Sally H. McCallum

Library of Congress Metadata Landscape

Sally H. McCallum

The Library of Congress (LC) has many of the same challenges as other libraries, especially large ones. LC has many different types of resources – books, journals, maps, music, manuscripts, audio, moving image, still image, artifacts, electronic – with large collections of each. Different levels of access are needed for this material: for some, collection level bibliographic description is adequate; for many, item level access is adequate; but for others, such as sound recordings, analytic, or sub unit access is highly desirable. The sizes of the LC collections are a major challenge – over 125 million non-electronic and over 3 million electronic items (and growing rapidly). And finally, electronic resources are presenting us with new issues – from metadata to preservation to storage to linking techniques.

LC has tried to approach these challenges from a service perspective. Access must be successful for the end user, which mandates as much coherence and consistency in the metadata as possible and access systems that are easy to use.

This paper focuses on the Library of Congress' perspective on metadata in the following three areas: (1) descriptive metadata in our current operations, (2) pathways that are developing that will support possible evolution in the future, and (3) broader metadata needs with digital material. The discussion is from a metadata element set and format point of view, not a cataloging data and cataloging rules view. Most acronyms used in this paper are expanded in an Appendix.

Descriptive metadata at LC

Currently the primary access tool for the LC collections is the OPAC (Online Public Access Catalog). Its over 13 million records have a relatively high level of consistency. The application of the same standards, with variation of levels where logical, is the key to this consistency.

OPAC record level	Content rules	Tagging system
Full level	AACR	MARC 21
Minimal level	AACR	MARC 21
Initial level (acquisition record)	AACR-like	MARC 21

Some collections that are represented by collection level records in the OPAC connect to supplementary finding tools that provide more detail or analytic level information to the user. These tools are built with the same standards as the catalog, or use closely harmonized rules.

Supplementary tools	Content rules	Tagging system	Used for special materials:
PPOC catalog	AACR	MARC21	Photograph collections
SONIC catalog	AACR-like	MARC-like	Sound recording collections
Finding aids	LC local, harmonized	EAD	Various collections of manuscripts, music, photographs
InQuery	mixed	internal	Digital conversion collections

The LC system environment also makes available to users analytical information for serial contents through access to abstracting and indexing data services.

In addition to the primary OPAC and supplementary catalogs indicated above the LC OPAC contains over 5 million name authority records (online Name Authority File [NAF]) and 300,000 records for subject authorities (online Library of Congress Subject Headings [LCSH] thesaurus). For building and maintaining all these access tools, LC has 450 catalogers and acquisitions specialists trained in the use of these standards, and a large integrated cataloging system and several smaller ones in place.

Like other libraries, LC tries to derive cataloging where possible even though LC does a great deal of original cataloging. LC gets a large number of initial level records from the book suppliers – MARC 21 records with varying levels of detail are sent along with the material purchased. The ISO information retrieval protocol Z39.50 is used to carry out copy cataloging from OCLC, RLG, and other Z39.50 accessible sites that can send back MARC 21 records. LC is starting to use ONIX records to augment the content of our MARC records with table of contents, reviews, and other enrichments. Several specialized sources are tapped for initial or minimal level records for sound recordings, maps, and moving images. For digital conversion objects the record for the non-electronic version of the item can often be multi-purposed. And LC is beginning to derive metadata from the electronic objects themselves, although that work has focused mainly on technical metadata thus far. LC's catalogers normalize derived records obtained from others as they are added to the LC catalogs.

With the electronic material, a new emphasis for the catalog record is enabling linkage to electronic resources from the catalogs. LC currently does this with explicit links in the bibliographic records. These elec-

tronic resource links are made persistent by assigning handle-type identifiers and accessing the resource through a handle server. LC is also experimenting with OpenURL linkage.

LC makes its bibliographic control task as easy and economical as possible by staying focused on stabile standards, AACR and MARC 21, as the primary ones for descriptive metadata. This has been very important for enabling us to derive cost savings from bibliographic metadata. However, some of the questions that arise with the opportunities and options in the new Internet environment are the following. Is this scalable to electronic resources? Do all resources need the same kind of treatment? How about proliferating metadata schemas? And the overarching question: how can both an evolutionary pathway and standardization be maintained?

Descriptive metadata evolution

The web and XML developments have spawned a great deal of diversity that could be useful and destructive at the same time. For data content, libraries use various rules, for example, AACR, RAC, EAD content guidelines, Dublin Core (DC) data, and ONIX content guidelines. These content rules have overlapping concepts but often-unique approaches to the specification of the concept. Does a MARC main entry = a DC creator = an ONIX contributor? However, it appears that the library community is maybe converging on AACR/ISBD-type descriptive content rules.

For markup, libraries are confronted with various tag sets: MARC 21, DC, ONIX, MAB, Unimarc, etc. In addition, the HTML tag set is the markup for the web and the EAD has its own set of tags. XML tag sets are so easy to establish by writing a schema or DTD that there will certainly be more. However, the library community has been converging on MARC 21 and EAD for the item level and collection level cataloging. Are publishers going to converge on ONIX, and has DC got staying power for cross-domain interoperability?

And finally, information specialists are seeing different structures for data records. MARC 21 uses the ISO 2709 structure, people are finding Microsoft Access to be convenient for certain metadata purposes, DTDs and schemas support the SGML, XML, and HTML family of structures. However, there seems to be a definite convergence on XML and schemas globally at the current time.

Thus with proliferating electronic material, an economically deep commitment to MARC data elements, proliferation of schemas beyond the library community control, and the rapidly growing XML tool environment there needs to be an evolutionary pathway forward for LC and libraries in general. At LC this has been interpreted to indicate that there is a need to take advantage of XML by establishing MARC 21 in an XML structure; a need for a compatible but simpler companion to MARC 21 in XML; a need for a coordinated set of tools for record transformations; a need for flexible transition options for the future. The architecture for an evolutionary pathway to XML by the MARC 21 community might be the following. The components of this suite are described below.

need to establish MARC 21 in an XML structure

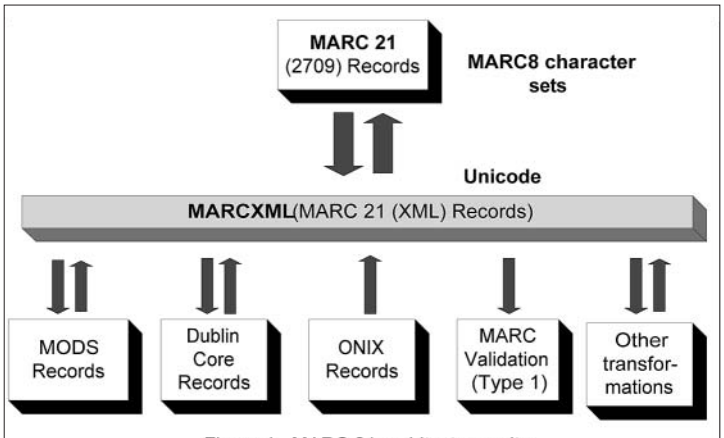

Figure 1: MARC 21 architecture suite

MARC 21 (2709) records. At the top of this architecture is the MARC 21 record in the ISO 2709 structure. This is the record as it has been known for more than 30 years. There is an installed base of thousands of MARC 21 based systems – OPACS, integrated full function, product production systems, etc. There are over a billion MARC 21 records in local and network systems worldwide. These records are accessible by 100s of Z39.50 clients, and thousands of librarians »speak« MARC 21. The record is compact and simple.

thousands of librarians »speak« MARC 21

A machine view of the MARC 21 record would be the following. The title of the item represented by this record is *Germany by Bike*. In this example the title field and the directory entry identifying and pointing to it (tag 245) are highlighted.

00637cam 2200193 a 4500001000900000005001700009008004200026020005300068
040001800121050002400139082002200163100003000185**24500740021**5260004400289
3000035003334400012003685000020003806500043004 00*
93047676*19990429094819.1*931129s1994 wauab 001 0 eng *
$a0898863872 (acid-free, recycled paper) :c14.95* $aDLC$cDLC$dDLC*
00$aGV1046.G3$bG47 1994*00$a796.6/4/0943$220*1 $aSlavinski, Nadine,
$d1968-***10$aGermany by bike :$b20 tours geared for discovery /**
$cNadine Slavinski.* $aSeattle, Wash. :$bMountaineers,$cc1994.* $a238 p. :
$bill., maps ;$c22 cm.* 0$aBy bike* $aIncludes index.* 0$aBicycle
touring$zGermany$xGuidebooks. *#
*(Graphic substitutions: subfield code: $; end-of-field: *; end of record: #)*

MARCXML (MARC21 [XML] records). The MARC record content and semantics can also be carried in an XML data structure. Since XML is a generalized structure, there are many different ways the MARC record could be adapted to XML – experimentation has occurred since the early 1990s, including an SGML version of MARC 21 published for trial use by LC in 1996 (and later converted to an XML DTD). Conversations with and requests from the community of users convinced LC that it would be important to establish a standard MARC in XML schema for the MARC 21 record and make it available from the MARC 21 documentation web site in order to avoid having many similar but not quite the same schemas in circulation. A key characteristic of this standard, called MARCXML, is that it is an exact equivalent of the MARC (2709) record so that roundtrip conversion to and from it is lossless.

lossless roundtrip conversion

```
xmlns="http://www.loc.gov/MARC21/slim">
<record>
        <leader>00637cam  2200193 a 4500</leader>
        <controlfield tag="001">93047676</controlfield>
        <controlfield tag="005">19990429094819.1</controlfield>
        <controlfield tag="008">931129s1994    wauab       001 0 eng   </controlfield>
        <datafield tag="020" ind1=" " ind2=" ">
                <subfield code="a">0898863872 (acid-free, recycled paper) :</subfield>
                <subfield code="c">$14.95</subfield></datafield>
        <datafield tag="040" ind1=" " ind2=" ">
                <subfield code="a">DLC</subfield>
                <subfield code="c">DLC</subfield>
                <subfield code="d">DLC</subfield></datafield>
        <datafield tag="050" ind1="0" ind2="0">
                <subfield code="a">GV1046.G3</subfield>
                <subfield code="b">G47 1994</subfield></datafield>
        <datafield tag="082" ind1="0" ind2="0">
                <subfield code="a">796.6/4/0943</subfield>
                <subfield code="2">20</subfield></datafield>
        <datafield tag="100" ind1="1" ind2=" ">
                <subfield code="a">Slavinski, Nadine,</subfield>
                <subfield code="d">1968-</subfield></datafield>
        <datafield tag="245" ind1="1" ind2="0">
                <subfield code="a">Germany by bike :</subfield>
                <subfield code="b">20 tours geared for discovery /</subfield>
                <subfield code="c">Nadine Slavinski.</subfield></datafield>
        <datafield tag="260" ind1=" " ind2=" ">
                <subfield code="a">Seattle, Wash. :</subfield>
                <subfield code="b">Mountaineers,</subfield>
                <subfield code="c">c1994.</subfield></datafield>
        <datafield tag="300" ind1=" " ind2=" ">
                <subfield code="a">238 p. :</subfield>
                <subfield code="b">ill., maps ;</subfield>
                <subfield code="c">22 cm.</subfield></datafield>
        <datafield tag="440" ind1=" " ind2="0">
                <subfield code="a">By bike</subfield></datafield>
        <datafield tag="500" ind1=" " ind2=" ">
                <subfield code="a">Includes index.</subfield></datafield>
        <datafield tag="650" ind1=" " ind2="0">
                <subfield code="a">Bicycle touring</subfield>
                <subfield code="z">Germany</subfield>
                <subfield code="x">Guidebooks.</subfield></datafield>
</record>
```

The schema is simple and flexible, there is no need to change it as fields or subfields, for example, are added to the format. Presentations of the data can be made from the MARCXML record with an XML stylesheet.

LC is also providing converters for transforming data from MARC 21 (2709) to MARCXML and back – converters that can be downloaded from the MARC website and used by others in their own systems where they can also shape them to their own data and needs (see www.loc.gov/marcxml). This key conversion software between MARC 21 (2709) and MARCXML is adapted from part of an extensive set of programs for manipulating MARC 21 data developed by Bas Peters in the Netherlands and made available by him as open source software.

MARCXML is appropriate for use in harvesting using the Open Archives Initiative (OAI) protocol. The OAI protocol version 2.0, issued in June 2002, changed its MARC recommendation from a project specific MARC schema (oai_marc) to the MARCXML schema. LC is exposing metadata associated with its American Memory digital collections for OAI harvesting in MARCXML. In 2003, LC is also planning to begin to offer distribution of MARC 21 cataloging records in the MARCXML schema, in addition to the ISO 2709 structure, even though LC's record distribution service expects 2709 to be the preferred format for a number of years to come.

Transformed via the conversions available from the MARCXML web site, the MARCXML version of the 2709 example record would look like this table (see on the left).

Metadata Object Description Schema (MODS). While MARCXML is the primary tool for MARC records, there has been a general request for a simpler data schema that could be used with MARC for, primarily, metadata associated with electronic documents. Thus a MARC 21 companion, the Metadata Object Description Schema (MODS) was developed with a group of MARC 21 users. It was available for review and comment for 6 months in 2002 and a new version for trial use, that incorporates the changes suggested during review, is now available (see MODS version 2.0 at www.loc.gov/mods).

MODS is a reduced data element set from full MARC 21, but it generally follows MARC semantics. MODS records are thus highly integrable with MARC 21 records, providing flexibility for use in large bibliographic data systems. The element set is richer than that for simple Dublin Core. The schema organization and tagging are user friendly – there are few coded values and words are used instead of numbers for tags. Based on the experience of reviewers from several digital projects, there are some special accommodations for electronic

resources. These features include: allowing external linking (the Xlink attribute) with data elements throughout the schema; a »related item« structure that supports the hierarchy needed for complex digital objects; a digital origin attribute; several data types specifically for digital projects (e.g., capture); and accommodation of e-resource identifiers, such as the DOI.

The Library of Congress is already using MODS in several ways – in web archiving projects, with a large project involving digitized audiovisual material, and for the folk life center's multimedia digital initiatives. As intended, it is particularly useful where LC needs to use technician or student input of metadata. LC is also using it with other XML-based projects such as descriptive metadata for METS documents, and, as an option, for the metadata sent by the OAI server.

The MODS schema is available from the MARC/MODS web site, which also offers transformation software to and from MARCXML for download and use (see www.loc.gov/mods). The above record transformed into MODS by the software from the web site looks like this:

Dublin Core (DC). In the MARC architecture suite, LC is also providing other downloadable transformations that are useful in the XML environment. One is MARCXML to DC and vice versa. LC has been providing a mapping between MARC 21 and DC for several years and these transformation stylesheets are an extension of that service. Some key DC application targets that make it useful for librarians to be able to transform the DC metadata to and from MARC 21 include cross domain initiatives and an expected wider use of DC for metadata in web document headers. Since the data element detail is vastly different between MARC 21 and DC there could be many different ways to map and convert data. By offering standard transformations for download the variation and incompatibilities that would inhibit interoperability could be minimized. LC is working only with the simple DC as qualified DC has many variations and may never be highly standardized since qualification is the mechanism by which projects adjust DC to meet their special needs. LC uses these transformations to offer a DC option for

mapping between MARC 21 and DC

```
<mods>
        <titleInfo>
                <title>Germany by bike :</title>
                <subTitle>20 tours geared for discovery</subTitle></titleInfo>
        <name type="personal">
                <namePart>Slavinski, Nadine</namePart>
                <namePart type="date">1968-</namePart>
                <role><text>creator</text></role></name>
        <typeOfResource>text</typeOfResource>
        <originInfo>
                <place>    <code authority="marc">wau</code>
                           <text>Seattle, Wash</text></place>
                           <publisher>Mountaineers</publisher>
                <dateIssued>c1994</dateIssued>
                <dateIssued encoding="marc">1994</dateIssued>
                <issuance>monographic</issuance></originInfo>
        <language authority="iso639-2b">eng</language>
        <physicalDescription>
                <form authority="marcform">print</form>
                <extent>238 p. : ill., maps ; 22 cm.</extent></physicalDescription>
        <note type="statement of responsibility">Nadine Slavinski.</note>
        <note>Includes index.</note>
        <subject authority="lcsh">
                <topic>Bicycle touring</topic>
                <geographic>Germany</geographic>
                <topic>Guidebooks</topic></subject>
        <classification authority="lcc">GV1046.G3 G47 1994</classification>
        <classification authority="ddc" edition="20">796.6/4/0943</classification>
        <relatedItem type="series">
                <titleInfo><title>By bike</title></titleInfo></relatedItem>
        <identifier type="isbn">0898863872 (acid-free, recycled paper) :</identifier>
        <recordInfo>
                <recordContentSource>DLC</recordContentSource>
                <recordCreationDate encoding="marc">931129</recordCreationDate>
                <recordChangeDate encoding="iso8601">19990429094819.1</recordChangeDate>
                <recordIdentifier>93047676</recordIdentifier></recordInfo>
</mods>
```

the format of metadata it exposes for OAI harvesting, in addition to MARCXML and MODS. The example record used above, processed through the MARCXML to DC transformation looks like this:

```
<rdf:Description xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
        <dc:title>Germany by bike : 20 tours geared for discovery /</dc:title>
        <dc:creator>Slavinski, Nadine, 1968-</dc:creator>
        <dc:type>text</dc:type>
        <dc:publisher>Seattle, Wash. : Mountaineers,</dc:publisher>
        <dc:date>c1994.</dc:date>
        <dc:language>eng</dc:language>
        <dc:subject>Bicycle touring</dc:subject>
</rdf:Description>
```

ONIX to MARCXML transformation

Other transformations. There are other transformations that LC is experimenting with and has made (or plans to make) available in the MARC 21 architecture suite for others to use in their own projects. These include an ONIX to MARCXML transformation. LC uses this transformation to pick certain data such as descriptions and tables of contents out of ONIX records to augment LC's bibliographic records. Data can easily be moved from ONIX to MARC21 (2709) via MARCXML and merged into the records in the LC integrated library system. Other tools posted or under consideration include:

— tagging transformations: oai_marc to MARCXML (posted), name instead of number tags for MARCXML, different language tags for MODS

— character set transformations: MARC8 to and from Unicode (posted), precomposed to decomposed characters

— MARCXML to FRBR display tool

— MARCXML record validation tool (draft posted)

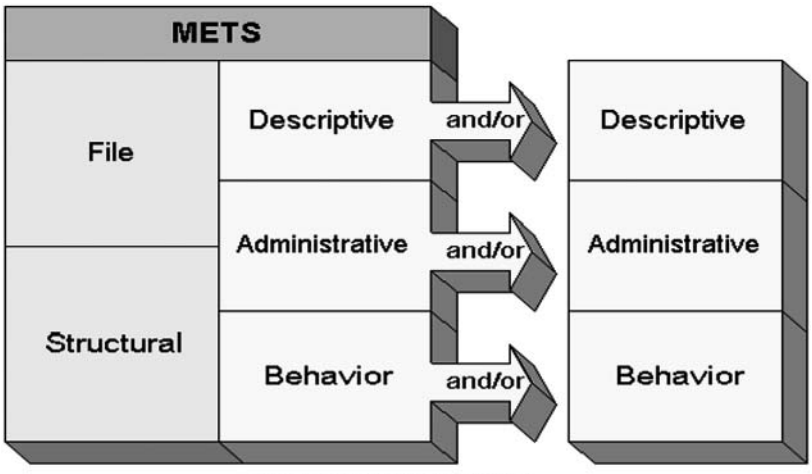

Figure 2: METS architecture

LC sees these services provided from the MARC 21 maintenance agency as being valuable to the community of users to help maintain the savings and interoperability built up through use of a common format. The various schema and transformations will help to standardize MARC across this community for XML communication and manipulation, open MARC 21 to XML programming tools and presentation style sheets, standardize MARC 21 for OAI harvesting, and standardize transformations to and from other standard formats such as DC and ONIX. They provide a basis for evolution while maintaining standardization.

Broader metadata needs

More broadly, descriptions of digitized items require technical and rights metadata not appropriate for MARC. LC is focusing on use of the emerging standard METS (Metadata Encoding and Transmission Standard) for packaging descriptive, administrative, and structural metadata into one XML document for interactions with digital repositories. METS data enables resource retrieval and manipulation, object validation, preservation, rights management, etc. when the requisite metadata is present in the METS document. The METS schema is actually a framework for combining several internal metadata structures with external schemas (such as MODS or MIX). METS is non-proprietary; it is being developed by the library community from experience gained from recent digital projects. It is (relatively) simple, extensible, and highly modular. The METS website gives a good overview of the concepts behind the schema and illustrates its use (see www.loc.gov/mets).

At LC METS is used for a big moving image project and for other mixed media digital collections. To support these projects a record derivation and creation utility was developed that LC may offer open source to others, if it proves useful. Some other projects using METS include Bibliotheque Nationale de France's web archiving and digital preservation projects, OCLC's web archiving project, and a large digital project at Harvard University involving audio, in addition to initiatives based at the National Library of Wales, Michigan State University, and the University of California at Berkeley.

In summary

LC has always focused on use of standards that provide maximum interoperability. Thus the following indicates LC's commitments and directions.

— LC uses AACR, MARC 21, and EAD for *primary* access to its collections, that is the foundation from which the institution builds.

— For LC, as for the library community in general, new development is evolutionary.
— LC is employing XML through a MARCXML architecture and tool kit.
— LC is finding the MARC derivative, MODS, an efficient format for describing a part of the large volume of electronic material LC is collecting.
— For broader metadata, LC is working with METS and appropriate extension schema.

For more information on several of the standards described above, see the following web sites
— www.loc.gov/marc
— www.loc.gov/marcxml
— www.loc.gov/mods
— www.loc.gov/mets

Die Verfasserin

Sally H. McCallum koordiniert die Aktivitäten der Library of Congress im Bereich Format- und Strukturentwicklung von Metadaten und für digitale Ressourcen. Sie koordiniert außerdem die interne Anwendung technischer Standards.
Chief Network Development and MARC Standards Office, Library of Congress, Washington, DC, 20540, USA
smcc@loc.gov

Appendix of acronyms and abbreviations

AACR – Anglo American Cataloging Rules
DC – Dublin Core
DTD – Document Type Definition
EAD – Encoded Archival Description
MARC 21 – MARC formats for bibliographic, authority, holdings, classification, and community information data
MARCXML – MARC 21 XML schema
METS – Metadata Encoding and Transmission Standard
MIX – NISO Metadata for Images in XML Schema
MODS – Metadata Object Description Schema
OAI – Open Archive Initiative
OCLC – Online Computer Library Center
ONIX – ONline Information eXchange
OpenURL – URL Framework for Context-Sensitive Services
PPOC – Prints and Photographs Online Catalog
RLG – Research Libraries Group
SGML – Standard Generalized Markup Language
SONIC – Sound Online Inventory and Catalog
XML – eXtensible Markup Language
Z39.50 – ISO 23950 Information Retrieval Protocol