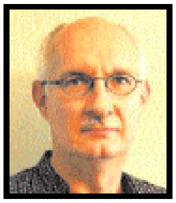
LIBRARY JOURNAL netconnect

XML: Libraries' Strategic Opportunity

Dick R. Miller explains how this innovative code can help unlock library-based data from ILS and MARC



XML (eXtensible Markup Language) is fast gaining favor as the universal format for data and document exchange -- in effect becoming the lingua franca of the Information Age. Currently, "library information" is at a particular disadvantage on the rapidly evolving World Wide Web. Why? Despite libraries'explorations of web catalogs, scanning projects, digital data repositories, and creation of web pages galore, there remains a digital divide. The core of libraries' data troves are stored in proprietary formats of integrated library systems (ILS) and in the complex and arcane MARC formats -- both restricted chiefly to the province of technical services and systems librarians. Even they are hard-pressed to extract and integrate this wealth of data with resources from outside this rarefied environment. Segregation of library information underlies many difficulties: producing standard bibliographic citations from MARC data, automatically creating new materials lists (including new web resources) on a particular topic, exchanging data with our vendors, and even migrating from one ILS to another.

Why do we continue to hobble our potential by embracing these self-imposed limitations? Most ILSs began in libraries, which soon recognized the pitfalls of do-it-yourself solutions. Thus, we wisely anticipated the necessity for standards. However, with the advent of the web, we soon found "our" collections and a flood of new resources appearing in digital format on opposite sides of the divide. If we do not act quickly to integrate library resources with mainstream web resources, we are in grave danger of becoming marginalized.

XML's huge upside

With its exceptional flexibility, generality, and convergence of functionality, XML offers an unprecedented opportunity. Business interests recognize that users prefer to search a single resource and are working round the clock to prepare enticing information portals complete with their "brands" of information. Libraries, however, have the unique advantage of well-known and long-held values of impartiality, trust, confidentiality, thoroughness, and lack of commercial interest and should not make the mistake of selling our good names as some professional societies have tried with unfortunate consequences. Libraries can add the unifying technical infrastructure of XML to their arsenal far more easily than the business world can convince its customers that they too share our values and high standards.

An open Internet standard

XML is becoming the de facto Internet standard for representation of information content (not format) -- optimized for web delivery. Technically, it is a recommendation approved in February 1998, by the World Wide Web Consortium (W3C). It is a metalanguage for defining an unlimited number of specific markup languages, each of which may contain an unlimited number of tags, hence extensible. It is a subset of SGML (Standard Generalized Markup Language), ratified by the International Organization for Standardization (ISO) in 1986 and used widely in Europe and in the publishing industry to assist in the electronic delivery and publication of text-based documents.

By 1996 it was clear that SGML was too complex to be handled on the fly by web applications, and similarly HTML (Hypertext Markup Language), an application of SGML, was too limited in handling digital presentation. In fact, vendors of popular browsers were adding nonstandard tags to HTML for commercial advantage, causing incompatibilities in the display of documents on the web. To address such problems, XML streamlined SGML, adopted HTML’s popular syntax, added web efficiencies, and influenced the introduction in February of XHTML (eXtensible HTML), which is poised to supplant HTML.

Simplicity is elegance

Documents/records encoded to conform with XML have both a logical and physical structure. Logically, they consist of a hierarchy of named elements, which may be likened to fields, with nested elements akin to subfields. Each instance of a document has a single root element to which other elements are subordinate. Container elements simply contain text and/or other elements. This may be thought of as an inverted tree with one root, many branches, and leaves representing actual data values. Elements must be delimited by matched pairs of angle brackets (start/end tags) a la HTML, with the important difference that end tags are required. A document is said to be well formed when its elements, marked by their start and end tags, are nested properly within one another. Unlike HTML, XML must be well formed (Figure 1).

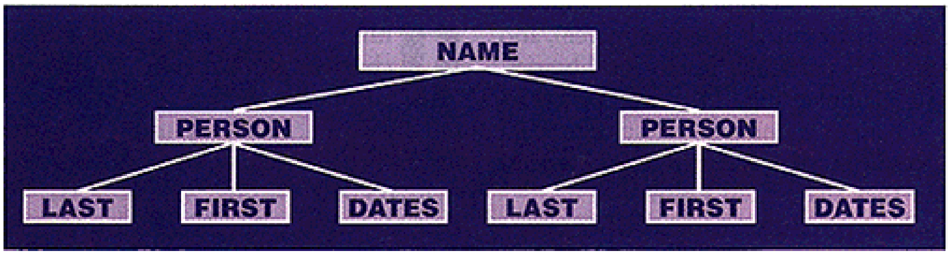


FIGURE 1 **XML’S HIERARCHICAL STRUCTURE:** XML's elements are most easily characterized as an inverted tree with actual data values occupying the leaves (at the bottom).

To provide information about an element’s properties, named attributes can be embedded in its start tag. For example, a unique identifier (id) could be referenced within a document by another attribute, identifier reference (idref) (Figure 2).

Physically, entities allow components of a document to be named and stored separately, permitting information reuse and non-XML data referencing, e.g., images. Usually, entities are declared at the top of a document and then referenced within the document. Other XML features are more esoteric and beyond the scope of this article.

A free-for-all?

Does all this flexibility portend chaos? Not necessarily. Groups with similar interests can develop a suite of DTDs (document type definition) to accommodate their shared needs, with local extensions readily providing for unique needs. A DTD declares each of the permitted entities, elements, and attributes and the relationships among them, basically forming a template for the logical structure of associated XML documents. It expresses the hierarchy and granularity of data, allowable attribute values, and whether elements are optional, repeatable, etc. When an XML document conforms to a DTD, it is said to be valid, although it can be well formed without being valid.

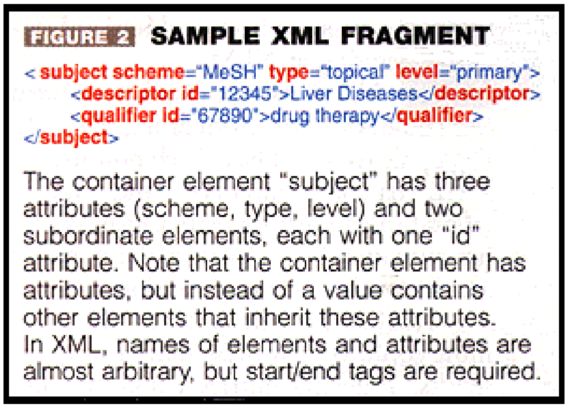


FIGURE 2 SAMPLE XML FRAGMENT

```
< subject scheme="MeSH" type="topical" level="primary">
      <descriptor id="12345">Liver Diseases</descriptor>
      <qualifier id="67890">drug therapy</qualifier>
</subject>
```

The container element "subject" has three attributes (scheme, type, level) and two subordinate elements, each with one "id" attribute. Note that the container element has attributes, but instead of a value contains other elements that inherit these attributes. In XML, names of elements and attributes are almost arbitrary, but start/end tags are required.

A DTD is not required, i.e., browsers can read properly tagged XML documents without one, but ideally should be a companion to XML documents. XML editors can be configured to enforce adherence to a DTD.

Although analyzing data and establishing logical relationships requires considerable intellectual effort, we can be encouraged by the example of other groups, which have overcome inertia and differences to reach consensus. DTDs have been used to define a Biosequence Markup Language, as well as ones for astronomy, chemistry, and mathematics. A Music Markup Language supports inclusion of sound and the display of text encoding as sheet music!

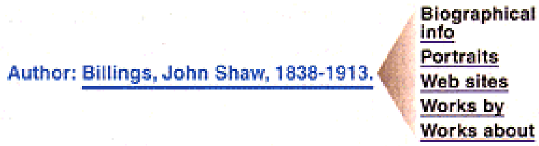
Potentially more powerful than DTDs, schemas use XML's syntax and permit complex data types (e.g., integer, decimal, time, language) to validate XML documents more effectively. There are several proposed schema languages under consideration by the W3C, which plans to issue its proposal this year.

Divide and conquer

XML is actually at the core of a family of specifications, each optimized to deal with a discrete aspect of document management on the web. These adjunct standards are in various stages of development. For example, XSL (eXtensible Stylesheet Language) is similar to Cascading Style Sheets and separates display instructions from the content designation covered by XML. The separation of presentation from content is one of the most powerful aspects of XML. The same data can be displayed in as many different formats as style sheets are defined for various purposes.

Other specifications include XLink (XML Linking Language) to accommodate hypertext linking between documents. It goes beyond simple hotlinking by permitting a single link to reference multiple related documents. An XML namespace is a collection of names that are used in XML documents as element types and attribute names. Within the defined context, the names are thus guaranteed to be unique. Since a DTD defines a single namespace, this comes into play when elements from different DTDs are needed in one document (Figure 3).

FIGURE 3 INTELLIGENT HYPERLINKING



This hypothetical fragment of a catalog or webliography entry shows how clicking on a single hotlink can yield multiple options for users seeking information about this famous physician/librarian. XLink permits embedded resource locators to create functionality in documents, such as conditional links to scope notes and cross-references in authority records, as well as triggered database searches.

A foundation for efficiency

XML offers the power, precision, control, and flexibility that should appeal to librarians at the gut level. It provides a unifying foundation for future development due to its extensibility (suitable for evolving standards), sophisticated hyperlinking, modularity (permitting reuse of information in multiple contexts with different requirements), and relative simplicity. It is often characterized as self-describing in that tagging is intended to be readable by nonspecialists.

In addition, XML is computer platform and software application neutral. This makes it ideal for handling everything from highly structured database records to semistructured documents in order to facilitate their exchange. While it has been used mostly in web publishing, XML's support for interoperability underpins its wide acceptance in computing, business (particularly in e-commerce), and science. Of particular interest to libraries, XML is supplanting EDI (Electronic Data Interchange) standards. Disparate data sources are easier to integrate and process when they share XML's syntax.

New products with XML support are appearing almost daily. In May, Oracle released the long-awaited Internet File System (iFS), which can automatically build a relational database from an XML DTD and easily output XML documents. Also in May, Intel announced new products designed to speed up processing of XML documents at the hardware level. Sun is adding an extension to Java to accommodate XML, which incorporates SAX (Simple API for XML), a popular Java-XML interface. Microsoft's Internet Explorer 5 was the first to support XML display, and Netscape has announced that it wants to turn Navigator into an XML platform. There is plenty of hype, but there is also persuasive momentum.

Much open-source software is freely available, including SAX and the developing DOM (Document Object Model), a standardized interface to XML data from W3C, which will allow programs and scripts to access and update dynamically the content, structure, and style of documents. The document can be further processed, and the results of that processing can be incorporated back into the presented web page. All most of us need to know is that there will be a wide array of products to create, manipulate, and leverage XML data easily. A related driving force is economics: some estimate cost savings in data distribution of 30 to 60 percent by using XML. Examples below illustrate current trends and the potential benefits of wide application of XML in libraries.

Electronic content preservation

Unlike SGML, XML has a fixed character set, Unicode. Unicode incorporates venerable ASCII's 256 one-byte values and, by using two bytes for each character, expands this to more than 65,000 possible values. XML has also adopted the ISO 10646 character encoding format that uses up to four bytes, providing for over two billion unambiguous possibilities. These universal standards allow XML to handle diacritics, special characters, and non-Roman data just like ordinary text, both within documents and in computer operating systems and applications. This is of critical importance

Behind XHTML

is a new standard that brings HTML 4 into conformance with XML (i.e., readily viewed, edited, and validated with standard XML tools), inheriting its benefits:

- Extensibility. Relatively easy to introduce new elements and attributes. HTML's fixed tag set was not the solution for controlling display.
- Flexibility. Regains power of SGML without most of its

to libraries as well as to the internationalization of data networking.

Due to Unicode support and platform neutrality, XML offers the greatest promise of data longevity (or future-proofing) as hardware, software, and network protocols continue to change. Lane Medical Library at Stanford is using XML to convert e-mail and possibly other electronic documents in obsolete formats where the programs that created them are no longer available. The deteriorating, magnetic tape archives contain e-mail, project, and computer program documents of Nobelist Joshua Lederberg, Edward Feigenbaum, and others who developed the "expert systems" branch of artificial intelligence from the 1970s. Analysis of the e-mail will allow mapping it into elements, such as date, to, from, re, message, etc. Personal names can be inverted, and threads to and from related messages may be preserved with XML linking features. Device-independent, XML-based e-mail systems that are in the works will eliminate the need for such conversions.

SGML without most of its complexity.
- Interoperability. Eliminates present incompatibilities in extensions to HTML for various browsers.
- Precision. Requires end tags, proper nesting (no overlapping elements), lower-case element and attribute names, attributes always quoted.
- Dynamic. Supports applications that rely on the Document Object Model (DOM).
- Modularity. Can be configured to support personal digital assistants, cellular phones, etc.
- Compatibility. Technical features largely backwards compatible with HTML.

Bibliographic databases

Database management is a complex topic. XML provides for unambiguous identification of complex data structures that can be treated as objects. Namespaces can be used to unite parts of DTDs or schemas to help manage greater complexity. And database interface products supporting XML, such as Oracle 8i (relational) and Ozone (object-oriented), are available.

Behind XML MARC

- Converts MARC records to XML documents.
- Includes sample DTDs for bibliographic records and authorities.
- Includes flexible maps (also in XML) to permit alternative mappings, localization.
- Free for noncommercial use.
- Released: December 29, 1999; announced: February 15, 2000.
- Over 300 licensees from over 40 countries.
- Produced by Lane Medical Library, Stanford University Medical Center.
- Available at: http://xmlmarc.stanford.edu

The best argument for the feasibility of using XML in conjunction with database management is the example of the National Library of Medicine (NLM). As part of a project to modernize its computer systems, NLM chose XML as the format for disseminating MEDLINE bibliographic citation data and will spend much of this year converting more than 11 million records. XML serves as the input/output mechanism to a commercial relational database product.

NLM took the opportunity of switching formats to make organizational changes in data, such as separating errata and retraction information from titles and providing for new elements, e.g., corporate author. XML will be the only distribution format for MEDLINE beginning in 2001. NLM plans to produce an XML version of MeSH (Medical Subject Headings) and eventually of its MARC cataloging. XML is also used internally for the forthcoming NLM Gateway, an intelligent search tool that can query the multiple back-end retrieval systems operating at NLM.

The NASA Astrophysics Data System also chose XML for reformatting all its bibliographic records. DialogWeb now uses an XML database interface, and WIPO (World Intellectual Property Organization) has announced that XML is the preferred format for document submission.

XML offers the potential for even more sophisticated presentation of query results. Indexing entries could be converted on the fly to construct a search result in XML that has structure and functionality, e.g., an author browse could display surnames, with an option to "open" selected ones to see an

Links

GENERAL
LC MARC SGML
http://lcweb.loc.gov/

alphabetical subarrangement of forenames with multiple hotlinking and hit counts. XMLMARC?

Despite its simple hierarchical structure, XML reveals a remarkable accommodation for complex bibliographic data. Librarians must take a very serious look at MARC and AACR2 in view of the many advantages afforded by XML. Creating DTDs for MARC would be a considerable task, but can we afford to do nothing with a format designed for card production in the 1960s? It appears possible to incorporate the best of MARC into a modern format with correlated "cataloging" rules for the 21st century.

Beginning in September 1998, Lane Medical Library undertook the Medlane Project, which involved converting catalog records to XML for integration with other web resources. Lane developed sample DTDs to explore restructuring and simplifying MARC and released XMLMARC software to demonstrate conversion feasibility. Currently, the project focuses on indexing and interface development.

In January, a French government agency released BiblioML, which converts Unimarc to XML. The Library of Congress produced a literal mapping, MARC SGML, from 1995 to 1998. Logos Research Systems'MARC to XML to MARC Converter and other mappings are also literal.

An elegant solution

A fully XML-based integrated library system is feasible within three to five years. ILS vendors, notably Endeavor, are beginning to incorporate XML into existing systems. The American Library Association could use the model of W3C to expedite development of the required standards, specifically creation of DTDs or schemas for records underpinning ILS modules: bibliographic, authorities, holdings, users, vendors, transactions, interlibrary loans, check-in, etc. Financial DTDs already exist. Although a daunting task, businesses are embarking on such development in other areas. Why not for library systems?

XML ILS?

XML affords an elegant solution to what at first appear to be complex and unmanageable problems. If ILSs and MARC were XML-based, we would be unshackled and free to concentrate on enhancement of functionality and cross-system integration, far surpassing today's systems. This single change to an open, universal format, which includes a role for library system vendors, can transform the time-honored MARC format to prevent its obsolescence and put librarians in the mainstream -- better positioned to serve our users.

marc/marcsgml.html

MARC and SGML/XML
http://www.oasis-open.org/
cover/marc.html

NLM MEDLINE XML
http://www.nlm.nih.gov/
bsd/xml_announce.html

XHTML.org
http://www.xhtml.org

XML.commune
http://www.xml.com

XML Cover Pages
http://www.oasis-open.org/
cover

XML.org
http://www.xml.org

STANDARDS

Unicode
http://www.unicode.org

W3C XHTML
http://www.w3.org/TR/
xhtml1

W3C XML
http://www.w3.org/TR/
REC-xml

PRODUCTS

BiblioML
http://www.culture.fr/
BiblioML

Logos MARC XML
http://www.logos.com/marc

XMLMARC
http://xmlmarc.stanford.edu
Includes library-oriented
webliography.

Dick Miller (dick@stanford.edu) is Systems Librarian and Head of Technical Services, Lane Medical Library, Stanford University Medical Center.

This article originally appeared in the Summer 2000 issue of netConnect.

For more information about netConnect, see http://www.ljdigital.com/netconnect.asp.

TOP