Emerging New Roles of Union Catalogues

Janifer Gatenby, Research Integration and Standards, OCLC, Leiden, Netherlands

Abstract

As union catalogues have embraced the web and as pan-union catalogues have emerged and become accessible via Google, Yahoo and other search engines, the role of the union catalogue has transformed. End users are now accessing union catalogues directly. This has had an effect on the way data is presented (FRBR, facets, data to assist selection, specialised catalogues on a national and international scale, international authorities and WorldCat identities). It has also affected physical delivery, electronic delivery and reference services both in presentation and how the services are interlinked and enacted in the background. The standards that underlie the linking of different systems are examined. In the web 2.0 world, services are being created that dynamically combine data from different sources, termed “mashups”; this trend has attracted new users to union catalogue data. The various non library uses of OCLC’s WorldCat data are discussed. The role of union catalogues in collection development and management is also discussed, including collective storage and weeding, identification of rare materials, identification of comparable collections and the significance of data mining to derive new data in a scalable way.

Traditional Roles of Union Catalogues

Originally the main motivation for developing and maintaining most machine readable union catalogues was to achieve efficiency through copy cataloguing. Original cataloguing was expensive and slow and resources sat for months, sometimes years in the back rooms as the cataloguing backlog. Thus union catalogues started as metadata pools, consisting of data loaded from national sources such as Library of Congress and British National Bibliography and members’ original contributions. As a by product, union catalogues were the centre for inter-library loans for the privileged few and were also the base for some cooperative collection management.

The library collection was typically in two parts, monograph and serial. Library catalogues and union catalogues handled monographs well, including books, audio-visual materials, cartographic materials, electronic resources, and so forth. However, the recording of serial holdings was, and still mostly is, only at a title level with article level discovery being done by commercial abstract and indexing services (A&I indexes). Libraries’ serial holdings are recorded in summary in their catalogues and union catalogues, and again in more machine usable detail in their resolvers that lead from articles found in the A&I indexes and the texts themselves.
New Roles – library oriented

Cooperative cataloguing is still highly important and is embedded in the library infrastructure of most countries. Cataloguing backlogs are still a problem as reported by the Research Information network in the UK (2007) and new ways are being sought to improve cataloguing efficiency as explored by the Library of Congress Working Group on the Future of Bibliographic Control (2007). OCLC has launched a pilot project on Next Generation cataloguing. Even so, the focus is shifting away from cataloguing towards discovery, delivery and far more serious cooperative collection management.

End User Services

End user discovery used to be the province of the local library’s online public access catalogue (OPAC) and the union catalogue was a librarian’s tool. This remained the norm until the public started to use Internet search engines in a big way from 2000 onwards. OCLC’s report on The perceptions of libraries (2005) indicated that only 2% of users used a library OPAC or portal as the first point of call whereas 84% started in a search engine. These statistics have been largely digested and as a result many union catalogues have developed freely available public interfaces and they are increasingly exposing their collections in search engines. Examples of union catalogues with public search interfaces include WorldCat, Sudoc in France, GBV in Germany, Bibliotek.dk in Denmark, BibSys in Norway, Libris in Sweden, COBISS in Slovenia, Libraries Australia and Ester in Estonia, but there are many more.

Figure 1 Libraries Australia user interface

OCLC started the movement towards web exposure of library data in Dec 2004 with the launch of the Open WorldCat (OWC) program that was presented to ALA mid winter in Jan 2005. This program includes the major search engines (Google, Yahoo and Microsoft) and also Google Book Search, Google Scholar, online books stores, antiquarian book stores and other sites. Many other libraries and
union catalogues have contracted directly with Google Scholar including union catalogues that have made their data available to Google Scholar as part of its union catalogue program. As at August 2006, these nations were Australia, China, Czech Republic, Denmark, Ireland, Israel, Hungary, Lithuania, Netherlands, Taiwan, United Kingdom and United States. To achieve maximum exposure particularly of the digitised part of the collection and realising the importance of size and comprehensive coverage in attracting end users, many have recently contracted to load to WorldCat including Australia, Czech Republic, Denmark, Finland, Germany, New Zealand, Poland, South Africa, Sweden, United Kingdom and more.

OCLC now exposes WorldCat data in multiple ways; via provision of harvested records to the search engines and other sites, via permalinks which are direct URL links to individual records, and via worldcat.org that makes access available to the complete catalogue (over 95 million bibliographic records in December 2007). In the month of October 2007, in Worldcat.org users passed from the result set page to full record display more than 12.9 million times and from there passed through to local OPACs nearly 1 million times. Of interest is the figure for referrals to WorldCat from Google Book Search. This Google service can only display snippets for in-copyright material, proposing to users either purchase or location in a library. More than a million referrals a month come to WorldCat via this path.

Users are finding WorldCat.org directly as well as via Google and when they find the site they are staying to look around more. This underlines the importance of web presence in terms of user discovery and acceptance.

And so user discovery has moved from the local OPAC to national union catalogues and to the global level, in WorldCat. The union catalogue is becoming more closely tied with local systems: using local resolvers to connect users directly with full text, accessing the local catalogue for circulation availability and
passing users to local reference services. One step further, WorldCat Local, which is in beta implementation in the United States, allows users to access local, regional and international resources in sequence within the one modern interface. Thus the union catalogue is now serving as a discovery tool at the national or regional level and linking in two directions; to local catalogues and also globally. The registry of libraries has become of central importance in linking the union catalogues to local services. As discovery has evolved internationally, so there is now a need for the global library registry being made available by OCLC as a free service.

Additional services for the end user are evident. Some union catalogues are accepting direct user requests for all material in the union catalogue. Examples are Bibliotek.dk in Denmark, Libraries Australia’s copies direct service, NCC in the Netherlands, SUDOC in France, GBV and Hebis in Germany, and BibSys in Norway. Some provide links to reference services and some include personalized services such as the Danish “mit bibliotek”. The Danish, Norwegian and Estonian services include reading lists for academic courses and research projects. The Danish and Slovenians collect loan statistics from their nation’s local systems and combine them with the union catalogue to provide services such as the Slovenian “best read books”.

Direct linking to online booksellers is another new service. There are on average about 100,000 links a month to Amazon from WorldCat resulting in approximately 4000 sales. Libraries Australia is also providing a similar service.

Many union catalogues have also responded to the challenge to provide a modern interface to meet expectations that users have from their experiences with web sites in general. Many are building special interfaces such as Libris, WorldCat and
Libraries Australia. Multiple databases are being accessed including articles, repositories, images and collections from archives and museums to provide a more comprehensive response to users’ searches. The presentation of results is improved by clustering works in the short display and allowing navigation from any one manifestation to all others of the same work (so called FRBR features), and by data that assists in refining results sets including facets, and by data to assist selection, such as audience level, holdings count (indicating rareness or indicating popularity and relevance). Authority data are being presented to the user in more attractive ways, such as People Australia and WorldCat Identities and authorities are being linked on an international scale via the VIAF project.

![WorldCat Identities](image)

**Figure 4** WorldCat Identities
Cooperative collection management

In recognition of user preferences for digital material that can be delivered immediately and remotely, libraries are now realizing that they must shift their emphasis from the physical to the digital collection (National and State Libraries Australasia 2007). To this end they are now increasingly collecting digital materials, buying license access to digital materials and creating their own digital content.

Lorcan Dempsey (2007) separates the library collection now into three parts; the bought collection, the licensed collection and the digital collection. The bought collection is mostly physical, including both monograph and serial, but serials are increasingly part of the licensed collection. The digital collection is largely monographic.

As libraries respond to the challenges of a 3 part collection and a user community with expectations of national and increasingly global discovery and delivery, they are forming into larger cooperatives. Consortia are forming making shared physical collections, cooperating on digitisation and physical and digital preservation. Examples of digitisation cooperatives are numerous, including The European Digital Library, Digital Libraries Federation of Poland and DLF / OCLC’s Registry of Digital Masters. The consortia may operate as independent union catalogues or as subsets of larger catalogues like WorldCat and TEL. Cooperative stores are also emerging and now in the U.S., 21% of library stores are for collectives (Payne 2007 and OCLC Programs and Research 2007). Consortia are also sharing digital resources and rationalizing digital subscriptions, though licensing too often poses barriers to inter-library access.

With physical collections, close library cooperation was geographically bound, often constrained by the limits of an internal courier service. This is not so anymore: libraries can make alliances to share electronically with other libraries anywhere on the globe with similar or complementary collections. Thus international cooperation has moved from the wings to centre stage. This has resulted in a phenomenal growth of WorldCat in 2006 and 2007 with the loading of large union catalogues from Europe, Australasia and Africa and the former RLG database. It has also made possible consortia with an international scope, e.g. based on scripts or materials for special needs groups.

New services to support collection management have appeared. OCLC offers the collection analysis service that allows a library to compare its collection with the entire global collection represented by WorldCat and also with a selection of other libraries. This new knowledge supports purchase, re-location and digitisation decisions.
New mined data such as estimate of copyright and holdings count and new harvested data such as resource usage statistics also support collection management.

More and more materials are being digitised, creating multiple versions of resources in greater proportions than before. This is making the application of FRBR grouping more important than previously and highlighting the importance of reliable international identifiers at all levels of the FRBR tree (authorities, works, manifestations and items).

Library collections are not only growing, they are getting richer and better linked with Internet resources. Many union catalogues now include journal articles or search for them in unison with the union catalogue; WorldCat has 30 million loaded. Catalogue entries are enriched with book covers, reviews, tables of content and URL links to reviews. Authority records lead to works by and about an author, timelines, statistics and biographies in Wikipedia. Worldcat.org allows end users to add reviews, comments and also to create lists. In the first 7 months that worldcat.org enabled users to create lists, 35,000 lists were added. The monthly contribution is growing; in December 2007 the rate was 6,000 a month. Large union catalogues (whether regional, national or international) are capable of attracting user interest and contribution. Enrichment is economically more sensible and viable if done collectively.

New Roles – outside the library arena

Exposed union catalogues have attracted not just end users but other institutions and their systems. New users include search engines, encyclopedias, blogs and book suppliers. In the web 2.0 world, services are being created that dynamically combine data from different sources, termed “mashups”. The use of the union catalogue has multiple faces; union catalogue data may be harvested (e.g. by search engines), linked into via direct link to a record or via a search box, or via permanent searches (e.g. online encyclopedias). The data is working harder than ever before.
The search main engines, by their choice, are harvesting between 3 and 4.4 million records, but covering about 75% of the 1,200,000,000+ holdings in WorldCat. 3 million of the 4.4 million are clustered “work type” records with holdings consolidated from related manifestation records. Google Scholar accepts more, mapping its subset to 67 million clustered records and the mapping from Google Book Search to WorldCat is close to 100%.

In August 2006, WorldCat.org provided a search box facility which is freely downloadable from the OCLC web page. 14 months later, the box had been acquired by over 3,500 sites, many of them blogs that are producing an impressive traffic flow, over 5% of the monthly searches. These sites include Squiggler.com, a general discussion blog which is sending over 4000 referrals a month, mysticbourgeoisie.blogspot.com over 13,000 referrals and websearch.cnn.com over 6,000 referrals.

Book trade organisations are becoming increasingly valuable as union catalogue partners. Publisher’s data, loaded early can considerably lessen cataloguing effort, offering already keyed descriptive information, rich biographical and visual data. OCLC has launched an ambitious pilot project with the publishing, vendor and library communities to capture ONIX data at an early stage in an effort to achieve considerable cataloguing economies (Next Generation cataloging OCLC 2007). If libraries could ease the load of cataloguing of mainstream publications they could give more attention to their special collections and backlogs. Another program, in production at OCLC, WorldCat selection, collects suggestions provided by a growing number of suppliers; libraries consult these centrally and make their selections which are then downloaded to their local system for completion of the order process. This circumvents each library needing to either pre-load the suggestions or to visit each supplier’s web site individually.

Other parties who are now collaborating include online encyclopaedias subscribing to list services and the antiquarian book trade with an interest in rarity for pricing; collection analysis for seeking materials; assessing deceased estates and also contributing enriched metadata.

Museums and archives as holders of unique materials have only recently become interested in cooperation with each other and libraries, as they too build digital collections that are sharable. The Australian Federated Search Project enables web users to search the collections of Australian archives, galleries, libraries and museums through a single search. OCLC’s Programs and Research department is operating a program called Collection sharing beyond libraries that is focused on museums.

Abstract and Indexing databases are accessing the library registry directly to locate resolvers for connecting users directly with full text.

In the Netherlands, the DAI project enables digital repositories to use the name authority file of the Nederlandse Centrale Catalogus (NCC).
Future New Roles?

As libraries cooperate more closely to create common collections, perhaps the acquisitions processes should be centralised in these cooperatives as well. Such common data as suppliers and serial prediction patterns could be shared at a general level and selection/rejection information could be shared among members in a cooperative.

Perhaps now is the time to examine all library data to see if sharing at regional, national or international levels could bring benefits. The factors that indicate that sharing would be beneficial are:

- data needs exposure and is more effective in a web scale site (e.g. bibliographic metadata)
- data is being managed multiple times redundantly (e.g. suppliers details, supplier suggestions)
- through collective knowledge, tasks can be simplified or improved (e.g. serial prediction)
- through a combined database, new knowledge can be deduced or mined (e.g. holdings count indicating rareness and popularity, supplier performance, copyright)
- links and imported enriched data are more efficiently managed through a combined database
- through collective data collections can be compared and assessed and combined collections can be managed

During the examination, the optimum location needs also to be considered:

- globally sharable data (e.g. bibliographic metadata, holdings, issue level holdings, suppliers, statistics, reference query and answer pairs)
- Data that can be shared within one or more cooperatives to which the library belongs (e.g. selection/rejection decisions, weeding reasons, data with specific technical requirements such as scripts, transliteration schemes or specific interface needs)
- Local data that is not shared (e.g. sensitive data such as budgets, invoice details, some user information, dynamic information)

Extending Interoperability – New standards

Further cooperation and data sharing will entail closer systems interoperation between union catalogues and local systems. Exposing library data collectively on national and international scales entails a responsibility for providing a cohesive delivery and reference service. As Kent Fitch (2007) put it: “The ultimate goal for using a discovery service is getting... [and libraries are becoming] great at finding but getting needs work”
This fact points to a level of urgency in identifying standards gaps and filling them as the essential foundation stone of systems interoperability. The fastest way to achieve this layer, and possibly the best way, is to encourage adoption by using existing extensible standards.

To take enquiry as an example, SRU\(^1\) is arguably the most suitable standard to use as it has the best architecture capable of handling result sets and their manipulation. It includes metadata about the result set including record count and result set position. Most important is its extensibility. The two main extensibility components are context sets that allow definition of how to search, especially which indexes are searchable and XML record schemas that determine the format of the records in the response. Some examples of SRU extensions that have already been identified as needed are:

- defining extra result set data to include facets in the form of SRU URLs
- record schema for work level records including embedded SRU URLs for “opening the cluster”
- SRU record context sets for
  - Registry enquiry
  - Holdings
  - Reference Knowledge base
  - Request
- Record schemas for
  - Registry enquiry
  - Holdings (the new standard ISO 20775 has been designed for this)
  - Reference Knowledge base
  - Request

For each distinct data set a search context set and a response data schema will meet the requirements. The process for registering and achieving consensus is considerably easier than for a full standard and can incorporate a trial use period.

Similarly for data addition and maintenance there are existing standards that can be extended to encompass new data schemas. Such standards include OAIPMH, ATOM syndication format and RSS, so called PULL mechanisms where a database makes the data available for external systems to harvest and does not monitor the subsequent use of the data. The widely employed FTP is either a PUSH or a PULL mechanism. SRU update and the ATOM publishing Protocol are PUSH mechanisms designed for a system to update another in background as if it were an online client. SRU update allows for the exchange of diagnostics and linking identifiers. All these standards support multiple data schemas.

Arguably, all interoperation can be modeled as either enquiry or as an update. Even transactions can be divided into notifications of action taken or request for action and these can be conveyed as changed data fields. Thus, if a data schema can be developed and agreed for each class of data to interoperate, the brunt of the standardization will have been accomplished. This has the potential

\(^1\) NCIP does not handle a result set, OpenSearch does not have standard searches and can be viewed as a subset of SRU and OpenURL is not intended to be a search mechanism
to significantly reduce standards efforts and to achieve interoperability as a reality in a more timely way.

**Conclusion**

Union catalogues have evolved to perform a central role between libraries’ local systems and the Internet at large. And now there is also a clear requirement for a combined presence for union catalogues such as TEL in Europe and WorldCat internationally.

Centralised data has diversified, now including not only bibliographic data but also clustered data, enriched displays, external links, authority and statistical data. Institutional registries provide a link between the union catalogue and local library services.

As libraries strive to meet the challenges of a three part collection; physical, digital and licensed, they are relying more heavily on union catalogues to achieve better, richer, shared collections and economies of scale in services and exposure. There is a potential to share more data among cooperatives which will necessitate a new layer of standards. The challenge will be to achieve not only the necessary policy changes in a timely way but also a standards layer to facilitate the interoperability of systems both between union catalogues and local systems and between union catalogues and global services and data stores. There is still a lot to be done to achieve smooth and simple to use interoperable services at national and international levels.

**Glossary**

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>bibliotek.dk</td>
<td>Public interface of the Danish National Union Catalogue <a href="http://bibliotek.dk/?lingo=eng">http://bibliotek.dk/?lingo=eng</a></td>
</tr>
<tr>
<td>COBISS</td>
<td>National union catalogue of Slovenia <a href="http://www.cobiss.net/cobiss_platform.htm">http://www.cobiss.net/cobiss_platform.htm</a></td>
</tr>
<tr>
<td>Collection Analysis</td>
<td>Service provided by OCLC to allow libraries to analyse their collections in relation to other libraries <a href="http://www.oclc.org/collectanalysis/default.htm">http://www.oclc.org/collectanalysis/default.htm</a></td>
</tr>
<tr>
<td>Collection sharing beyond libraries Program</td>
<td>Program of OCLC Programs and Research for investigating needs of museums’ digital collections <a href="http://www.oclc.org/programs/ourwork/collectivecoll/sharecoll/default.htm">http://www.oclc.org/programs/ourwork/collectivecoll/sharecoll/default.htm</a></td>
</tr>
<tr>
<td>Danbib</td>
<td>National union catalogue of Denmark <a href="http://www.dbc.dk/top/top_danbib_eng.html">http://www.dbc.dk/top/top_danbib_eng.html</a></td>
</tr>
<tr>
<td>Digital Libraries Federation, Poland</td>
<td><a href="http://fbc.pionier.net.pl/owoc">http://fbc.pionier.net.pl/owoc</a></td>
</tr>
<tr>
<td>DNB</td>
<td>Deutsche Nationalbibliothek <a href="http://www.d-nb.de">http://www.d-nb.de</a></td>
</tr>
<tr>
<td>Libris</td>
<td>National union catalogue of Sweden <a href="http://librishelp.libris.kb.se/help/highlights_eng.jsp">http://librishelp.libris.kb.se/help/highlights_eng.jsp</a></td>
</tr>
<tr>
<td>NCC</td>
<td>Nederlandse Centrale Catalogus – National Union catalogue of Dutch libraries <a href="http://picarta.pica.nl/DB=2.4/LNG=EN/">http://picarta.pica.nl/DB=2.4/LNG=EN/</a></td>
</tr>
<tr>
<td>OCLC</td>
<td>Nonprofit, membership, computer library service and research organization <a href="http://www.oclc.org/about/default.htm">http://www.oclc.org/about/default.htm</a></td>
</tr>
<tr>
<td>OpenSearch</td>
<td>a collection of technologies that allow publishing of search results in a format suitable for syndication and aggregation developed by A9, an Amazon subsidiary <a href="http://www.opensearch.org/Home">http://www.opensearch.org/Home</a></td>
</tr>
</tbody>
</table>
| Registry of digital masters | Registry of digital preservation masters maintained by the digital Libraries Federation DLF and OCLC  
http://www.oclc.org/digitalpreservation/why/digitalregistry/ |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>RSS</td>
<td>formally &quot;RDF Site Summary&quot;, known colloquially as &quot;Really Simple Syndication&quot;) is a family of Web feed formats used to publish frequently updated content <a href="http://www.rssboard.org/rss-specification">http://www.rssboard.org/rss-specification</a></td>
</tr>
</tbody>
</table>
| SRU                         | Search and Retrieve via URL. Search mechanism hosted by the Library of Congress  
http://www.loc.gov/standards/sru/ |
| SRU update                  | Update mechanism developed by the SRU community, hosted by the Library of Congress  
http://www.loc.gov/standards/sru/record-update/ |
| SUDOC                       | Système Universitaire de Documentation; union catalogue of French universities and institutes of higher education  
http://www.exlibrisgroup.com/voyager.htm |
| TEL                         | The European Library. Catalogue of 44 European national libraries.  
http://www.theeuropeanlibrary.org/portal/index.html |
| VIAF                        | Virtual International Authority File |
| WorldCat                    | Union catalogue of global dimensions managed by OCLC  
http://www.worldcat.org/;  
http://www.oclc.org/worldcat/statistics/default.asp |
| WorldCat Identities         | An author look at WorldCat based on data mining and data mash-ups  
http://orlabs.oclc.org/Identities/ |
| WorldCat Local              | A localized view of worldcat.org  
http://www.oclc.org/worldcatlocal/default.htm |
| WorldCat Selection          | A centralised selection service managed by OCLC  
http://www.oclc.org/selection/default.htm |

References

Dempsey, Lorcan (2007) The network reconfigures the library systems environment. Lorcan Dempsey’s weblog  

Fitch, Kent (2007) New paradigms for getting  


http://www.loc.gov/bibliographic-future/ |


