The Advantages of XML and Linked Data in Cataloging for Small Libraries

by Catherine Sebastian

This white paper is intended for local/small library catalogers and librarians and will familiarize them with the basic concepts of XML and linked data for use in their institutions. It will discuss the benefits and practical realities of using MARC versus XML/linked data for encoding bibliographic information, the value of linked data for authority control and information sharing, and some potential problems or liabilities of moving to a linked data model.

Some Benefits of XML and Linked Data

- Bibliographic data that is immediately understandable by humans and simultaneously readable by computers
- Improved searching capabilities
- Increased interlibrary resource access
- Decreased authority control issues and increased collocation of related resources

The Problem with MARC

One of the great leaps forward in library cataloging was the creation of MARC (Machine Readable Code) by Henriette Avram in the 1960s. It allowed for the standardization and sharing of cataloging records like never before, and it is still the dominant encoding container for libraries worldwide. Despite its revolutionary effect on library catalogs, MARC has increasingly
fallen short in this era of pervasive internet access and the shift toward online access to information.

MARC was never intended for an online environment, and there are numerous issues with continuing to use MARC in the future. MARC does not allow for full implementation of RDA, and MARC does not make it easy to put bibliographic data online. Users and patrons now expect to be able to access catalogs and even many resources online, and MARC is not the best choice for this. MARC also does not have the ability to reflect relationships between multiple resources; we must use search queries or other tools outside the scope of MARC to bring together related materials.

**What is XML?**

XML is an acronym for eXtensible Markup Language, and it is a programming language similar to HTML, the language that is used to create web pages. HTML uses tags as its “code” to create web documents and to give the computer instructions about how to present the text and other content of these web documents. XML also uses tags, but it was specifically designed to store and transport data and also to be readable by both computers and humans. So whereas HTML tags do not make any sense to a person who is trying to read it, XML is able to use natural language tags so that data makes sense to the person reading it even without any need for a computer to read and display it. See Figure 1 for a comparison of HTML tags, XML tags, and MARC fields.

XML presents a valuable option for improving upon MARC coding. MARC codes do not have any inherent meaning; we rely on RDA, AARC, or other cataloging rules to assign meaning to MARC fields and subfields. So when librarians try to view a MARC record in its
“raw” format, they must know what the fields, indicators, and subfields mean in order to understand the information without any translation.

### Figure 1: Comparison of HTML, XML, and MARC “codes”

<table>
<thead>
<tr>
<th>HTML tags</th>
<th>XML tags</th>
<th>MARC fields</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>&lt;p&gt;</code></td>
<td><code>&lt;WorkTitle&gt;</code></td>
<td>247 00 $v`</td>
</tr>
<tr>
<td><code>&lt;li&gt;</code></td>
<td><code>&lt;Contributor&gt;</code></td>
<td>340 #7 $2</td>
</tr>
<tr>
<td><code>&lt;img&gt;</code></td>
<td><code>&lt;ShelfMark&gt;</code></td>
<td>765 01 $b`</td>
</tr>
<tr>
<td><code>&lt;a&gt;</code></td>
<td><code>&lt;GenreForm&gt;</code></td>
<td>010 <code>##</code></td>
</tr>
</tbody>
</table>

**What is linked data and the semantic web?**

The semantic web uses the general structure of the World Wide Web in order to share data. The web sites and web pages that we use daily are sophisticated shared documents; the semantic web uses the same model to share data, and to do so in a way that makes the data easily understood both by humans and by computers. Most of the electronic data that we currently use is either readable by humans or readable by machines. Making data understandable to both humans and computers simultaneously will allow for widespread, easy sharing of bibliographic data.

Linked data establishes connections between many different types of data (or “records” in the case of library cataloging). The catalog record for one fiction item could have links to authority records for the author, title, characters, subject headings, format, language, and many others. It could also link directly to other resources related to the item. This would essentially
turn the web into one giant catalog, allowing users to make connections between related resources that might not have been possible in a small library’s catalog. Linked data uses URIs, Uniform Resource Identifiers, to connect data. The creation of a semantic web through linked data utilizing URIs could result in more powerful searching capabilities and more extensive and precise query returns.

Linked data can be used to resolve authority control issues; rather than searching for an authority record and copying it into original cataloging, catalogers can link directly to that authority record’s URI. This has far-reaching benefits for both eliminating errors and for greatly increasing efficiency of search queries. Inter-institutional resource sharing can be dramatically improved with the use of linked data. Additionally this model of linked data more accurately reflects the relationship models for cataloging defined by FRBR and FRAD.

Catalogs today rely on relational databases to enable searching. A query searches through pieces of comparable data from numerous records that are stored in tables along with record identifiers. A search returns the record identifiers for any data that meets the requirements of the query. This is why “author” or “title” searches are usually quicker and more reliable than keyword searches; they search only the specific tables that contain author or title data. Keyword searches will search multiple tables which takes more time and will usually return some records that are not useful for your needs, even though they did meet the requirements of the query.

XML and linked data still allows for individual libraries to tailor the look of their catalog to their specific needs, but has the potential to reduce some of the issues of conflicting software packages. Cascading Style Sheets (CSS) are used to make the content of HTML documents look a certain way. Different style sheets can be used to make the exact same HTML document
display very differently on the screen. Extensible Stylesheet Language (XSL) can be used in the same way to style and display XML data. XSL Transformations (XSLT) can be used to transform XML documents into other computer languages, making XML a very versatile language for storing and using data. Javascript and other scripting languages are also used to add dynamism to web pages, and these tools can also be used with XML data to help tailor your library’s catalog to be exactly what you want.

**Concerns and Liabilities**

Currently linked data is being explored, researched, and tested in prototype cases in some libraries. A full implementation of linked data in order to create a semantic web of bibliographic data is still years away. Some linked data sets can be seen and examined through both the Library of Congress and OCLC in order for current librarians and catalogers to familiarize themselves with linked data. Some specific challenges and liabilities include:

- Continuing education in XML and online computing concepts will be needed for library professionals who are not already proficient.

- Transitioning MARC21 data to a linked data framework will inevitably result in some loss of data and other challenges.

- One of the largest concerns is the robustness of our internet infrastructure. Loss of connectivity could mean loss of complete catalog function until online service is restored.

- Hacking and cyberattacks could also present a serious concern. A larger system reliant on numerous connections would be more vulnerable to this sort of threat.
Summary

MARC is becoming outdated and will need to be either fundamentally overhauled or replaced. XML offers an attractive option to catalogers for encoding bibliographic data in a form that is simultaneously readable by people and by computers, and is expected to reduce libraries’ cataloging costs. Using linked data to create a semantic web of bibliographic data offers new opportunities for information sharing, fruitful and efficient searching, and effective authority control. Although linked data is not ready to be fully implemented at this point, XML is already being widely used in conjunction with MARC and linked data projects, such as the Library of Congress’s BIBFRAME, are being actively pursued.
Bibliography


