

Annotated Bibliography on Extensible Markup Language (XML)

Virginia Shields

LI 804 Theory of the Organization of Information

Dr. Cathy Perley, Instructor

School of Library and Information Management

Emporia State University

March 2003

Annotated Bibliography on Extensible Markup Language (XML)

The topic of this annotated bibliography is Extensible Markup Language (XML), sometimes written as eXtensible Markup Language. The standards for XML were written by a workgroup from the World Wide Web Consortium (W3C) and published February 10, 1998. Although based on Standardized General Markup Language (SGML), this document markup language is simpler to use, readable by both humans and computers, and can be used in a variety of applications. XML's structure allows it to describe the content of a document, organizing the information so it is more retrievable. There are many facets to learning about XML. These include, but are not limited to, understanding: the structure (markup tags for elements and attributes); the concepts of well-formed and valid; the differences between Document Type Definitions (DTDs) and XML Schema; the various uses of XML, such as MARCXML; the other standards being created to aid XML (sometimes called the XML family of technologies); and the current debate over how to store XML content. The goal of XML is to persistently and accurately retrieve and share content in context.

XML is related to the broader subject area of the organization of knowledge as an organizing tool for content and information. The structure and the language of XML allow both computers and humans to look at an XML document, and put context to the content. Just as we organize our kitchens or bookshelves so that what we want can be retrieved easily, XML is a tool to accomplish this in our document-rich world. XML is already used in content management, workflow and e-commerce.

The audience for this bibliography is the graduate student enrolled in the SLIM program. In article after journal article, it is written that XML is ubiquitous. If this is so, SLIM students will need an understanding of XML, no matter what arena they end up working in. I know that

SLIM students have a range of skills in searching out information, depending on where they are in the program. The books and articles I chose to annotate should be accessible to my audience through databases, library holdings and the internet. My audience will seek out this information to educate themselves and to have a knowledge base on XML. While they may never write in XML, SLIM students know that they will be assisting others in searches. There is material for those interested in libraries, as well as non-library fields. XML promises to make information more retrievable in the future for everyone.

Searching for the documents that were utilized in this annotated bibliography was like going on a treasure hunt. I searched through the catalogs of three library systems for books to first educate myself about the topic. I found another source for books when I used the Google search engine and found two internet bibliographies about XML. One of these bibliographies was Charles F. Goldfarb's 'All the XML Books in Print ...or nearly so.' Goldfarb is known as the 'father of meta languages' because he helped invent SGML, which gives his opinions some authority. Certainly using Goldfarb's recommendations helped me weed out my own choices of what was available to read.

I used the 2001 Library of Congress Subject Headings (LCSH) books to help find the common terms used so that my searches would be more effective. The Subject Headings use XML as the term, with the broad term as Document Markup Language, which is under the broader term of Text Processing (computer science). Finding the LCSH books was a search in itself. At the local library branch, the most current edition they held was 1989, and XML was not covered since it hadn't been invented yet. I then discovered the more current edition at Johnson County Community College's Billington Library. The LCSH certainly are helpful in discovering both relationships between subjects, and terms that might aid a search that you might

not have thought of. If you were totally unfamiliar with a subject, the LCSH would be a very useful place to start in your search.

The next part of my search was very frustrating at times. I used databases available to me through my public library system, through Emporia's Kellogg Library, and through SLIM's access to Dialog. My experience with databases had been limited before joining SLIM, so this was still a learning experience. I experimented with Boolean searching, as well as using truncation. I discovered how to use advanced searches and to look for subject headings in the databases. I found which databases would most likely cover my subject, which would have full text, and which would just point the way to an article.

Discovering that ERIC documents or journal articles were often not available in full text or abstracts, and that sometimes the meta records didn't contain the author's name or the journal it had appeared in, discouraged me initially. However, I tried alternate databases to find them in full text, including InfoTrac Web: General Reference Center Gold, Dialog's Business and Industry, and the Gale Group and Trade. I found most of my journal articles on databases. However, while looking at those articles that I had already found, it was revealed that many of them were from the same sets of magazines or periodicals. Using the Google search engine to locate those magazines' web pages, I was then able to search on their websites for other articles that might pertain to XML and retrieval.

Another source that proved to be useful in my search was the Special Libraries Association (SLA) website at <http://www.sla.org>. As a member, I was able to access electronic information resources, which included a subject heading of metadata. Some of the articles that I chose to use in this annotated bibliography were originally pointed to by the SLA website.

When I started to work on the citations and the critical analysis, there were several problems with the full plain text versions, as well as the electronic versions of the works I had chosen. Often figures, diagrams, and other explanatory tools were missing from the article I had printed. Electronic versions of periodicals rarely included their volume or issue number, or their page numbers. Sometimes I could find this on a database, but not always. Fortunately, I discovered rich depositories of periodicals and magazines at William Allen White Library in Emporia, the Central Resource Library in Johnson County, as well as the JCCC Billington Library. I queried all three online library catalogs to find out their holdings concerning the 14 journals I was interested in. It was then simple to make a plan of attack, to visit the necessary library, and to check out the periodical or microfiche where I needed additional information.

I learned several things about the organization of information in this process. The major point is to never give up, never surrender! While information abounds around us, there are many ways to access the same information, if it is well organized. The key these days often lies in the metadata tags describing the information, which the tool of XML will improve as its use increases. I found in many cases I was able to access the same article through a database, then find it on the WWW, and then seek out its print version in a library. When that occurred, I usually cited the print version, because I feel it is the most stable and reliable. I gained a better understanding of the subject of XML and how information is structured and organized, by being flexible enough to approaching it from many different angles.

This annotated bibliography is designed to help with an understanding of Extensible Markup Language (XML) and its many facets. It is organized to provide first a basic understanding of XML and its structure, and then to delve into the potentials of this tool, especially in terms sharing information and information retrieval.

Bos, B. (2003, January 10). *XML in 10 points* (update). Retrieved March 24, 2003, from

<http://www.w3.org/XML/1999/XML-in-10-points.html>

A snapshot of XML, this introduces the basic concepts with numbered points and brief explanatory paragraphs following. Produced by W3C, this document has authority. With hypertext links on technical jargon, it is useful as an overview or to present the subject to others.

Tittel, E., & Boumphrey, F. (2000). *XML for Dummies (2nd ed)*. Foster City, CA: IDG Books Worldwide, Inc.

This is a detailed guide on learning XML concepts for the purpose of writing in XML. The structural concepts and their relationships covered in chapters 1 through 5 include elements, attributes, well-formed documents, valid documents, and Document Type Definitions (DTDs). Icons, diagrams and an in-depth glossary are useful. Tittel and Boumphrey also add information on XML family of technologies (including XLink and XPath) and on XML languages (MathML and WordDocs). A CD-ROM is included, which has ability to access examples and URLs by chapter; several web browsers that can recognize XML; and some XML editors. This book is written with humor.

Morrison, M. (2001). *HTML and XML for beginners*. Redmond, WA: Microsoft Press, 282-329.

Although the bulk of this book is about Hypertext Markup Language (HTML), the last three chapters concern XML. Clearly written, the opportunities to capture the contents of documents with XML are presented in the first of these chapters, "Understanding XML." The other two chapters are aimed at readers learning how to write XML, and also how to add style to XML through XSL (eXtensible Style Language) and XHTML (eXtensible HTML).

Harrold, E.R., & Means, W.S. (2001). XML in a nutshell: A desktop quick reference.

Sebastopol, CA: O'Reilly & Associates, Inc.

XML fundamentals are covered in depth, with additional sections covering XML's ability to be global; on the web; and as a data format. The XML technologies covered in their own chapters include XSL Transformations (XSLT), XPath, XLinks, and XPointers.

There is a thorough index and the sections are tabbed on sides of pages, making it easy to access the section wanted.

Balas, J.I. (2002, September). What is this XML thing and why do I need to know about it?

Computers in Libraries, 22(8), 39-41.

From a librarian's point of view, this article covers learning XML as a new subject.

Useful tools are mentioned, such as Webopedia, a free online dictionary with technological terms. Other resources, such as website tutorials, are discussed. The XML4Lib electronic discussion group and its archives were mentioned, but the given link didn't work. The corrected link is here: <http://sunsite.berkeley.edu/XML4Lib/>. The archives themselves have three years of listserv information, retrievable by date, subject, or author. There are other resources available to learn about XML and libraries at this site. The author of this site is Roy Tennant.

Extensible Markup Language (XML) 1.0 (1998, February 10). Retrieved February 19, 2003, from <http://www.w3.org/TR/1998/REC-xml-19980210>

The actual specification for XML, this document is very structured. Hyperlinks from the table of contents to the sections themselves within the document are useful. The original ten design goals are presented in section 1.1., and the names of the W3C XML Working Group (Non-Normative) are in Appendix G. The latter is pointed out since many of the XML Working Group members have authored articles about XML and their work carries authority. The rest of the specification may appeal to those who need the nitty-gritty details about this subject.

Harold, E.R., (1999). *XML Bible*. Foster City, CA: IDG Books Worldwide, Inc.

A detailed guide for writing in XML, this is the book to have at hand for that purpose. What is helpful, even for those not ready to write in XML, are the summaries at the end of each chapter. A CD-Rom comes with this book, and is invaluable for seeing examples on the computer screen. Examples include Shakespeare's plays, baseball statistics and players, the Old Testament, and the periodic table, all written in XML.

McDermott, I.E. (2002, February). The third wave of the information age: Internet librarian conference November 2001. *Searcher*, 10(2), 54-60. Retrieved February 24, 2003, from Academic Search Elite database.

In the section XML for Libraries, there is an overview of XML, its structure, and its potential for libraries with presenter Roy Tennant. (Tennant's XML4Lib electronic discussion group was mentioned in a previous annotation.) Benefits of coding in XML include: long-term storage across platforms; easy migration; and enhanced searchability of documents.

Scharf, D. (2002, December). XML under the hood. *Information Outlook*, 6(12), 20-27.

A thorough look at XML and its power, Scharf includes an overview of XML structures, DTD, XML Schema, and history. Scharf considers XML to be the key to resource sharing and calls it a set of rules for organizing information. 'Definition of extensible ' is put in context for XML. Describes some of the ways the Library of Congress has embraced XML, mentioning Encoded Archival Description (EAD) and MARCXML. The Semantic Web, related technologies, and references are presented. A chart comparing English text, HTML and simple XML facilitates understanding.

Davis-Tanous, J.R. (1999, December). XML: A language to manage the World Wide Web.

ERIC Digest, ED437941. Retrieved March 10, 2003, from

http://www.ericfacility.net/databases/ERIC_Digests/ed437941.html

This article presents a simple understanding of XML's potential use on the web, with the differences between HTML and XML pointed out. Discusses the Gateway to Educational Materials (GEM) Project and how XML would add value to it. Davis-Tanous looks at the reasoning behind using DTDs with XML. Some of the problems discussed about XML working on web pages have been resolved since the article's publication. Bibliography at end of article provides 12 references.

Byrne, T. (2002, September). Heeding the call of reusability [Electronic version]. *EContent*, 25(9), 17-23.

Content management systems use XML as a powerful new tool. The content structure concepts that XML impacts are: reusability, addressability, predictability, and granularity. Byrne includes a Structure Thesaurus, example of a chunked press release, and review of some current technologies, especially for the non-technical content owner.

Mace, S., Flohr, U., Dobson, R., & Graham, T. (1998, March). Weaving a better web. *Byte*, 23(3), 58-67.

This is a look at XML as a tool to improve the World Wide Web. XML's relationship with HTML is touched on, and what will drive the acceptance of XML as a standard. If possible, retrieve this article in print, rather than plain full-text from a database because of the many diagrams and sidebars. These included: information on the web's 1998 and (projected) 1999 capabilities; HTML problems; XML's structure; applications, syntax power, and name spaces; new tools for the web; and creating XML objects. One of the difficulties of this article is the age of it. It goes into great detail about Dynamic HTML (DHTML), but currently XHTML and HTML 4.0 are more prevalent. It is nevertheless a valuable look at XML.

Lamont, J. (2001, May 1). Behind the scenes, XML sizzles. *KM World*, 10(5). Retrieved February 24, 2003, from

http://www.kmworld.com/publications/magazine/index.cfm?action=readarticle&article_id=1007&publication_id=1

Lamont looks at the benefits and drawbacks of XML. Concepts covered are interoperability, flexibility, and data exchange. This leads to applications in content management, access to legacy data, and e-commerce. Some of the difficulties in creating standardized schemas for companies are examined. Drawbacks include the file size and the technicality of XML beyond its basic concept. Also discussed are products for XML and standards based on XML. Lamont has short interviews with many corporate leaders in this article, allowing the reader to see real life applications of XML.

ven Eman, J. (2002, October/November). What can you do with XML today? *Bulletin of the American Society for Information Science and Technology*, 29(1). Retrieved February 24, 2003, from http://www.asis.org/Bulletin/Oct-02/ven_eman.html

Ven Eman examines uses for XML, including commerce, research and development, the information industry, and information research. The article looks at how using markup tags as metadata adds to the understanding of information; and there is a concise look at the differences between content, context, and format markup. In-depth history of evolution from Generalized Markup Language (GML) to XML included. XML syntax and schemas are discussed. The underlying theme in this article is how XML can brew the morning coffee, and according to ven Eman, it just might be possible.

Margulius, D. (2002, October 28). XML everywhere. *InfoWorld*, 24(43), 44-45. Retrieved February 24, 2003, from http://www.infoworld.com/article/02/10/25/021028feundatatci_1.html

Use of XML is impacting the field of content management by leading to standards for modeling content. Some of XML s technologies are discussed in context with content management. Margulius also touches on the question of what kind of database to use; and on proprietary products being created for use with XML.

Dyck, T. (2002, January 21). XML standards updated [Electronic version]. *eWeek*, Article 0,3959,31992,00. Retrieved February 24, 2003, from <http://www.eweek.com/article2/0.3959,31992.00.asp>

Dyck looks at how XML family of technologies (specifically XPath and XQuery) will aid in search and retrieval of data from XML documents. Advantages and disadvantages of both standards are discussed, as well as how updates to standards will affect XML

databases. Dyck presents his concerns that the lack of standardization of updates may lead to fragmentation, because corporations will come up with proprietary solutions in the meantime.

Dragan, R.V. (2002, November 5). Structuring XML documents [Electronic version]. *PC Magazine*, 21(19), IP01-IP04.

Advantages of new XML Schema over DTDs: DTDs not written in XML; with XML Schema can use XML tools to design both structure of data and data itself; and XML Schemas save time, add control and power. Disadvantages are that XML Schema is more complex to learn, and that namespaces are necessary. Dragan also describes two XML-based standards for communicating on the web: Web Services Description Language (WSDL); and Simple Object Access Protocol (SOAP). Two figures compare and contrast a DTD and an XML Schema, which is useful for visual learners.

Udell, J. (2002, December 2). Modeling biz docs in XML. *InfoWorld*, 24(48), 18-19. Retrieved January 30, 2003, from Academic Search Elite database.

Udell examines XML Schema in detail, with both pros and cons presented. The Microsoft product currently being created (Office 11) will support XML Schema. There are two concise lists that give the Top 5 Reasons to Fear XML Schema and the Top 5 Reasons to Embrace XML Schema. This article also offers a test center perspective.

Dougherty, D. (1999, June). XML's Achilles heel. *Web Techniques*, 4(6), 88. Retrieved March 26, 2003, from InfoTrac Web: Expanded Academic ASAP.

Dougherty examines the difficulties of putting DTDs theory into practice. The W3C XML working group's thoughts, according to Dougherty, were that industry trade groups or the domain specialists would collaborate to create DTDs. He believes that this isn't

happening currently due to politics and human nature. Other concepts covered include XML as syntax and that semantics are not standardized.

Fichter, D. (2001, July/August). XML and intranets: Fact, fantasy, or both? *Online*, 25(4), 69-71.

Driving forces behind XML's diffusion include arenas of electronic publishing; e-commerce; information delivery; automatic processing after receipt; and acceptance of XML by the major players (Microsoft, IBM, and SAP). Covers number of different XML specifications for news formats; document publishing and ebooks; library-specific initiatives, including MARCXML; and data interchange between applications on intranets.

Banerjee, K. (2002, September). How does XML help libraries? *Computers in Libraries*, 22(8), 30-34.

Describes why librarians are excited about XML, especially with the view moving away from libraries as a centralized repositories of information. Information is given about XML's relationships with the Encoded Archival Description (EAD), MARC, and Digital Libraries, as well as other library initiatives.

Jasco, P. (2002, September). XML and digital librarians. *Computers in Libraries*, 22(8), 46-49.

Retrieved February 16, 2003, from Business & Industry database.

Jasco promotes XML as a vehicle for information retrieval and information exchange on the internet. An understanding is offered that MARCXML will neither displace the MARC communications format nor dumb it down. Jasco also explains some of MARC's previous problems involving the limitations of the Dublin Core. Examples of XML uses are given, including PubMed and HR-XML (human resources). Several open source

(free) resources involving MARC, Dublin Core, and Microsoft's Notepad XML editor are examined.

Radosevich, L. (1997, August 25). Health care uses XML for records. *InfoWorld*, 19(34), 51-52. Retrieved March 10, 2003, from <http://www.infoworld.com/cgi-bin/displayStory.pl?features/970825xml.htm>

Radosevich looks at how XML can be used to create portable electronic medical records, thereby helping the health-care industry. Looks at the Kona Proposal, which encompasses this idea. It is stressed that non-proprietary systems are needed to keep patient medical records out of the control of health-care institutions and insurance companies, who could use the information to discriminate against patients. The idea of patient records on the web has both drawbacks (confidentiality issues) and advantages (ER physicians having access to records). Although this article is specific to health-care business, Radosevich points out how vertical industry groups could cooperate to customize XML for their own industries.

Bosak, J., & Bray, T. (1999, May 6). XML and the second-generation web. *Scientific American*, 280(5), 89-93.

Bosak and Bray, both members of the W3C workgroup who created XML, stress XML's ability to make information self-describing, thereby making information more retrievable. XML is designed for document exchange, with efficiency gains in workflow and more effective searches being the result. Advantage of XML's reliance on Unicode allows it to be used globally, as well as between different computer platforms. Discusses XLink (a standard for XML-based hypertext), which allows links to central databases, rather than just linked pages. A diagram shows how one document marked up with XML

can be retrieved in vastly different formats, such as on cell phone, computer screen, or hand-held display; another diagram shows the power of an XML hyperlink. Side box about new languages, sometimes referred to as XML vocabularies, including MathML, Chemical Markup Language (CML) and Astronomical Instrument ML (AIML).

Ludwig, M. (2003, January). Breaking through the invisible web: Mark Ludwig discusses the University at Buffalo's attempts to move its catalog content to the web's surface. *Library Journal*, 128(1), S8-S10. Retrieved March 24, 2003, from InfoTrac Web: General Reference Center Gold database.

Ludwig describes using XML as a way to have access to library content, which is hidden behind proprietary database interfaces. Concepts covered include putting library content on the web, making them bookmarkable persistent documents. Examines Ludwig's experiment of how search engines on the web retrieved XML records. In the experiment, over two million MARC records (coded in XML) were put on a single website. The experiment proved that many records can be put on a single website, but that most search engines don't have the ability to find them all. Inktomi Enterprise Search Engine is mentioned, which the author found had both XML capability and ability to index the site with over two million records.

Miller, R. (2003, March). Get it together: integrating data with XML. *EContent*, 26(3), 20-24. Retrieved March 24, 2003, from InfoTrac Web: General Reference Center Gold database. XML described as a data integration tool across platforms and databases. Benefits of XML are touted, but its flexibility may also be a negative since collaboration among corporations and industries are needed. The term "middleware" is defined. Reasons are given why native XML databases might be preferred over relational databases. Real

world uses of XML in different industries are looked at, including a US government initiative.

Boeri, R.J. (2002, December). Content-centric XML: Where we've been, where we're going in 2003. *EContent*, 25(12), 68-69. Retrieved January 31, 2003, from Academic Search Elite database.

Content-centric XML stressed over data-centric XML (interchanging information between systems). Four areas of XML standards were set in 2002: XHTML; format independence; semantic web; and multimedia. Discusses increasing XML support for business users, and an increase in XML management tools as products.

Fichter, D., & Cervone, F. (2000, November/December). Documents, data, information retrieval, & XML. *Online*, 24(6), 30-36.

XML's structure of documents can lead to better retrieval of documents and data. Seven XML goals listed; history of XML; XHTML structure compared to HTML; Rich Site Summary (RSS); Data Documentation Initiative (DDI); and libraries and XML is touched on.

Adams, K.C. (2001, March/April). The web as a database. *Online*, 25(2), 27-32.

This is about Information Extraction (IE), with a brief section on XML's role as a step towards solving IE difficulties. Other topics covered, and their relationship with IE, include: the differences between IE and information retrieval (IR); natural language processing (NLP); wrapper induction; content management; and the hidden web. Useful definitions and 14 references.

Luk, R.W.P., Leong, H.V., Dillon, T.S., Chan, A.T.S., Croft, W.B. (2002, April). A survey in indexing and searching XML documents. *Journal of the American Society for Information Science and Technology*, 53(6), 415-437.

Scientific research paper covers structured indexing, searching, and information retrieval of XML documents. Reviews four major classes of database structures; examines advantages and disadvantages of an information retrieval (IR) engine; and scrutinizes ways to index (full-text, position-based or multidimensional), and ways to search (full-text, XML assisted, or multistage) in terms of XML. Includes models and over 125 references.

Schlieder, T., & Meuss, H. (2002, April). Querying and ranking XML documents. *Journal of the American Society for Information Science and Technology*, 53(6), 489-503.

Scientific research paper discusses taking advantage of XML document structure to improve retrieval precision. Classical models and proofs, computer algorithms, and 28 references are included. Although the majority of the paper includes very technical formulas for the uninitiated, the text is clearly written and provides a look at the computer science part of making information retrievable.

Leon, M. (2001, April 2). Where s the XML? *InfoWorld*, 23(14), 36. Retrieved March 10, 2003, from <http://archive.infoworld.com/articles/hn/xml/01/04/02/010402hnxml.xml>

Examination of three types of databases (relational, object, and native-XML) to store XML content in. A look at the main vendors and their direction; uses of XML by businesses. Graph of how organizations plan to store XML data.

Carr, D.F. (2001, July 15). XML-native databases. *Internet World*, 7(14), 54-55. Retrieved March 10, 2003, from MasterFILE Premier database.

Reasons behind push for XML-native databases involve XML's structure. Gives arguments against object-oriented databases; definition of 'native'; and examination of relational databases and XML.

Dougherty, M.S. (2003, February 1). Are XML databases necessary? XML databases may help development efforts in some situations but an RDBMS may still be the best choice. *Intelligent Enterprise*, 6(3), S16-S18. Retrieved March 24, 2003, from InfoTrac Web: General Reference Center Gold database.

Choices of types of databases for storage and retrieval of XML. Includes look at relational databases (RDBMS) and native XML databases (NXDB). Two main architectures of NXDBs are text-based and model-based. Data- and Document-centric XML documents are defined.