A Bibliographic Metadata Infrastructure for the 21st Century

Roy Tennant
California Digital Library
roy.tenant@ucop.edu

Author's version, published in Library Hi Tech, vol 22 (2) 2004, pp.175-181.

Abstract

The current library bibliographic infrastructure was constructed in the early days of computers – before the Web, XML, and a variety of other technological advances that now offer new opportunities. General requirements of a modern metadata infrastructure for libraries are identified, including such qualities as versatility, extensibility, granularity, and openness. A new kind of metadata infrastructure is then proposed that exhibits at least some of those qualities. Some key challenges that must be overcome to implement a change of this magnitude are identified.

Without question, the development of the Machine Readable Cataloging (MARC) standard in the 1960s was a revolutionary advancement in modern librarianship. It formed the foundation for moving libraries into the computer age by providing a common syntax for recording and transferring bibliographic data between computers. In association with the Anglo-American Cataloging Rules (AACR), MARC allowed libraries to share cataloging on a massive scale, and thus greatly increase the efficiency of the cataloging task as well as set the stage for the creation of centralized library databases such as those managed by OCLC and RLG that are now major worldwide resources.

But that was then. This is now. The technical environment has completely changed from the first days of MARC. When MARC was created, computer storage was very expensive – so expensive that every character was treasured. Very few people had access to a computer – not at work, and most certainly not at home. The Internet was no more than an idea. XML was decades away from being an idea.

In addition, we are no longer dealing only with library catalog systems. Bibliographic records are being used in a variety of computer systems within libraries; for example, interlibrary loan systems, working paper repositories, and directories of online resources such as e-journals and databases. In many cases, MARC is not a good fit for such systems, and the lack of a rich metadata infrastructure finds libraries making up solutions that may prevent them from building an integrated metadata management system.

Also, our cataloging practices have been focused completely on the physical item, rather than the intellectual one. This has led to the creation of, in some cases, dozens of records for items with identical content, thereby sowing confusion and frustration among the users of our systems. Only through the application of the principles laid out in the Functional Requirements of Bibliographic Records (FRBR) do we have some hope of knitting this mess back together on behalf of our clientele. But clearly we can — and must — do better.

We now have the opportunity to recreate our foundational bibliographic standards to take advantage of a new array of opportunities, as well as to fix problems with our current set of standards. It will not be sufficient to tweak our existing standards, since we have been using that method and it is unlikely to provide the scope and scale of change proposed here. We require computer systems, policies, and procedures that allow libraries to create bibliographic metadata, ingest bibliographic metadata from others, make enhancements to it, output it in both complex and simple forms, and do all of this and more with facility and effectiveness. We require a bibliographic metadata infrastructure

that likes any metadata it sees, and can easily output simple records when needed, or complex records when called upon to do so.

What I'm suggesting is different in scope and structure than is implied by my "MARC Must Die" column in Library Journal, although I alluded to it in the follow-up "MARC Exit Strategies" column. What must die is not MARC and AACR2 specifically, despite their clear problems, but our exclusive reliance upon those components as the only requirements for library metadata. If for no other reason than easy migration, we must create an infrastructure that can deal with MARC (although the MARC elements may be encoded in XML rather than MARC codes) with equal facility as it deals with many other metadata standards. We must, in other words, assimilate MARC into a broader, richer, more diverse set of tools, standards, and protocols. The purpose of this article is to advance the discussion of such a possibility.

Infrastructure Requirements

The qualities of the bibliographic metadata infrastructure we require are many, varied, and in some cases, may be in opposition to each other (e.g., simplicity and versatility). Our challenge is therefore not only to build a sophisticated set of standards, protocols, and tools, but also to do it such a way that balances competing priorities. When faced with competing priorities, the needs of our users and our ability to serve those needs should weigh heavier in the balance than our needs for ease of implementation or maintenance.

Versatility

A modern metadata infrastructure should be capable of ingesting, merging, indexing, enhancing, and presenting to the user, metadata from a variety of sources describing a variety of objects. A simple example would be accepting an ONIX record for a book in press, then enhancing that record with information from an OCLC record when it becomes available. Formats as simple as unqualified Dublin Core must be accommodated, as should be records in more complex, granular, and qualified formats. We require an infrastructure that can take in any arbitrary set of metadata and be able to do something useful with it.

Extensibility

Our needs today will not be our needs tomorrow; therefore, we need an infrastructure that will allow for extensions to be developed and applied without breaking the whole. There must be room at the edges for experimentation, since it is often through such experimentation that the way forward is demonstrated. Extensibility can also be a problem, however, when it allows for differentiation beyond what can be accommodated by those relying on the infrastructure. Therefore, extensibility should be crafted to allow metadata consumers to ignore extensions should they wish, without rendering the base metadata unusable. For example, with a metadata record format that allows for multiple, discrete "packages" of metadata within it (e.g., as does the Metadata Encoding and Transfer Syntax or METS standard, see below), if a consumer of such a record wishes to ignore one or more of those packages in favor of others, they can easily do so. A specific example would be a record that has both an ONIX and a MARC or MARC-like package

(e.g., MODS). A library may choose to ignore the ONIX package, while a publisher may choose to do the opposite, and a third party might use both.

Openness and Transparency

To facilitate implementation and extensibility, standards, protocols, and software should be open and transparent as much as possible. Efficiencies of sharing solutions and code can be realized if solutions are offered to others as open source without restrictions that prevent their useful implementation. Transparency is important for potential implementers to see how systems work (e.g., sharing of source code, human-readable metadata formats, etc.).

Low Threshold, High Ceiling

We need a metadata infrastructure that will allow as many people and organizations to participate as possible, which means a system that can accommodate simple uses. But that same infrastructure should also support the more complex requirements of those needing a more full-featured system. The challenge will be to architect a system that can accommodate such diversity without needless complication for low threshold users, nor prevent more complex activities for those requiring a high ceiling.

Cooperative management

No single organization should own the essential pieces of a new bibliographic infrastructure. In particular, the creation and ongoing management of new metadata standards should occur in as cooperative and inclusive process as is practicable. The METS draft metadata standard is a useful example of such cooperative standards

development, in which a number of research libraries are participating through the Digital Library Federation in a process managed by the Library of Congress.

Modularity

The systems we use to create or ingest metadata, and merge, index and serve up or export that metadata should be modular in nature. That is, with a modular system it is possible to replace a component that performs a specific function with a different component, without breaking the whole. For example, a metadata infrastructure that uses XML should be constructed in such a way that whichever XML parser is being used can be swapped out for a different one when needed, without adversely affecting other parts of the infrastructure.

Hierarchy

A modern bibliographic metadata infrastructure must be capable of handling hierarchical information. For example, the table of contents of a book is inherently hierarchical, and there is no good place to put this data in the MARC record. But given an appropriate metadata infrastructure (see below), hierarchy could be handled very easily.

Granularity

Granularity is a key quality of metadata. If a personal name is encoded as:

```
<person>
    <name>Gabriela García Márquez</name>
</person>
```

rather than something like:

```
<person>
```

```
    <name type="family">García Márquez</name>
    <name type="given">Gabriela</name>
</person>
```

it will be difficult for software to process names consistently and correctly. Therefore, metadata must be of a sufficient granularity to support all intended uses. Metadata can easily be insufficiently granular, while it would be the rare case where metadata would be too granular to support a given purpose (for more discussion of granularity, see "The Importance of Being Granular", *Library Journal* 127(9) (May 15, 2002) p. 32-34).

Graceful in Failure

After experiencing the rather forgiving search systems offered by Internet search systems such as Google™, many of our users are likely dismayed to learn how easy it is to fail when searching our library catalogs. Many of our systems will return zero hits rather than do the best that can be done with what is entered. Modern search systems are capable of offering alternate spellings, returning hits ranked by the number of entered terms that are found in the records, or even performing the search using a different index after failing in the selected index. But such features are still rare in most of the bibliographic metadata search systems we offer our users.

A Proposal

We do not need a bibliographic record *format*. We need a bibliographic metadata *infrastructure* that has a number of components, each of which may have multiple variations. Our systems must be able to accommodate a great diversity of record formats to provide us with the flexibility and power that only such diversity can provide.

Therefore, although I touch on specific metadata formats that are in use today, or that promise to be useful in the future, it is not meant to be an inclusive and exclusive list. Rather, this proposal is aimed at creating an environment that is welcoming to — and effective for — metadata formats yet to be created. Should we do our work well, choosing to use a new metadata format will not require us to make substantial changes to our underlying infrastructure. A robust metadata infrastructure should be able to accommodate new metadata formats by creating or applying tools specific to that format, explained in greater detail below.

Transfer Schema

The transfer schema (for which clearly XML is the most reasonable solution) must be able to accept any arbitrary package of metadata. We need a method to pass records that may have metadata containers using ONIX, MODS, Dublin Core, or virtually any other format.

A draft standard that does just this is the Metadata Encoding and Transfer Syntax (METS, see also related articles in this issue). Figure 1 illustrates a METS record with all major segments of the record collapsed. Note how one container holds a MODS record, consisting of a translated MARC record from the UC union catalog, while another holds

a record called "ucpress", consisting of bibliographic metadata from an in-house database at the University of California Press.

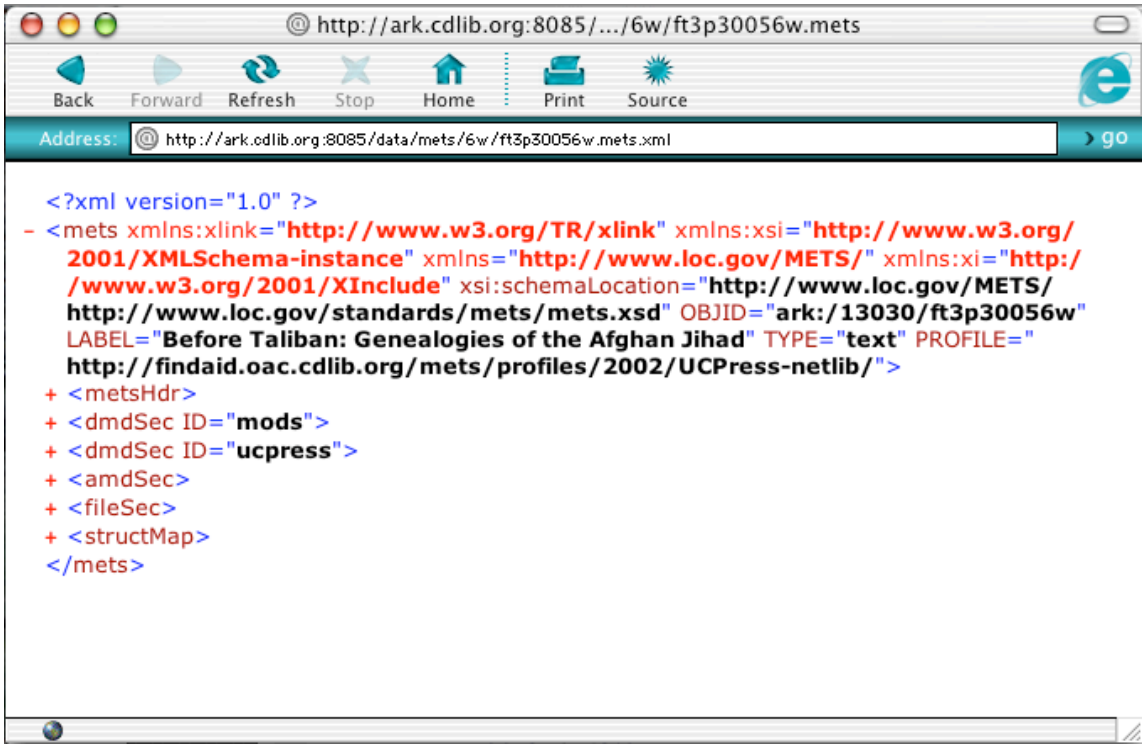

Figure 1. A collapsed view of a METS record.

This example illustrates how a transfer syntax like METS can carry containers of metadata adhering to different standards, or indeed no standard at all, and be associated with the same object. In this particular case, fields are indexed from both records for user searching and display.

Bibliographic Schemata

As mentioned above, we need the ability to ingest, manipulate, and output metadata in a variety of formats. Some of these formats will initially include MARC, MODS, Dublin Core, and ONIX. There are many others, and still more that have yet to

be developed, all of which may eventually need to be accommodated in some way. These various bibliographic schemata must be welcome within our bibliographic metadata infrastructure, and be able to be made searchable, displayable, and exportable.

Application Rules

Schemata alone will be insufficient — we will also require rules and guidelines on their application and use. We will likely need general rules, as well as schema-specific rules, similar to the way that MARC has been the encoding and transfer syntax of the cataloging rules expressed in AACR2.

Best Practices

Beyond specific rules that must be followed for compliance, there exists a grey area where implementations may vary. This is both a good and bad thing. The good aspects have to do with the ability to experiment, to make adjustments for local needs, etc. Where this becomes "bad" is when local variances harm interoperability. Therefore, it will be helpful to build a set of "best practices" beyond the scope of application rules, that illustrate the best ways to implement a given infrastructure component.

Crosswalks

I have recently said that librarians must be able to say "I've never metadata I didn’t like" — or that we can walk, talk, eat, and drink metadata of all varieties. To be proficient at this will require crosswalks, or algorithms for translating metadata from one encoding scheme to another in an effective and accurate manner. A number of crosswalks already exist for formats such as MARC, MODS, and Dublin Core. Besides using

crosswalks to move metadata from one format to another, they can also be used to merge two or more different metadata formats into a third, or into a set of searchable indexes.

Indexing and Display

A heterogeneous metadata infrastructure presents particular challenges to effective indexing and display. When can a field in one metadata format be treated the same as a field in another? How can we logically deal with significant variances in the metadata we wish to search and display as a unified whole? How do we rectify differences in metadata quality, encoding practices, and granularity? Likely we will need to use a variety of strategies depending on the situation. Crosswalking may be sufficient in some cases, while on the other extreme we may find that only human intervention will fix some problems.

Enrichment

A robust metadata infrastructure will offer opportunities for metadata enrichment — both human and machine-based. For example, book records could be enriched with such things as book reviews, cover art, and the table of contents. These items are already making it into some library systems, but with a robust infrastructure they could also be augmented by such things as robot-collected metadata — wherein software queries other systems and collects relevant metadata to add to the record, in a special encoding for what may be only partially trusted information.

Tool Sets

As we begin to build and use this new metadata infrastructure (as is already happening at OCLC, RLG, and large research libraries), we will begin to accrete tools

that can be used to create and manage our metadata systems. For example, XSLT stylesheets for parsing records from one format to another, from XML to an HTML screen display, etc. These tools can be made available to others, and thus enable other libraries to implement this new infrastructure with greater facility and ease. We are already seeing this happen with the Library of Congress making available tools for translating MARC records into MODS, OCLC making available its FRBR algorithm, and METS implementers offering tools for METS record creation and translation.

Relationships with Other Standards and Protocols

Given an appropriate container/transfer format, virtually any bibliographic metadata format could be accommodated by a well-architected metadata infrastructure. Therefore, existing standards such as MARC (as expressed in XML), Dublin Core, as well as emerging standards such as MODS can all be used as carriers of bibliographic metadata. This will enable us to absorb our legacy systems while also offering new opportunities hitherto impossible.

Interoperability and access standards such as the Open Archives Initiative Protocol for Metadata Harvesing (OAI-PMH) and the Simple Object Access Protocol (SOAP) are likely candidates for support in a full-featured metadata infrastructure. These protocols offer a low-overhead way to make bibliographic metadata available to others, for services such as federated searching.

Implementation Issues

Large professional organizations such as OCLC, RLG, and ARL, the Library of Congress, large research libraries, and imaginative and committed individuals must lead

the way. Luckily, they mostly already are. One of the prime examples of leadership in this area is the development of METS. Springing from a real need to have a metadata container capable of ingesting and preserving the richness of a variety of metadata standards, as well as the structure of a complex digital object or set of objects, the METS development effort holds great promise for the kind of metadata infrastructure I envision here. The leadership in developing this standard comes from the sources named above, which is no surprise. Those kinds of organizations are both the best suited for such activities (having generally more resources to apply), as well as the most in need of such cutting-edge solutions for digital library problems.

Challenges

Moving from a bibliographic infrastructure that is relatively homogenous (MARC21 and AACR2) into a diverse universe of metadata managed and controlled by a variety of library and non-library groups will clearly have its challenges. This short list of challenges is unlikely to be complete, but it may serve as the beginning of an honest assessment about what we must address to achieve the desired state as outlined in this article.

Adapting to a Diversity of Record Formats

In moving into the brave new world I describe here, we will be leaving the familiar shores of MARC and venturing out into an ocean where we must be able to deal with just about anything that comes our way. For example, if we want to provide searching of working papers to our clientele, we will need to be proficient with the OAI Protocol for Metadata Harvesting and the Dublin Core metadata standard. If we wish to make tables of contents, book covers, book reviews, and other types of information

available for the items we own, we will find a need for new metadata standards that will more easily and effectively accommodate such features (yes, many libraries and vendors are making MARC stand on its head to do these things now, but if they are based on MARC, they are stop-gap solutions that do not provide a strong foundation for the future).

OCLC has already begun laying the foundation for a diversity of bibliographic records formats and types, by rebuilding WorldCat® from the bottom up. "Extended WorldCat" as it is called by OCLC staff, stores records using an internal XWC (for Extended WorldCat) XML-encoded format in an Oracle 9i database. Although presently only taking in USMARC and Dublin Core records, this infrastructure can potentially include records of a variety of types. The goal is to be able to accept virtually any bibliographic record, provide searching and display of the record, and output it in its original format when called upon to do so. This effort appears to be one of the first major projects to create something similar to the bibliographic infrastructure described here and will likely provide some early lessons on what works and what does not.

Cross-walking and Merging

Taking records for the same object from different input streams and formats and making a merged record that retains the best of the granularity and qualification of the original records is clearly a challenge. But add to that the necessity of creating indexes, search result displays, etc. and the breadth and depth of the challenge begins to become clear.

OCLC has done some interesting work in the area of crosswalking in their Metadata Switch project. The idea is to create a software service that can take a record in

one format as input, and output that record in a different metadata format. This service would logically be offered via a Web Services interface, so that the entire interchange can happen using software only. Such a service would allow distributed systems to take advantage of a robust central infrastructure for record translation and crosswalking. Early findings in this project suggest that while some records can be crosswalked in a straightforward manner, others will require first mapping them to an "interoperable core" before the translation process can be completed. (Godby, Smith, and Childress, 2003). As is the case in many situations, to appear simple from the outside there must be sufficient internal complexity. OCLC's experience appears to indicate that we have not yet plumbed the full extent of the required internal complexity to create a simple service for metadata translation.

Accurate record merging is a challenge even with a relatively homogenous data stream (e.g., MARC and AACR2), but with heterogeneous record formats and rules for applying those formats, it is a challenge that may only be partially met for quite some time. The International Standard Text Code (ISTC) may help, as may perhaps the algorithms being developed in support of implementing the concepts of the Functional Requirements for Bibliographic Records (FRBR). But widespread implementation will take time, and meanwhile we'll need to do the best we can with what we have.

In addition, "merging" can have different meanings depending on the result desired. One type of merging takes two or more metadata records for an item and merges them into one record that is not intended to be displayed or exported as separate records again (i.e., "unification"). Another type of merge would retain the information required to reconstruct the separate records again (i.e., "federation"). Federation of records would

be required if a system must be able to provide the original records from which the merged version was created (for example, if different contributing organizations needed to maintain their version of the record).

Indexing different record formats into a single index will require crosswalking different fields into the same virtual index for searching. Where record formats have fields not found in other formats, or that have metadata that is of a different granularity (e.g., no distinction between first and last personal names), there will be problems.

The challenge of display can conceivably be met by the provision of different display profiles for different types of records, but doing this in a way that will not be confusing to the user will again be a challenge. It may be easier to create summary displays or brief records that appear relatively homogenous, but full record displays will likely exhibit more divergence.

System Migration

To migrate from systems based on MARC/AACR2 to the infrastructure proposed here is clearly a significant undertaking. As anyone who has ever been involved with migrating from one integrated library system to another knows, even moving from one system based on MARC/AACR2 to another can be daunting. Within this context, the changes proposed here must clearly be fostered by cooperation at a national, and perhaps international, level and carefully staged. However, this proposal is about *inclusion* if it's about anything, and therefore our existing records can certainly be included, albeit in an envelope that can accommodate other record formats.

But despite the very real challenges of a systemic and widespread migration to a new kind of metadata infrastructure, I believe that it is both necessary and achievable. We

can no longer afford to have systems that are inadequate to meet both the challenges and opportunities that currently face libraries.

Staff Retooling

One of the most significant barriers to the implementation of this proposal is ourselves. Most of us in the profession today have never known anything but MARC and AACR2 as an online metadata infrastructure. But now we must dramatically expand our understanding of what it means to have a modern bibliographic metadata infrastructure, which will clearly require sweeping professional learning and retooling. Such a vision may be daunting when viewed as a whole, but when attacked piecemeal over time, there is indeed hope for achieving it.

There are already hopeful signs that librarians are rising to the challenge before them, whether by participating in metadata standards development activities such as the Dublin Core and METS efforts, or simply in learning more about metadata issues by reading and attending conference presentations.

The Once and Future Infrastructure

With a robust bibliographic metadata infrastructure as a foundation, many things become possible that may have been more difficult or even impossible with the type of single-stream infrastructure we presently have.

There is no doubt that engineering such an infrastructure will be a long and difficult task. However, the potential benefit to both libraries and library users is likely to be both substantial and long-lasting — particularly if it is constructed with the essential qualities of extensibility and flexibility.

Also, we are apparently already on the path to a better future, with important early

work in process both within key organizations (e.g., OCLC) and among them (e.g., the cooperative METS effort). Likewise, individual librarians are learning how to use technologies like XML and XSLT that will form the foundation of their new bibliographic tool set.

These are hopeful signs that we are beginning to muster both the political will and technical skill to support the type of massive change proposed here. Having not been a part of the effort to create MARC those many decades ago, I cannot imagine what conditions fostered its birth. But in my ignorance I imagine that the opportunities created by computers inspired Henriette Avram and company to rise to the challenge of recreating our professional infrastructure in a revolutionary and farsighted way. We would do well to look to our past for the inspiration we need to create a future that our descendants will look back upon with similar amazement.

References

Dublin Core, <http://dublincore.org/>.

Functional Requirements of Bibliographic Records. International Federation of LibraryAssociations, München: K.G. Saur,1998. <http://www.ifla.org/VII/s13/frbr/frbr.htm>

Godby, Carol Jean, Devon Smith, and Eric Childress (2003), "Two paths to interoperable metadata," 2003 Dublin Core Conference, Seattle, WA, September 28-October 2, 2003, <http://www.ischool.washington.edu/dc2003/>.

International Standard Text Code (ISTC), <http://www.nlc-bnc.ca/iso/tc46sc9/istc.htm>.

Machine Readable Cataloging (MARC). Library of Congress,

<http://www.loc.gov/marc/>.

Metadata Encoding and Transmission Standard (METS), <http://www.loc.gov/standards/mets/>.

Metadata Object Description Schema (MODS), <http://www.loc.gov/standards/mods/>.

Metadata Switch, <http://www.oclc.org/research/projects/mswitch/>.

Online Information Exchange (ONIX), <http://www.editeur.org/onix.html>.

Open Archives Initiative — Protocol for Metadata Harvesting (OAI-PMH), <http://www.openarchives.org/OAI/openarchivesprotocol.html>

Scorpion Project, OCLC Office of Research, <http://www.oclc.org/research/software/scorpion/>.

Simple Object Access Protocol (SOAP). World Wide Web Consortium, <http://www.w3.org/TR/SOAP/>.

Tennant, Roy. "The Importance of Being Granular," *Library Journal* 127(9) (May 15, 2002) p.: 32-34, <http://libraryjournal.reviewsnews.com/index.asp?layout=article&articleId=CA216337>.

Tennant, Roy. "MARC Must Die," *Library Journal* (October 15, 2002), p:26-27 <http://libraryjournal.reviewsnews.com/index.asp?layout=article&articleid=CA250046>.

Tennant, Roy. "MARC Exit Strategies," *Library Journal* (November 15, 2002), p:27-28 <http://libraryjournal.reviewsnews.com/index.asp?layout=article&articleid=CA256611>.