Running Head: XML: Web 2.0 Fad or the Open Source Solution?

XML: Web 2.0 Fad or the Open Source Solution to Interoperability?

Amy E. Neeser

San Jose State University

The first markup language, GML (Generalized Markup Language), was created in 1969 by a group of IBM researchers. Its original purpose was for document publishing, text editing and formatting, and allowed basic information retrieval systems to share documents (Kay, 2005, p. 30). As these technologies progressed, GML expanded in 1986 and became known as SGML (Standard Generalized Markup Language) which was created to, "provide a set of rules that describe the structure of an electronic document so that it may be interchanged across various computer platforms" (Chowdhury, 2004, p. 323). In addition, SGML allowed users to add editorial comments to files, create different versions of a document in a single file, identify where to place various types of illustrations and how to incorporate them into text files, and provide basic information to supporting programs.

Despite these important advancements, markup languages were nevertheless highly complex to the average user and still mostly unknown until Tim Berners-Lee, inventor of the World Wide Web, created HTML (Hypertext Markup Language) in 1990 (Kay, 2005). With the rise of the Internet, data needed to be displayed in a Web browser so HTML was created to incorporate information that dealt with presentation; amongst other shortcomings, this is one of the major reasons why HTML is criticized as a much too limited markup language. HTML uses fixed tags that are completely unrelated to the actual data of the resource; instead, these tags simply describe how that data should be displayed (Fichter & Cervone, 2000). Because of the great need for an advanced generalized markup language that focused on the data itself, XML (Extensible Markup Language) was finally created and is being used in a wide variety of settings today.

XML is a combination of SGML and HTML; it is less complex and resource-intensive than SGML yet surpasses the ability of HTML in that it does much more than tells a browser

how to display data and link to other items. "XML is intended for computers to generate data, read data, and ensure that the data structure is unambiguous" (Chowdhury, 2004, p. 325). XML is not directly tied to any particular program or application; instead, it simply describes and structures the data so it may be interpreted by whatever program happens to be using it. XML is fully interchangeable and customizable and can be adapted for the particular needs and terminology of individual fields (Saunders, 1998, p. 45). "This means that the same application could display information on a Web browser, hand-held computing device, or cell phone simply by using a different style for each device type" (Fichter & Cervone, 2000, p. 32).

XML's self-describing tags provide a highly detailed representation of documents and data, which inherently ties this markup language to information retrieval. "An XML database enables this information to be indexed for powerful, detailed search … It also supports multi-criteria sorting and delivers multiple options for ordering results" (Rogers, 2004, p. 19). Because of these descriptive names and labels that are assigned through tagging, information can be accessed and retrieved by a number of different systems and a multitude of applications, making this an optimal tool to facilitate information retrieval. By breaking down traditional silos which were barriers to information sharing, XML enables information to be reused by, "… integrating text and data from different sources and by searching and linking across these sources…" (Adler, Cochrane, Morar, & Spector, 2006, p. 210).

There are many factors, applications, and characteristics about the markup language itself that have made XML, "…the predominant mechanism for electronic data interchange between information systems…" (Adler, Cochrane, Morar, & Spector, 2006, p. 207). Some of these that will be further explored in this paper include the advantage of being an open source program, using XML to solve information searching and retrieving dilemmas on the Web, and the

examination of XML both inside and outside of the library setting. Many information specialists are in favor of the open source extensible markup language XML as its purpose is to aid information systems in sharing structured data; however, this is still a fairly new technology and there are also criticisms of it in addition to the large amount of excitement and praise. The intent of this paper is to examine the ever-increasing role that XML as an open source entity plays in the field of library and information science, specifically in regards to information retrieval. This paper will mostly focus on the use of XML in libraries; however, due to the multi-system accessibility of XML, I will also explore a wide range of studies and criticism that fall outside of the library setting as well.

Definitions

Throughout the course of this paper, I will be returning to a few specific terms that need to be defined in order to understand their relationship to each other and to the broader field of library and information science. First and foremost, the term *information retrieval*, "…came to mean retrieval of bibliographic information from stored document databases" (Chowdhury, 2004, p. 1). Furthermore, the function of an information retrieval system is "to retrieve the information – either the actual information or the documents containing the information – that fully or partially match the user's query" (Chowdhury, 2004, p. 2). Information retrieval systems are diverse and can range from digital libraries, to OPACs (Online Public Access Catalogs), to online databases, to various types of web search engines.

Markup languages, on the other hand, are defined as, "A scheme that allows the tagging and describing of individual structural elements of text for the purpose of digital storage,

appropriate layout display, and retrieval of individual components" (Taylor & Joudrey , 2009, p. 463). *XML*, a specific type of markup language, can be distinguished as, "A subset of SGML, designed specifically for Web documents, that omits some features of SGML and include a few additional features (e.g., a method for reading non-ASCII text); it allows designers to create their own customized tags, thus overcoming many of the limitations of HTML" (Taylor & Joudrey , 2009, p. 478).

Finally, *open source software* (OSS) can be defined as "… software that is … free to download, free to use, and free to view or modify. Most OSS is distributed on the Web and you don't need to sign a license agreement to use it." (Schneider, Free for All. Week 3). According to the Open Software Initiative, OSS must also comply with the following criteria in order to fall into this category: free redistribution, inclusion of source code, must allow the creation of derived works from the original product, integrity of the author's source code, no discrimination of persons, groups, or fields of endeavor, distribution of license, the license must not be specific to a product, the license must not restrict other software, and the license must be technology-neutral (http://opensource.org/docs/osd). Although these definitions are good starting points for beginning to understand these important terms, their relationship to each other and their broader implications will become much clearer throughout the remainder of this paper.

The Open Source Advantage

As the need for a more useable yet simultaneously more advanced markup language became more apparent, it was also becoming more evident that many differing, "…specialized languages suited to specific domains were required to represent the numerous bodies of data used in those

domains" (Adler, Cochrane, Morar, & Spector, 2006, p. 208). These individual languages also needed to be able to be shared and maintained by differing technologies and by a wide body of users for different purposes. If each separate domain developed its own language and accompanying system, it would be a very inefficient use of time and resources because this data could not be shared across systems. XML was the solution to this problem in that it provided a very general approach for satisfying these common requirements. "It allowed the definition of languages in which information is encoded as tagged text and in which different encodings and tags support different domains of discourse" (Adler, Cochrane, Morar, & Spector, 2006, p. 208). Because XML was developed to be used by a wide array of disciplines and by both large and small scale interests, it needed to be open source, or nonproprietary, so multiple parties could contribute and ultimately share technology, information, and resources. According to the *Online Dictionary for Library and Information Science* (n.d.), open source is defined as, "A computer program for which the source code is made available without charge by the owner or licenser, usually via the Internet, to encourage the rapid development of a more useful and bug-free product through open peer review. The practice also allows the product to be customized by its users to suit local needs. To be certified 'open source' under the Open Source Initiative (OSI), software must meet certain established criteria that include no restrictions on access".

As XML standards were being developed, a community of diverse groups and individuals was also being established. There were a wide variety of reasons why both individuals and companies not only favored but also endorsed and participated in XML's open standards: "…to gain the benefit of an open community to supplement their own development resources, to take advantage of the positive marketing perceptions surrounding the participation in nonproprietary solutions, and to benefit from the vast market opportunities created" (Adler, Cochrane, Morar, &

Spector, 2006, p. 209). New standards and prototypes were now easier to develop with the work and support of an entire community with various abilities and background knowledge. "The desire not to be left behind the competition, customer requirements for interoperable solutions, and the simple economics of sharing in a common pool and community of interests all led to the rapid development and adoption of open standards" (Adler, Cochrane, Morar, & Spector, 2006, p. 209).

The fact that XML is open source greatly contributed to the markup language playing an important role in information retrieval. One of the great challenges in information retrieval is being able to successfully share and retrieve documents and information from across many databases and disciplines. Largely because it is nonproprietary, XML and its related standards were allowed to enable, "…data interoperability, content manipulation, content sharing and reuse, document assembly, document security and access control, document filtering, and document formatting across all disciplines and for all types of devices and applications" (Adler, Cochrane, Morar, & Spector, 2006, p. 209). When information is more accessible, it is available to a larger audience and ultimately achieves one of the major goals of information retrieval: true interoperability.

Solving Information Searching and Retrieving Dilemmas on the Web

XML's relationship to information access and retrieval is especially evident when it comes to resolving information and content retrieval errors with the use of XML and the many programs associated with this increasingly-used markup language. XML is commonly used by many differing types of information retrieval systems such as digital libraries, online databases and

OPACs, but the Web is one of the greatest challenges when it comes to information retrieval. "The Web is the world's greatest repository of information … if you can efficiently find what you're looking for" (Rogers, 2004, p. 19). When irrelevant documents and information are retrieved after a query, one must first consider what the problem is, and then how to remedy that error. "…the focal point for most content retrieval errors is the data itself" (Yager, 2000, p. 88). XML is a powerful tool that can aid users in correcting this flawed data to make it more accessible and thus ultimately more retrievable.

Information and documents on the Web are often described as being in silos or stranded on information islands when this data is not searchable beyond a single site. XML helps ease this problem and, "…makes it easy to expose information in a content management system to other sites, enabling searches that cover multiple databases across many sites" (Rogers, 2004, p. 19). XML "…enable(s) information reuse by integrating texts and data from different sources and by searching and linking across these sources, thereby breaking down traditional silos, which were barriers to information sharing" (Adler, Cochrane, Morar, & Spector, 2006, p. 210). With the help of this user-friendly markup language, users are able to characterize text within a document with the use of tags and labels so they have the ability to simultaneously search across multiple information retrieval systems, making search results on the Web much more accurate.

In addition to examining information retrieval on the Web, it is also important to consider information extraction (IE) when discussing the use of XML to aid in locating and retrieving documents and information. "Information extraction software identifies and removes relevant information from a variety of sources, pulling information from a variety of sources, and aggregates it to create a single view. IE translates content into a homogeneous form through technologies like XML" (Adams, 2001, p. 27). While there is certainly interplay between the

two, information retrieval mainly focuses on document retrieval whereas information extraction focuses on the retrieval of facts. Nevertheless, both must overcome difficulties of retrieving information on the Web such as language ambiguities including synonymy, polysemy, morphology, and homogeny (Lancaster, 1991). "XML is important because it facilitates increased access to and description of the content contained within the documents. The technology separates the intellectual content of a text from its surrounding structure, meaning that information can be converted into a uniform structure" (Adams, 2001, p. 30).

Because of XML's flexibility, it has the ability to work in conjunction with and be employed by other programs, techniques, and applications. AJAX (Asynchronous JavaScript and XML) is a perfect example of this kind of interoperability. While AJAX functions in a number of differing capacities, it is best known for its ability to retrieve data from a server (asynchronously) in the background without affecting the functions on the current page that is being displayed. This is achieved by making, "…a request to a server via Hypertext Transfer Protocol (HTTP) and continue to process other data while waiting for the response" (Clark, 2008, p. 32). While this may initially sound superfluous, it is being more commonly used in online information retrieval systems such as Web search engines, online databases, and digital libraries to aid in the searching, browsing, and retrieval processes. AJAX is responsible for simple techniques that many users take for granted such as username and password verification without having to lose or reload the data on the entire screen. Another example is keystroke matching; in other words, as users search for keywords, subjects, and titles in information retrieval systems, potential keyword matches are often displayed under the form field in order to help with faster entry and spelling correction for more efficient and successful information retrieval.

XML in the Library Setting

In addition to the myriad of ways that XML, specifically when being used in AJAX, can aid in perfecting searching and information retrieval on the Web, it also has practical applications in the library setting. For example, many of the ways that AJAX aids in more efficient Web searching could also be used to improve libraries' OPAC systems. Possible keyword matches, verification of a user's personal information or settings, and digital library search applications could be faster and done without loss of data or having to wait for pages to reload. "AJAX can also make searching and browsing library resources easier" (Clark, 2008, p. 32). Rather than having to click through the OPAC's various subject pages, AJAX enables the user to browse through the possibilities by simply rolling the mouse over the links. "Ajax reduces the need to click through to more information, bringing data into the user's working environment" (Clark, 2008, p. 32). In many cases, what may seem like a single search to a user is actually an unseen complex task as AJAX accesses multiple databases and libraries in order to bring the requested information to one location; this greatly improves the quality of the users' searches without them ever being aware of it. An increasing number of libraries are beginning to incorporate AJAX into their systems for various reasons as it aids in the improvement of information access and retrieval and it would be impossible without the multi-functionality of XML.

XML also plays an important role in the library setting with its strong impact on electronic records. An increasing number of libraries have adopted the use of electronic records rather than hard copy records because they are, "…easier to transmit, store, and access than the paper records they represent" (Winters, 2005, p. 64). Because of this wide acceptance, libraries

needed to determine how to manage these new records, and in many cases, XML has been adopted for this purpose. More specifically, XML aids in handling an item's format which is how the item is displayed, the structure which shows how to treat each item, and an item's meaning which interprets each item based on the given tags (Winters, 2005, p. 65). Using XML to handle electronic records is relevant to both the open source movement and information retrieval because once an item is tagged using XML, various other applications are able to interpret this data thus making it usable to more systems. In addition, more people are able to use these records because XML is human-readable in addition to being machine-readable; this is important because, "…it is possible to interpret an XML document without special training or a glossary of tags" (Winters, 2005, p. 66). Because XML is open source, users are able to customize the markup language to suit their own needs while still being able to share their resources. Finally, using XML not only allows libraries to share records and other types of resources, but also helps ensure that these electronic records are tagged and labeled consistently, ultimately making them more accessible for retrieval.

MARC (Machine Readable Catalog) standards, "…a suite of data element sets that provides the mechanism by which computers exchange, use, and interpret bibliographic data" (Radebaugh, 2007, p. 15) are commonly used in many major libraries. It is the foundation for most library catalogs that are used today and was introduced by the Library of Congress in the 1960s. In an effort to increase the sharing and exchange of bibliographic data, the Library of Congress' Network Development and MARC Standards Office has and is continuing to develop "… a framework for working with MARC data in a XML environment. This framework is intended to be flexible and extensible to allow users to work with MARC data in ways specific to their needs. The framework itself includes many components such as schemas, style sheets, and

software tools" ("MARC 21 XML Schema", 2009). MARCXML, or using MARC in XML syntax, is advantageous to libraries and information retrieval as a whole because in the past, only other libraries that used MARC were able to share data in these catalogs; they will continue to do so with MARCXML, but also be able to share it outside of that particular setting as well. Because of the integration of XML, this is now be a fairly easy transition because, "The MARCXML schema supports all MARC-encoded data regardless of the format" (Radebaugh, 2007, p. 15). Finally, libraries are now able to take advantage of the many tools that XML has to offer without abandoning MARC's advantages by adopting the hybrid MARCXML.

As we have seen from the above sections, one of the major goals in information access and retrieval is, "…the creation of data that is sharable, transformable among different systems, and can be remixed in part, or its entirety, in innovative ways" (Walker, 2007, p 28). Aggregation, which is to gather and reassemble separate sets of data, is one form of this remixing and can be observed at many levels in the library setting. Libraries aggregate books and electronic resources to provide the most and best possible resources for their patrons. "Many libraries offer aggregation through metasearch capabilities so that a user can quickly enter a single query and receive information that is drawn from a very broad and diverse set of sources (each of which is potentially a large aggregation)" (Walker, 2007, p. 28). Aggregation is closely tied to XML because as more information resources are being converted into XML, the cost and time of aggregating these resources is dramatically reduced. "Once the data is marked up with XML tags, it can be sliced, diced, and reconstructed into the features and presentation formats that are compelling to users" (Walker, 2007, p. 28).

Xrefer is a classic example of aggregated information sources that many libraries use today. "This online platform provides both Google's swift, electronic convenience and an

accuracy and focus that the search engine's web-scouring mechanisms cannot deliver" (Guz, 2007). With Xrefer, libraries also have the ability to download MARC records, which allows terms entered into Xrefer's search engine to find resources within the library's own collection. Xrefer uses XML compatible formats such as Xreferplus and SFX (Special Effects), the most widely-known OpenURL link server in the library and publishing community (Desmarais, 2004), to aggregate the many differing information resources into compatible formats.

Implications for the Field of Library and Information Science

Because of its foundation in the open source movement and its wide acceptance and broad uses, XML is certainly not limited to information retrieval or even to the overarching field of library and information science. Nevertheless, despite being a relatively new technology, it has already made a significant impact on the field and profession through its universality, interoperability, and strong economic impact. For these reasons and many more, "XML has become one of the most important and widely used paradigms in distributed computing" (Adler, Cochrane, Morar, & Spector, 2006, p. 209).

First and foremost, universality is an overall goal of the open source movement as well as in library and information science. To highlight just a few examples, information professionals strive to provide multilingual access to all programs and applications, specifically electronic, and to allow information systems to share resources. XML allows these opportunities with its inherent universality. "Any of the world's languages can be used for XML markup or content, due to its incorporation of Unicode as a key component" (Adler, Cochrane, Morar, & Spector, 2006, p. 207). XML defies regional, cultural, and linguistic barriers in that the user creates the

meta-language that is best suited to him or her, regardless of the language, profession, or purpose. Interoperability can also fall under the overarching objective of universalism; XML also aids in the library and information science field achieving that goal. Different users and communities have varying needs but also need to work together in order to share resources; XML provides the tools to fulfill this opportunity allowing more people to share information and materials. What is unique about XML is that it can be uniquely customized to the needs of individual users, yet has the ability to transcend the barriers of individual programs and systems allowing for optimal interoperability. "One of the most compelling aspects of XML's evolution was the intense and spirited collaboration of communities from different disciplines" (Adler, Cochrane, Morar, & Spector, 2006, p. 215). XML's inherent interoperability has united people of different experiences, expectations, backgrounds, and fields of study and has allowed them to work together to share ideas, expertise, and resources.

Finally, the economic impact of XML has also greatly affected the field of library and information science. As discussed earlier, because it is an open source technology, differing communities have all worked together to achieve the common goal of creating and using XML. In a strictly for-profit market, only large companies and firms would have been able to partake, thus eliminating the purpose of XML to unite systems of all purposes and sizes for the sake of sharing resources. By working together as a community rather than competitively, groups are able to drastically reduce costs and not be hesitant to implement these new technologies that benefit the individual, the community, and the field as a whole. As more systems incorporate XML in the future, "Distance, time, language and communication barriers will be vastly reduced" (Adler, Cochrane, Morar, & Spector, 2006, p. 219) in addition to reducing costs.

Conclusion

Although there have been a lot of positive and productive articles written and studies completed relating the importance of the XML to information retrieval in the field of library and information science, there are also some criticisms and the issue calls for additional research. Many professionals remain skeptical of XML's impact on the Web, the information technology field, and the library and information science community because this is still a relatively new technology. Others feel it is a facet of the Web 2.0 craze and simply adds more layers of abstraction or is a program without any real meaning or solid business model. Nicholas Petreley (2001) is not alone when he states, "XML is great as a standard way of saying, 'This next thing is a widget.' But XML doesn't require that you describe what the widget does, how it works or that the widget itself conforms to a standard" (Petreley, 2001). This is a classic example of the criticism that it is a fancy application, but has no solid meaning and will quickly disappear. As XML continues to be used in more settings and by more diverse groups of people, it will continue to be tested to see if it endures as a legitimate technology. For those in the field of library and information science, it is worthwhile to learn about and test these new technologies and in order to study the impact they may have on our field, regardless of their permanence. "… we must keep a firm grounding in the technologies that drive digital library development, even though they are changing fast. That means keeping current with HTML, XML, and the overall Web site administration" (Huwe, 2004, p. 41). Regardless of if these technologies will disappear in a few years or not, they are shaping our current understanding of the field as a whole so they are nevertheless worthwhile to continue to learn from and examine.

Works Cited

Adams, K. C. (2001). The Web as a database: new extraction technologies and content management. *Information Today, Inc.* (27-32).

Adler, S., Cochrane, R., Morar, J. F., & Spector, A. (2006). Technical context and cultural consequences of XML. *IBM Systems Journal*, 45 (2), 207-223.

Chowdhury, G. G. (2004). *Introduction to Modern Information Retrieval*. London: Facet Publishing.

Clark, J. A. (2008). AJAX (asynchronous JavaScript and XML): this isn't the web I'm used to. *Online*, 30 (6), 31-34.

Desmarais, N. (2004). XML in action. *Against the Grain*, 15 (3), 102-103.

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

Guz, S. S. (2007). The promise of Xrefer. *Library Journal*, n.d.

Huwe, T. K. (2004). Keep those Web skills current. *Computers in Libraries*, 24 (8), 40-42.

Kasdorf, B. (2008). The XML advantage. *Net Connect*, 12-15.

Kay, R. (2005). Markup languages. *Computerworld*, 30.

Lancaster, F. W. (1991). *Indexing and abstracting in theory and practice*. Champaign, Ill: University of Illinois, Graduate School of Library and Information Science.

MARC 21 XML Schema. (2009). Retrieved April 21, 2009, from http://www.loc.gov/standards/marcxml.

Open source. (n.d.) In Online Dictionary for Library and Information Science. Retrieved April

29, 2009, from http://lu.com/odlis/search.cfm.

Open Source Initiative. (2009). *Open Source Initiative*. Retrieved October 27, 2009 from http://opensource.org/docs/osd

Petreley, N. (2001). Ontology and the web. *Computerworld*, 35 (41).

Radebaugh, J. (2007). MARC 21 / MARCXML. *Computers in Libraries*, 27 (4), 15.

Rogers, B. (2004). Solving Web Search Dilemmas. *EContent*, 19.

Saunders, L. (1998). Not your mother's HTML: moving on to XML. *Computers in Libraries*, 18 (10), 45.

Taylor, A. G. & Joudrey, D. N. (2009). *The Organization of Information*. Wesport: Libraries Unlimited.

Walker, J. (2007). Sliced, diced, and reconstructed. *Netconnect*, 28.

Winters, R. (2005). XML marks the future for electronic records. *The Information Management Journal*, 39 (6), 64-68.

Yager, T. (2000). Search no further! Content retrieval is gaining XML boost. *Infoworld*, 21 (46).