letters, expanding abbreviations, removing stopwords, or converting all letters to upper or lower case.

Faceted classification

Faceted classification allows the assignment of multiple classifications to an object, enabling the classification to be ordered in several ways, rather than in a single pre-determined order. The idea is that one can decide to divide objects by name, subject, language, format, location, availability or other facets rather than browsing through a pre-determined hierarchy that may not precisely suit the searcher's way of thinking.

Tagging

With the advent of Web 2.0, social bookmarking tools became available to searchers. When information is found, tags (keywords) can be added to provide additional access points. That facilitates refinding of the information not only for the taggers, but also other searchers. The use of tagging does not alter MARC or other records, but provides access by emerging, colloquial, and specialized terminology not found in formal subject headings.

Tag cloud

A tag cloud is a visualization of word frequency. It provides a compact overview of search results in the form of clusters, maps, or graphs. The popularity of tag clouds is attributable to the fact that they are easier to read and understand than lists. Tags may be displayed in alphabetical order and visually weighted by font size. The most common words in a language are usually ignored. They may also be grouped into topical clusters based on relationships or their use in various disciplines. To highlight a tag on a cloud, one clicks on it with a mouse.

Personalization

Personalization can consist of alerts of materials due, the availability of requested materials, or other patron-specific messages.

Collaboration

An important component of a discovery platform is the ability of searchers to add not only tags, but also ratings and reviews.

Internationalization

Accommodation of multiple languages and differences in the approach to searching is important not only for vendors that seek to market globally, but also to libraries that increasingly serve very diverse communities.

Consortium support

Libraries in a consortium may choose to share a system, yet want to establish and maintain their own policies and procedures.

Major vendors

Not all of the major vendors target public libraries, but even those that focus on other markets are identified in the following paragraphs.

AquaBrowser has been implemented by academic, public, and special libraries since 2000, but the vast majority of the more than 410 libraries that have purchased it are public libraries. By the time R. R. Bowker—a division of Cambridge Information Group, a company that also owns ProQuest and Serials Solutions—acquired Medialab in mid-2007, a number of libraries were using it not only to improve access to their own patron access catalogs, but to a wide variety of other resources within and outside of the library, including a library's own Web site, digital repository, ERM system or image database(s). It could also be used to search other libraries' patron access catalogs, other institutional Web sites, database aggregators' services, individual subscription databases, and library selected URLs. All of the features discussed in this TechNote were available except for collaborative tools. A year later, AquaBrowser was augmented with collaboration tools that were named "My Discoveries." It allowed users to create reviews, ratings, and view personal tags. An underlying patron access catalog is not required unless a library wants to offer patron services. The size of a library's collection is the basis for pricing.

Endeca (www.endeca.com), a software company that has developed discovery platform software for a wide range of industries, had its product adopted by the North Carolina State University Library in 2007. As of the end of 2008, eight other university libraries had installed the product, one now known as the Information Access Platform. The company does not appear to target the public library market. That may explain why it failed to respond to repeated requests for information for a publication of the Public Library Association. However, the Phoenix Public Library has been identified by the company as a customer.

ExLibris (www.exlibrisgroup.com) its Primo discovery platforms in 2006 and introduced it in general release in late 2007. From the beginning, the goal was to extend the scope of searching to a wide range of electronic resources. ExLibris targets the academic and special library markets and did not choose to respond to repeated inquiries. There were approximately 130 Primo installations as of the end of 2008.

Innovative Interfaces' (www.iii.com) Encore was first released in November of 2007 and had more than 150 installations by the time it introduced Release 3.0 in early 2009. It can be used to access all electronic resources within and outside of a library. All of the features described in this TechNote were included. All of the relevant standards except RDA were supported. It does not require an underlying patron access catalog, but can work with either the vendor's own patron access catalog or that of any other vendor. The size of a library's collection or circulation is used as the basis for pricing.

OCLC (www.oclc.com) announced its TouchPoint in June of 2008. As of the first quarter of 2009 a pilot of the product was underway in the Netherlands.

Polaris (www.polarislibrary.com) does not offer a discovery platform, but it did introduce some elements of a discovery platform in June of 2008 partially as an enhancement to its PowerPAC patron access catalog and partly as a separately priced

add-on product called Fusion. As of the second quarter of 2009, it was possible for users to search a library's own patron access catalog, digital repository, or image database(s), but not its own Web site or ERM system. It could also be used to search the patron access catalogs of other libraries, individual subscription databases, and library-selected URLs. MARC, Dublin Core (Fusion only), OAI-PMH (Fusion only), and Z39.50 standards were supported. The available features as of the second quarter of 2009 were federated searching, patron authentication, personalization, fuzzy matching, faceted search results, normalization, relevancy ranking, patron activity alerts, and consortium support. Harvesting was available only to purchasers of Fusion. The basis for pricing is the size of a library's collection.

SirsiDynix (www.sirsidynix.com) launched its Enterprise in the third quarter of 2008. The initial version supported only the searching of its own patron access catalog, but access to a wide range of other electronic resources was planned. By the second quarter of 2009, access to a library's own Web site was included, but access to all other electronic resources was still planned for future release. The features that were available at that time were a single point of discover for all source types, fuzzy logic, faceted searching (based on the GlobalBrain data retrieval technology from BrainWare), relevancy ranking, patron activity alerts, internationalization, and consortium support. All of the other features were still in development. The standards supported were MARC and Dublin Core. The OAI-PMH standard for harvesting and Z39.50 support were planned for a future release. An underlying patron access is required, but it must be the vendor's own. Pricing is based on the size of a library's collection and the size of its patron base. It had signed 18 contracts as of the end of 2008, all for a hosted service or SAAS (software as a service).

TLC (www.tlcdelivers.com) introduced the first version of its LS2 in October of 2008. Searching of a library's own patron access catalog or Web site was supported, as was searching of individual subscription databases and library selected URLs. Access to a library's own digital repository and own image database(s) was planned. There were no plans to provide access to the patron access catalogs of other libraries, other

institutional Web sites, or database aggregators' services. Only MARC was supported, but support of Dublin Core and OAI-PMH for harvesting was planned. Features still in development were fuzzy matching, harvesting, guidance and recommendations, collaborative tools, and internationalization. No underlying patron access catalog is required. Pricing is based on the size of a library's collection and the number of patrons. There were 12 installations as of the end of 2008.

VTLS (www.vtls.com) introduced its Visualizer in 2008. It is designed to work with all of the company's products. The discovery platform has not only a facet-based search engine, but a knowledge base that adds information facets to the data. This allows searchers looking for information about Africa to also discover information about Kenya or those looking for information about jazz to discover information about Thelonious Monk. Visualizer can be used to search a library's own catalog, Web site, image database(s), and digital repository and external targets systems. As the name "Visualizer" suggests, tagging and tag clouds are an essential component of the discovery platform. Fuzzy matching, personalization, patron activity alerts, and collaboration are supported. Pricing is based on the size of a library's collection. There were three installations as of the end of 2008.

Completed April 28, 2009