Next Generation Catalogue: A User’s expectation

Vinit Kumar

Abstract

Since the days of Cutter, tools to access the resources of libraries are changing their structure and interface rapidly and dramatically to fulfill the dynamic user needs. Today almost every library user comes with expectations set and defined by their experience of using the Web. So the catalogues, which are offered by the libraries, need to operate at the same level of sophistication as other popular Web destinations. The "next generation” library catalog is a tool designed to fit into this shifting environment and move librarianship into a more active role when it comes to increasing the sphere of knowledge. The purpose of the paper to examine the present developments and explores the likely future developments in re-designing the OPAC to support resource discovery. The different ongoing developments follow a unique approach, but one thread that is common in all of them is that they involve a desire to go far beyond the capabilities of legacy catalogues and give library users more powerful and appealing tools.

Keywords: Next Generation Catalogues, Online Catalogues, Future Catalogues, OPAC, Web OPAC

1. Introduction

In the past decade people’s expectations regarding search and the access to information have dramatically changed. The proliferation of desktop computers, the increasing amount digital content, and the advent of the Internet have all contributed to this phenomenon. People expect to enter one or two words (or maybe a phrase) into a search box, click a button, get a list of relevancy ranks results returned, select an item from the results, and view/download the information. The process is quick, easy, and immediate. Even though people know the information available from libraries is free and authoritative they often use libraries as a last resort because they see library systems are overwhelmingly associated with books, confusing, difficult to use, time consuming, and inconvenient.

In library parlance, an OPAC is an "online public access catalogue." The operative word in this phrase is “catalogue”. Traditionally speaking, the OPAC is thought of as an index to the things owned or licensed by a library. It is an electronic version of the card catalogue. As such it contains “cards” pointing to books, not the books themselves. It contains “cards” pointing to article indexes but not the articles themselves. Because libraries do not own them, library OPAC’s, for the most part, do not contain pointers to Internet resources because they are too dynamic in nature to maintain. These things, along with the bibliographic indexes, are relegated to a library's website. Metasearch technology has tried to bring these things together under one search interface but with little success. As a catalogue, OPAC is a tool allowing people to search for and identify things in or available from libraries. Beyond identification, OPAC do not provide very many services against holdings besides the ability to borrow.
The proposed “next generation” library catalogue is intended to address the issues outlined above. It builds on the good work previously done by the library profession through collection, preservation, organization, and dissemination. It takes into account the changing nature of user expectations and tries to meet them. It exploits computer technology and harnesses the knowledge built up by the information retrieval and digital library communities. The goal of the “next generation” library catalogue is to create a transparent system enabling library users to get their work done more quickly and efficiently. It will not be seen as an impediment to learning, teaching, and scholarship, but rather as useful tool for getting an education and increasing the sphere of knowledge.

2. Problems associated with the present catalogue

Today almost every library user comes with existing expectations set by their experience of the Web. So the catalogues that are offered by the libraries need to be operated with the same level of style and sophistication as other popular web destination.

Some of the shortcomings of the legacy catalogues (OPAC) might be enumerated as:

♦ complex search interface that might not be sufficiently intuitive
♦ not consistent with the well established user interface conventions
♦ unable to rank the results according to relevancy or interest
♦ limited in scope: does not include
  i) article level searching
  ii) online display of article content
  iii) search and display of content from local digital library collections: photographs, manuscripts, local newspapers, genealogical materials, and the like
  iv) Contents of an institutional repository.
♦ limited to print materials and are less able to address electronic content
♦ unable to deliver online to the users
♦ lacks social networking features in engaging library users

3. What actually are they using?

In 2005, according to Perceptions of Libraries and Information Resources, a major study done by OCLC revealed some threatening statistics

Question: Please indicate if you have used the following electronic information sources, even if you have used them only once.

♦ 72% used Web search engine. 30% used library catalogue
♦ Where do you typically begin your search for information on a particular topic (all participants)?
84% Web search engine. 1% library

<table>
<thead>
<tr>
<th>Electronic Resource</th>
<th>Total Respondents</th>
<th>College Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-mail</td>
<td>74%</td>
<td>83%</td>
</tr>
<tr>
<td>Search engine</td>
<td>72%</td>
<td>82%</td>
</tr>
<tr>
<td>Instant messaging/online chat</td>
<td>51%</td>
<td>69%</td>
</tr>
<tr>
<td>Online news</td>
<td>50%</td>
<td>64%</td>
</tr>
<tr>
<td>Online bookstore</td>
<td>47%</td>
<td>62%</td>
</tr>
<tr>
<td>E-mail information subscriptions</td>
<td>46%</td>
<td>51%</td>
</tr>
<tr>
<td>Topic-specific Web sites</td>
<td>41%</td>
<td>50%</td>
</tr>
<tr>
<td>Library Web site</td>
<td><strong>30%</strong></td>
<td><strong>61%</strong></td>
</tr>
<tr>
<td>Electronic magazines/journals</td>
<td>28%</td>
<td>58%</td>
</tr>
<tr>
<td>Blogs</td>
<td>16%</td>
<td>29%</td>
</tr>
<tr>
<td>Online database</td>
<td>16%</td>
<td>34%</td>
</tr>
<tr>
<td>Electronic books (digital)</td>
<td>15%</td>
<td>31%</td>
</tr>
<tr>
<td>Ask-an-expert</td>
<td>14%</td>
<td>21%</td>
</tr>
<tr>
<td>Audio books (downloadable/digital)</td>
<td>10%</td>
<td>16%</td>
</tr>
<tr>
<td>Online librarian question service</td>
<td>6%</td>
<td>8%</td>
</tr>
<tr>
<td>RSS feeds</td>
<td>5%</td>
<td>7%</td>
</tr>
</tbody>
</table>

Table 1: Usage of Electronic Resources—By College Students and Total Respondents

4. How do we move forward?

"If libraries are to avoid further marginalization, they need to make a fundamental change in their approach to user services." (Peterson, 2007)

So, considering the inadequacies of the legacy catalogue, a number of approaches are going to bring out the solutions as “next generation catalogue” which is expected to gather a broader set of information, resources and services into a single interface that is more comprehensive in scope and more in modern presentation.

5. Terminology

5.1 It is not a catalogue

According to the Eric Lease Morgan in his “A next generation library catalog”, June 2006, “In two sentences, this catalog is not really a catalog at all but more like a tool designed to make it easier for students to learn, teachers to instruct, and scholars to do research. It provides its intended audience with a more effective means for finding and using data and information.” The “next generation” library catalogue is not a catalogue, per se.
The “next generation” library catalogue is not intended to be searchable list of the things a library owns or licenses. It is more than that. It is a list of the things deemed useful for achieving the goals of the library’s parent organization. The world of information has moved beyond books, magazines, and videos. The computers and the Internet has spawned mailing lists (full of names, addresses, and telephone numbers), data sets, images, full-text books and journal articles, archival finding aids, pre-prints & other gray literature, etc. To various degrees all of these things are important to meeting the needs of library users. To the best of its ability, the “next generation” library catalogue includes the full-text of these items or at least metadata describing these items in its collection.

5.2 It avoids multiple databases

The “next generation” library catalogue avoids multiple databases and indexes, thus increasing simplicity and reducing the problem of de-duplication. Multiple databases and multiple indexes increase complexity and necessitate the requirement of computer infrastructure. Even creates information silos and makes it challenging to create holistic systems. By avoiding the creation of multiple databases and indexes, it will be easier to do global searching and refinement. Reducing information overload from the too many search hits through the use of faceted browsing, an “intelligent” user interface, and the ability to create tightly focused queries against the indexes can be avoided.

5.3 It is bent on providing services against search results

The “next generation” library catalogue user interface is bent on doing things with found items beyond listing and providing access to them. Again, the “next generation” library catalogue is not a catalogue but a tool.

One could do all the things that he can do with a physical book. One can place a hold on it. One can request it, and borrow it even remotely through document delivery. One can renew it. With a tiny bit of work, we could turn the book’s metadata into a various citation styles for inclusion in a paper. We could allow people to review the book. Many of the same things can be applied to printed journal articles, but electronic journal article has a different story. It facilitates greater number of services such as download, print, save, annotate, and index along with a set of other downloaded things. With adequate metadata, it would be relatively easy to feed a book in a system or journal article and request a list of possible email addresses for the author.

In other words, the “next generation” library catalogue provides services against items in its collection. These are value-added services provided by the library, and in turn, saving the time of the reader and making them more productive.

5.4 It is built using things open

This “next generation” library catalogue is built using open standards, open source software, and open content in an effort to increase interoperability, modularity, and advocate the free sharing of ideas.
Librarianship is a highly collaborative profession. We share cataloguing records. We jointly purchase/license materials. We share in the creation of collections. Our professional meetings measure in the 10’s of thousands of people. Much of this is facilitated through the use of open standards like Z39.50, MARC, and AACR2. Hardware and software are crucial to the provision of the “next generation” library catalogue, and it is a well-known fact that “given enough eyeballs, all bugs are shallow”. Therefore open source software will be used in order to keep things transparent. Finally, open content such as articles from open access journals, full-text books from Project Gutenberg, theses & dissertations available from the Networked Digital Library of Theses and Dissertations, and preprints from open archives data repositories will be given priority in an effort to highlight advocate the use of information unencumbered by licensing agreements. This does not mean commercially available content will not be included, but access to the content or its metadata must be provided through some sort of open standard such as OAI-PMH or XML files.

6. **What are they going to serve?**

Features expected that next generation catalogue:

- will be a single entry point into all of the resources (print and electronic) that a library has to offer (catalogue, digital repository, and subscription databases);
- will be easy to learn and powerful for experienced users;
- would support the full range of metadata standards;
- more powerful search interface;
- guide users to precise, comprehensive results;
- check spelling;
- decrypt journal title abbreviations;
- connect expressions and manifestations with their respective works; and
- will offer faceted browsing of result sets interactively to users with appropriate results.

7. **Faceted Navigation**

A user-interactive technique that has proven it to be extraordinarily useful in the search process involves the use of facets that can be selected to narrow the results.

The idea is to navigate by ‘facets’. So in the first step users can select a facet out of the set of all facets. In the next step, the users can select another facet out of the remaining facets, which users will see as available navigation choices in dependent on the path they take.
For example, it will theoretically work well for ‘collections’ of things like recipes, flights, cars, hotels, CD’s, books and so on. However, if the collection has a traditional categorization that users will expect, Faceted Navigation may not be the best choice. If your site has a lot of ‘double-bound’ items (items that appear in more than one category), you may have to consider Faceted Navigation.

Faceted Navigation can also be seen as an alternative for Advanced Search.

In terms of navigation, there are two options for realizing Faceted Navigation:

1. Using a variation of Directory Navigation. In each step the number of categories is reduced by one.

2. A variation of Advanced Search. Instead of entering a keyword, the users set the characteristics and get the results matching the characteristics.

For both variations, it is important that the users should see how a selection leads to fewer results. Therefore, the minimum feedback is the number of results that match the current selections. Even better is to show the results directly so that users can select them if they see what they like.

Thus, instead of forcing one way to view the items, faceted navigation allow users to view the items in any way they want. At the same time, user learns how the items are structured and can decide on reconsidering the search strategies.

8. Relevance

Another key expectation well established by other web interfaces is that almost all web search engines return results according to some kind of relevancy. In legacy catalogue, search results are simply ranked according to last in/first out similar to the databases search and displays it accordingly entered into database. Somewhere the search result is sorted by “date”. Relevance ranking is just one of many basic search-engine functionalities missing from online catalogs.
It appears almost magical to the searcher, when relevancy ranking works well. But implementing a system that performs good relevancy ranking is very difficult. The problem becomes tenser for broad queries as the number of results can be very large.

Most of the search engines are using two techniques for relevancy ranking

♦ **TF**, the term frequency, measures the importance of the term in the item one retrieves. The more the term appear in the document the higher the document will be in the ranking. As the more number of terms shows that the term is more important for the document.

♦ **IDF**, the inverse document frequency, measures the importance of the word in the database one searches. The fewer times the term shows up in the entire database, the more important, or unique, it is.

Putting TF and IDF together—the importance of a term in a document, and the uniqueness of the same term in an entire database—basic relevance ranking can be achieved.

In library context, a number of other factors might be considered. For example, when ranking for books, the number of times that an item has been checked out could be considered as an indicator of popularity. In the same way, number of copies owned by the library, number of other libraries that own the same book according to OCLC could be used.

9. **Spell Checker (“Did You Mean...?”)**

Another feature, which has grown to be universally accepted, involves the ability for the search engines to detect common spelling errors in a query. For example, in Google search, misspelled words are suggested as “did you mean...?” with a term that will work. To implement this feature a number of technical algorithms are available. One can create an index of phonetically similar words that can be used to test the plausibility of a query. If a phonetically similar terms returns a much larger results set than the one provided by the user, then the system might present that term as a suggestion. The index can be updated by using the statistics available at Google trends website, by taking the most popular search strings and match them with the user query. In the library context, a spellchecker (did you mean...?) feature can help prevent a very large number of failed searches, which is much better than “No results found”.

10. **Recommendation Feature**

This feature is borrowed from e-commerce arena, involves proactively providing information about related materials. Microsoft in its download website, for example, provides a prominent recommendation feature: “Those who downloaded X also downloaded Y”. While this feature is highly used for merchandising motivation of online e-commerce websites, likewise libraries can use this feature for promoting other materials in the collection. But for implementing a lot of user behavior data is required on which to base these associations for recommendations.
11. **Web 2.0: Enabling User Contributions**

The tools that are available in recent years with the incoming of Web 2.0 can be used to enable user contribution. There are several ways in which this approach can be incorporated into next generation catalogues. In addition, a catalogue can rely on its users to contribute supplemental content. Users can be invited to rate the items or to write reviews that express their opinions regarding works represented in the catalogue. Other user might comment on these reviews or write their own.

(i) **Tagging**

Tagging is another practice associated with the Web 2.0 movement, provides a convenient way for users to assign their own informal terms (tags) to items of their interest. Completely unconstrained by formal rules, users chose their own tags, which they can use to find items later on. While some tags might relate to some communities, some may be related only to personal interests. A “folksonomy” is a set of tags developed among a community to help classify a body of information.

In library context, libraries can use this method of organizing their collection as some large collection websites like Flickr rely solely on user assigned tags. Likewise user assigned tags can serve as an interesting supplement to the subject vocabularies (subject headings) traditionally found in library catalogues.

(ii) **RSS**

RSS (really simple syndication or rich site summary) is another tool given by Web 2.0, highly used for distributing content. It provides the user with opportunities to use the content in more convenient way as was possible in conventional Web pages. RSS delivers a set of related items through a simple XML protocol and today finds widespread use.

In library context it can be used in following ways:

1. List of news items in the collection.
2. New library rules
3. List of relevant items in other environments.
4. Received items per day.

12. **Some developments in the Library World**

If we talk about first generation catalogues (1960s and 1970s) they provided the same access points as the card catalogue by dropping the user into a pre-coordinated index. Based on the assumption that most users were interested in known–item searching. Then came second generation of online catalogues based on keyword and Boolean searching. While system based
on Boolean algebra represented an advance over those that preceded them, Boolean is still a retrieval technique designed for trained and experienced searchers. Boolean systems were, however, simple to implement and economical in their storage and processing requirements, important at that time.

Soon after the fashion of combining free text terms across records wore off, the library community recognized that major problem with first and second –generation catalogues was the difficulty of searching by subject.

By the early 1980s libraries turned to next generation catalogue features. Out of the surge of interest in improving online catalogues, emerged a number of experimental catalogues that incorporated advanced search and matching techniques developed by researchers in Information retrieval. They typically did not rely on exact match (Boolean) but used partial-match techniques (probabilistic and vector-based). Since probabilistic and vector-based models were first worked out on document collections, not collections of MARC records some adaptations were made. These prototype systems included Okapi, which implemented search trees, and Cheshire II, which refined probabilistic retrieval algorithms for online catalogues.

In recent years the OPAC emerged as a module of an integrated library system (ILS) that allows patrons to search the collections, which were focused on the physical items. The concept of an OPAC that is tied solely to the physical inventory of the library and that doesn’t incorporate at least the basics of how people use the Web today.

Today, we see a variety of projects that are all intent on creating a new application that better suits the current expectations of library users. Some creative efforts are coming from the:

i) Roaster of traditional library automation companies,
ii) Some from other commercial sectors, and
iii) Some from librarians themselves.

This all became possible because of the almost limitless computational power, network bandwidth, and data storage at incredibly low prices. While in previous times, basic software infrastructure came at a dear price, provided by computer software giants like IBM and Unisys etc. In today’s environment, open source components abound for all of the layers of technology infrastructure. The Linux operating system; the apache Web server; relational database management systems such as MySQL, PostgreSQL, the Lucene and Solr search technologies; and many other open source projects provide essential technological infrastructure components with very low cost.

13. From traditional Library automation companies

Polaris Library System offers an ILS called Polaris. Polaris chose the path of enhancing the online catalogue it delivered as part of its integrated system with the state of features expected
in next generation library interfaces. The latest version (version 3.3) includes features such as relevancy ranking, faceted navigation, “did you mean?”, enriched content and book jacket images. This version also uses AJAX to improve the function of its user interface.

13.1 Katipo Communications Ltd — Koha (marketed by LibLime)

Koha is open source software first developed and deployed by Katipo Communications Ltd, but now, community develops it.

After the user enters a query, Koha returns a results list sorted in relevancy order. A drop down control offers the ability to resort the results in several different ways, including “field-weighted”, “relevance ranked,” “popularity,” “author,” “call number,” “dates,” and “title”.

13.2 Some from the library sector

Georgia Public Library System—Evergreen
Villanova University—VuFind

13.3 Some from other commercial sector

The most popular approach towards creating a next generation library interface involves making use of technology from other commercial sectors. These technologies are not especially meant for libraries but libraries can use them as a tool to achieve their goals. There are some product from the commercial market such as Talis—Prism, Medialab Solutions—AquaBrowser, Innovative interfaces—Encore, Ex Libris—Primo, LibraryThing—LibraryThing for Libraries. Apart from those the most important and pioneer improvement is brought by “Endeca – Guided Navigation” which was first implemented by NCSU (North Carolina State University) Library. Endeca was founded in 1999 and is based in Cambridge, Massachusetts. Endeca provides state-of-the-art search and navigation technology. In May 2006, Endeca was awarded a patent covering its “hierarchical data driven navigation system and method for information retrieval,” or “Guided Navigation.”

The Endeca search engine includes a wide variety of sophisticated features for determining matches to a query and for calculating relevancy ranking, determining appropriate facets, and giving spelling suggestions. Following are some features of the Endeca:

13.3.1 Functionality

Functionality of Endeca IAP falls in three main areas: relevance-ranked results, new browse capabilities, and improved subject access.

13.3.2 Searching and relevance ranking of results

Endeca’s MDEX search engine is capable of both Boolean and limited partial-match retrieval. Queries submitted to Endeca can use one of several matching techniques (e.g., matchall, matchany, matchboolean, matchallpartial). The NCSU implementation uses “matchall” technique for keyword
searching, an implied AND technique that requires that all search terms (or their spell-corrected, truncated form) entered by the user occur in the result.

Although classic information retrieval research tends to associate relevance ranking with probabilistic or vector-based retrieval techniques, Endeca includes a suite of relevance ranking options that can be applied to Boolean type searches (i.e., implied AND/OR). Each search index created in the Endeca software can be assigned a different relevance ranking strategy. This capability becomes significant when considering the difference in the data being index for ISBN/ISSN as compared to general keywords search. Since keyword anywhere index contains the majority of the fields in a MARC record and is the default search operator, its relevance ranking strategy received the most attention. This strategy currently consists of seven modules. The first five modules rank results in a dynamic fashion, while the final two modules provide static ordering based on publication date and total circulation.

The NCSU libraries, algorithm prioritizes results with the query terms exactly as entered (no spell correction, truncation, or thesaurus matching) as most relevant. For multiterm searches, results contacting the exact phrase are considered more relevant than those that do not. In addition, NCSU has created a field priority ranking, which provides the ability to define matches that occur in the title as more relevant than matches that occur in the notes fields. The relevance algorithm also considers factors such as the number of times the query appears in each result and the term frequency/ inverse document frequency (tf/idf) of query terms.

13.3.3 Spell correction, “Did you mean . . . ?” and sort

Several other features are included in the basic Endeca IAP application. These include auto-correction of misspelled words, which uses an index-based approach based on frequency of terms in the local database rather than a dictionary. Due to the presence of unique terminology in the database (particularly author names), the relevance ranking has been configured to display any matches on the user’s original term before spell-corrected matches. A “Did you mean...?” feature also checks queries against terms indexed within the local database to determine if another possible term has more hits than the original term in order to provide the user the option to resubmit the search with a different spelling. Various sort options are supported, including date, title, author, and “most popular.”

13.3.4 Browse

Many users site the serendipity of browsing the stacks and “recognizing” relevant resources as a key part of their discovery process. Key browsing features provided by the Endeca software are faceted navigation and the ability to browse the entire collection without entering the search term. Guided or faceted navigation exposes the relationships between records in the result set. Any of the dimensions (facet) can be used to browse the collection in this fashion, and the ability to assign item-level information (e.g., format, availability, new book), as well as bibliographic-record elements, to the dimensions further enhances the browsing functionality.
13.3.5 Improving subject access

The Endeca-powered catalogue, in addition to addressing classic keyword-searching problems by introducing relevance ranking, spell correction, and stemming, also leverages the "ignored" controlled vocabulary present in the bibliographic records—subject headings and classification numbers—to aid in improving topical searching.

13.3.6 Recommended features

Endeca provides two outstanding features viz; “most popular” sort option and the “more titles like this” feature.

The most popular sort option is powered by aggregated circulation data for all items associated with a title. It provides users a previously unavailable opportunity to define relevance. Apart from NCSU some other libraries are also implemented Endeca recently like Phoenix Public Library, McMaster University in Canada and some others.

Each of the above follow a unique approach, but one thread that is common in all of these are that they involve a desire to go far beyond the capabilities of legacy catalogues and give library users more powerful and appealing tools.

14. Conclusion

The “next generation” library catalogue is intended to be an evolutionary development. It is expected to build on the work already done by the library community, supplemented by the expertise of computer scientists, and driven by the expressed needs of library users. Libraries have often been described as the heart of colleges and universities. This was true because they contained the majority of the data and information needed to do learning, teaching, and scholarship. With the advent of globally networked computers used to create information “born digital” there has been a shift away from books towards smaller bits of electronic information. The “next generation” library catalog is a tool designed to fit into this shifting environment and move librarianship into a more active role when it comes to increasing the sphere of knowledge.

References


About Author

Mr. Vinit Kumar, Junior Research Fellow, Documentation Research and Training Centre, Indian Statistical Institute, Bangalore, INDIA
E- mail :vinit@drtc.isibang.ac.in