Joined up data and dissolving catalogues

Chris Todd
National Library of New Zealand

Abstract

External discovery applications, including “Next-generation” catalogues, are a relatively new feature on the library landscape. These 21st century systems, aimed at enhancing the user experience with a range of innovative services, are currently built on 20th century data created using AACR2 and encoded in MARC. But all this is about to change. RDA (Resource Description and Access) is the proposed successor to AACR and due for release in 2009. RDA drafts indicate a very new direction for cataloguing rules and the structure of catalogue records. This paper discusses how these new rules might affect both the design of retrieval systems and the work of cataloguers.

Keywords
RDA ; MARC ; AACR2 ; Next Generation Catalogues ; Semantic Web ;

Introduction

Google and other search engines have had a huge impact on the design of new catalogue interfaces. Interfaces that are much more intuitive for catalogue users, and have more of the features people have come to expect in any resource discovery system. However these interfaces can only work as well as the data they’re based on. While data created following AACR (Anglo-American Cataloguing Rules) and encoding applying the MARC (Machine-Readable Cataloguing) format may seem anything but vague, it is not always explicit enough for machine use. The combination of new interface systems with new rules for descriptive data has the potential to remove the repetition and redundancy from cataloguing and at the same time improve the quality of descriptive data.

Why OPACs Suck

Since before Christine Borgman’s 1996 article ‘Why are online catalogs still hard to use?’ [Borgman, 1996] and well before Google, librarians and library users have been concerned about the quality of OPACs. This concern has increased in recent years. The main problems seem to be that most OPACs are hard to search, give us information we don’t want, use obscure language to describe the things we do want, and then don’t actually deliver the items. Karen Calhoun in her 2006 report [Calhoun, 2006] states that “a large and growing number of students and scholars routinely bypass library catalogs in favour of other discovery tools, and the catalog represents a shrinking proportion of the scholarly information universe”. Karen Schneider’s much-cited blog, and the source of the title for this section, identifies features such as relevance ranking, word stemming, spell-checking, the ability to refine original queries, sort flexibility and faceting as lacking in many (but not all) OPACs. [Schneider, 2006] To this list we can add: logical grouping of results and what Michael Vandenburg describes as “sideways searching (suggestions, expansion of searches and search targets)” [Vandenburg, 2006].

Some commentators have suggested that combining inventory management and resource discovery in a single tool (the OPAC) is no longer effective and these two functions should be more clearly separated [Dempsey, 2006]. This idea is reflected in the effort being put into improving the resource discovery layer, the interface between the user and the data, and this is where many of the items on Schneider’s wish-list are being addressed.

A second area of development is addressing the data itself. The data content and data structures that support (or inhibit) resource discovery are being analysed, questioned and changed.

Next Generation Catalogues

One solution to the OPAC problem is the type of system often referred to as a “next generation” catalogue. [Breeding, 2007] At their simplest these systems provide a new type of interface between the user and the library catalogue and they are an example of the separation of resource discovery from inventory management. Next generation catalogues systems do not include the collection management functions of integrated library systems. They are resource discovery systems that usually link to an existing integrated library system and may also link to other data sources. Systems of this type include Endeca, Primo, Aquabrowser and Encore as well as the National Library of Australia' new catalogue which uses VuFind.

Typically the features of these interfaces can be divided into 3 groups: the first group improve the use of existing metadata in bibliographic records through features such as relevance-ranked searching, faceted navigation, clustering result sets and a range of sorting and limiting options. Clustering mechanisms are sometimes loosely based on the International Federation of Library and Information Associations (IFLA) model Functional Requirements for Bibliographic Records (FRBR). [IFLA Study Group on the Functional Requirements for Bibliographic Records, (1998)] There are also features of the "sideways searching" mentioned earlier such as identifying similar items and the 'did you mean?' option for reviewing a search. This group of features enable users to more easily find identify and select information resources. Find, Identify, Select, and Obtain are the four user tasks identified in the FRBR model.

A second group of features supports the Find Identify and Select user tasks by enriching the data associated with the bibliographic record. This group includes the addition of cover art, contents notes and publishers descriptions. It can also include direct links to external sources of descriptive or evaluative metadata such as reviews and subject tags. Amazon is common here and the National Library of Australia's catalogue at http://catalogue.nla.gov.au/ has links to LibraryThing. In addition there may be the facility for user contribution of reviews, subject tags or other data about a resource.

The third group of features relate to the Obtain user task. This includes direct links to online resources, links to circulation or interloan systems and links to commercial suppliers such as Amazon.

For users the search experience is potentially easier and more satisfying, particularly when a system is used to search across a number of previously separate databases.

AACR2/MARC21: text or data?

"Cataloguing is about control, the OPAC and how it (doesn't) function is about resource discovery. But if we are not getting the data right in the first place, it gets harder and harder to design systems that can use it effectively for resource discovery." [Delitt, 2006] These new catalogue interfaces are firmly based on data that was created following AACR2 and encoded using MARC. Exposing that data to more sophisticated search tools also exposes its flaws. As Delitt & Fitch (2007) say "MARC in particular has become so entwined with other standards, particularly AACR/RDA (and now does contain elements of a content and data standard) that it can be hard to tell where one ends and the other begins."

Some of the flaws exposed by the use of next generation systems are cataloguers' mistakes, others result from the fact that standards have evolved over time, and older records simply do not contain the data present in newer ones. There are also problems that result from the complex relationship between MARC, AACR and other standards used in the creation of bibliographic records. One example referred to in the development of the State Library of Tasmania's TALISPlus opac was the creation of a facet for fiction [Sokvitne, 2007]. Identification of fiction in a standard bibliographic record can be through a classification number (but not always), through subject headings (but not always) through a note (sometimes) and through the use of a MARC code (but only for relatively recent records). This is an example of an obvious piece of information that cannot be easily and consistently derived by a machine from our bibliographic records. Date of publication is another piece of information that raises similar problems.

Another problem is clearly stated by Karen Coyle "Many elements serve more than one function in the bibliographic description, and most of these functions are implicit, not explicit. This has always been the case with library data and it is definitely the case with data we have coded in MARC format" [Coyle, 2008] What this means is that people who understand the rules and codes can interpret the information but computers can't. The title proper in a bibliographic record serves to identify the item being described, it is a display element in most catalogues, it is an access point and it can be used to determine the sort order of a group of records. The Publisher element in a bibliographic record may contain the name of the actual publisher of a work, a sub-imprint or brand of the publisher, a fictitious publisher, or it could be the cataloguer's guess as to the name of the publisher. The way the data is currently recorded there is no sure way of knowing whether the publisher subfield in a MARC record contains the name of the real publisher or not. While this may seem trivial in some circumstances, it reflects the fact that the MARC format was designed for the exchange of records, it was not designed for resource discovery. A catalogue record created in MARC following AACR2 is a very efficient way of conveying a great deal of information, but that information is designed to be interpreted by a person, not a machine. Implicit data is not useful at a time when we want more bang for the bibliographic buck.

RDA

Where next generation catalogues or interfaces are improving resource discovery through better use of existing data, Resource Description and Access (RDA) is a development that addresses the data itself. RDA is the planned successor to AACR2, however it is more than just a revision.

Among the goals of RDA are: to provide a consistent and flexible framework, and to be compatible with internationally established principles, models and standards. [Tillett, 2008] These goals should result in the kind of data that will support the new resource discovery systems described earlier. RDA is based on the Functional Requirements for Bibliographic Records (FRBR) [IFLA Study Group on the Functional Requirements for Bibliographic Records, 1998] and the draft Functional Requirements for Authority Data (FRAD**)** [IFLA Working Group on Functional Requirements and Numbering of Authority Records, 2007]. These models both identify bibliographic entities along with their attributes and the relationships between them and then map these to user tasks. Other important influences are the International Standard Bibliographic Description (ISBD) and the draft Statement of international cataloguing principles [IFLA. Cataloguing Section, 2008]. Two principles of representation that should make cataloguing more efficient have been summarised as:

Take what you see – no more obscure abbreviations

Accept what you get – designed to facilitate the automated capture of data and the acceptance of data from other sources (such as publishers) [Tillett, 2008]

The use of the FRBR entities work, expression, manifestation, and item in RDA will provide the potential basis for logical clustering of records. Many new generation catalogues already provide some grouping of search results based on FRBR, but the new rules will make this much easier and more consistent. For example application of RDA rules will enable the various publications of the Lord of the Rings books to be grouped by language and form and quite clearly separated from the films based on those books.

The treatment of different types of resource in AACR2 combines identification of the content of a resource and the carrier of that content in what is called the General Material Designation (GMD). In RDA content and carrier are clearly distinguished and vocabularies developed for each attribute. For example, the content type text, can be stored on the carrier types sheet, volume, online resource, microfilm cassette, among others. Clarifying these distinctions supports the creation of facets and various browses and limits available in search interfaces.

Recording relationships is a major part of RDA. There are relationships connecting the elements associated with a single information resource, relationships associated with authority data and relationships that link different works, expressions, manifestations or items. Where an AACR2 record can show that there is a connection between, for example, a person and a publication, RDA will make specific whether the role of the person is an editor, illustrator or composer. This provides the potential for more specific role or relationship-based filtering in resource discovery.

One of the most important features of RDA is that it is a content standard and not a display standard or an encoding standard. Because of the close existing relationship between AACR2 and MARC it is anticipated that the MARC format will be used to encode RDA records, but it will also be possible to use other formats such as Dublin Core. This is a clear shift away from AACR2.

RDA the Semantic Web and Dissolving Catalogues

Early in 2007 representatives from RDA, Dublin Core and the W3C Semantic Web Deployment Working Group agreed to begin work on 3 tasks:

1. Definition of an RDA element vocabulary
2. RDA value vocabularies
3. RDA DC application profile

The first task involves making a list of the data elements that have been defined for RDA. Each element is defined and assigned a URI (Uniform Resource Identifier). Elements such as "Title","Language of content" and "Content type" appear in this list.

The second task involves identifying the lists of values, rather like "mini-thesauri", that occur within RDA. These are lists such as the forms of content that would appear in a list of content types, lists of languages, and other terms used in the physical description of resources [Hillman & Dunsire, 2008]. Each term in each of these lists will also have a URI.

The final task is the defining of a set of metadata elements within the Dublin Core framework, for use with the RDA rules for content.

The combination of these three tasks means that computers will be able to use library data much more readily. In a way it's like adding definitions from the rules into the online environment. With this shift our information becomes part of the Semantic Web. We're moving library data beyond libraries and into a world where other systems can interact with it. To use a very simple example: If "Country of publication" is a defined element, there will be a list of authorised country names, where each name has a URI. Using this data, it would be possible to create a map of the world showing the countries represented by publications in a library collection. This is taking data that is part of the bibliographic record and being able to re-use it in a completely different context, because that data has been defined in a way that makes it usable by non-library systems.

If these developments are successful, they have the potential to transform cataloguing. The separation of content and structure from display could remove a huge amount of repetition and redundancy from the cataloguing process. Cataloguers would be more focused on adding value by making connections and analysing content. The cataloguer of a new work might create records reflecting the work, expression, manifestation and item(s) associated with that work. Part of this task would be adding links (possibly in the form of URIs) to authority-like records that identify authors, performers, sponsoring bodies, publishers, etc. as required. The forms of content and carrier for that work could also be expressed as links to vocabularies. If the cataloguer is describing a new manifestation of an existing work, the only information they will be adding is new information. In this situation the work record, with its subject analysis and links to authors, etc. has already been created, so the new manifestation record would be linked to the existing work record. From the perspective of the cataloguer, full bibliographic records would be replaced by a network of connected data.

This means that the way the information is conveyed to the user can be determined more fully by the design of the interface. While the cataloguer may no longer create what we currently view as a full bibliographic record, this view can still be made available to the user through the catalogue interface. The author could be shown to the user using an English name in a Latin alphabet, in Cyrillic, or by an image. The form of the carrier could appear as text and/or symbol and/or cover image.

Conclusion

The initial release of RDA is scheduled for mid-2009 and key players such as the Library of Congress, British Library, Library and Archives Canada and the National Library of Australia are planning evaluations of RDA which could see implementation some time in 2010. And while all this may not happen overnight … if we develop catalogue interfaces and cataloguing standards in the directions identified here, we are likely to be developing systems that can deliver what our users are asking for and link library systems much more readily with the systems of other communities. We are then looking at Google **and** the catalogue, not Google **or** the catalogue.

Acknowledgements

Douglas Campbell for his help with Semantic Web concepts.

References

Borgman, C. (1996). Why are online catalogs still hard to use? *Journal of the American Society for Information Science*, 47 (7), 493-503.

Breeding, M. (2007). Thinking about your next OPAC. *Computers in libraries,* 27 (4), 28-31.

Calhoun, K. (2006). *The changing nature of the catalog and its integration with other discovery tools: Final report.* (2006) Retrieved Sept. 11, 2008, from http://www.loc.gov/catdir/calhoun-report-final.pdf

Coyle, K. (2008). *R&D: RDA in RDF, or Can resource description become rigorous data?* Paper presented at the Code4Lib conference 25-28 February 2008. Retrieved Sept. 2 from http://www.kcoyle.net/code4lib2008_w_text.pdf

Delitt, A. (2006, Sept.). Classification v. cataloging, was Murdering MARC [thread] Retrieved Sept. 7, 2008, from http://osdir.com/ml/culture.libraries.ngc4lib/2006-09/msg00048.html

Delitt, A., & Fitch, K. (2007) *Rethinking the Catalogue* Paper delivered to the Innovative Ideas Forum, National Library of Australia, 19th April 2007. Retrieved Sept. 7 2008, from http://www.nla.gov.au/nla/staffpaper/2007/documents/Dellit-Fitch-Rethinkingthecatalogue.pdf

Delsey, T. (2007). *RDA database implementation scenarios.* Paper prepared for the Joint Steering Committee for Revision of AACR 14 January, 2007. Retrieved Sept. 11, 2008 from http://www.collectionscanada.gc.ca/jsc/docs/5editor2.pdf

Dempsey, L. (2006, May 14). Lifting out the catalog discovery experience. *Lorcan Dempsey's Weblog*. Retrieved Sept. 5, 2008, from http://orweblog.oclc.org/archives/001021.html

Hillmann, D., & Dunsire, G. (2008). *DCMI/RDA Task Group Status Report, July.* Retrieved Sept. 12, 2008, from http://dublincore.org/news/communications/statusreports/2008/09/DCMI-RDA_Task_Group_report_20080720.pdf

IFLA. Cataloguing Section. (2008). *Statement of International Cataloguing Principles*. April 10, 2008 version. Retrieved Sept. 11, 2008 from http://www.ifla.org/VII/s13/icc/imeicc-statement_of_principles-2008.pdf

IFLA Study Group on the Functional Requirements for Bibliographic Records. (1998). *Functional requirements for bibliographic records: Final report.* Munich: K.G. Saur.

IFLA Working Group on Functional Requirements and Numbering of Authority Records. (2007). *Functional requirements for authority data: A conceptual model.* Draft. Retrieved on Sept. 11, 2008 from http://www.ifla.org/VII/d4/FRANAR-ConceptualModel-2ndReview.pdf

Schneider, K. G. (2006, March 13). How OPACs suck, Part 1: Relevance rank (or the lack of it). *ALA TechSource*. Retrieved Sept. 5, 2008 from http://www.techsource.ala.org/blog/2006/03/how-opacs-suck-part-1-relevance-rank-or-the-lack-of-it.html

Schneider, K. G. (2006, April 3). How OPACs suck, Part 2: The checklist of shame. *ALA TechSource.* Retrieved Sept. 5, 2008, from http://www.techsource.ala.org/blog/2006/04/how-opacs-suck-part-2-the-checklist-of-shame.html

Schneider, K. G. (2006, May 20). How OPACs suck, Part 3: The big picture. *ALA TechSource* Retrieved Sept. 5, 2008 from http://www.techsource.ala.org/blog/2006/04/how-opacs-suck-part-3-the-big-picture.html

Sokvitne, L. (2007). *Elephants description and travel: Producing a new OPAC using existing MARC data.* Paper presented at the Libraries Australia Forum, 6 September, 2007. Retrieved Sept. 10, 2008, from www.nationaltreasures.nla.gov.au/librariesaustralia/aum/laf07/documents/Producing_a_new_OPAC.PPT

Tillett, B. (2008). *RDA, Resource Description and Access overview, history, principles, conceptual models.* Paper presented at IFLA Satellite Meeting, Quebec City, August 8 2008. Retrieved Sept. 10, 2008 from http://www.collectionscanada.gc.ca/jsc/docs/iflasatellite-20080808-tillett.pdf

Vandenburg, M. (2006). *The future of the OPAC*. Presentation at the GEAC User Group Meeting, 2-5 May 2006. Retrieved Sept. 5, 2008, from www.gaug.org/conference2006/future_of_opac.ppt