XML and Better Web Searching

J. Jackson

Donald L. Gilstrap

XML and Better Web Searching

Joe Jackson and Donald L. Gilstrap

Abstract

This article addresses the implications of the new Web meta-language XML for World Wide Web searching. Compared to HTML, XML is more concerned with structure of data than documents. These XML data structures, especially when declared in document type definitions, should prove conducive to precise, context rich searching. Some of the directions the XML language is intended to move are briefly covered. Additionally, trends in World Wide Web development with respect to beta versions of the XML language are discussed.

Introduction

One aspect of the World Wide Web that has continued to frustrate online searchers accustomed to more organized data sets is the imprecision of search engines. Different search engines usually produce disparate sets of results for the same search terms, and the ranking of the sites that are retrieved is also often highly questionable. Even while search engines continue to add new features, there is little hope searching will ever improve dramatically as long as queries are confined to keyword searches of flat, featureless full-text files. Much of this problem resides in the typical HTML document, as the bulk of the markup in HTML simply determines how the browser will display the content; Gottesman describes HTML as "simply a presentation language" (Gottesman, 1998). Those few structural elements that are available in HTML, such as the header, have been reduced through usage to formatting options. Meta-tags provide a partial solution, but as we shall see, do little to improve actual searching.

During the past year, much has been written about XML and the potential impact it may have on the World Wide Web. Sperberg-McQueen writes, "It seems likely that XML is the future of digital libraries" (Sperberg-McQueen, 1998). Additionally, St Laurent writes, "XML is not an official replacement for HTML, but it definitely includes and extends HTML in a way that probably will stop HTML development in the fairly near future" (St Laurent, 1998). Whereas statements such as these may underestimate the staying power of HTML, XML is certainly becoming a priority, as both Microsoft and Netscape are incorporating XML functionality into their next generation of browsers. As Tim Bray, the editor of the XML specification for the Web says, XML and HTML will likely coexist "very intimately" ("XML editor bullish", 1998, p. 30).

Like HTML, XML is a derivative of SGML (standard generalized markup language), a complex and powerful meta-language that allows for the creation of platform-independent, structured documents, as well as other markup languages. But whereas HTML is an application of SGML, XML is a condensed, simplified version of SMGL, intended to provide similar features within a much easier and more manageable framework. In this manner, XML can both include and extend HTML, making it possible to convert existing HTML to XML, and to address the shortcomings of HTML. XML has one key

advantage over HTML that should have all Internet searchers awaiting its inception: XML's extensibility allows for the creation of custom tag sets that are in effect equivalent to data structures. XML, as Stanek notes, can "describe information as well as structure it" (Stanek, 1998). The gradual replacement of HTML- coded documents with data-centric, highly structured XML documents should greatly improve the precision of online searching on the Web.

XML and Web searching

In theory, Web authors attempt to create documents that are both rich in content and visually stimulating. The tools that have arisen for Web page development have made it easy for people used to desktop publishing to create attractive Web pages. However, while these Web pages rival their print counterparts in terms of visual organization through such devices as headings, tables, and style sheets, there is little in HTML that addresses content itself. HTML sacrifices power for ease of use, and as a result there is nothing in HTML to distinguish one table or one heading from another, except for the keywords enclosed within the tags.

At the present, most search engines pay little attention to markup, and focus instead on the content of the page. Consequently, results are produced mainly from the information found in the <TITLE> tags, or somewhere in the <BODY> of the document, the equivalent of full-text hits. For example, if one were to search "Mark Twain" on the Web, one could find the document shown in Figure 1.

This page might be ranked highly by certain search engines for the following reasons:

- it contains the exact term in the title of the document;

- the term appears early in the document;

- the term is repeated in the document.

Although this is not a comprehensive list of factors that will determine how highly a page is ranked by a search engine, they are the most common and generally serve as the foundation of the ranking algorithm for the major search engines. Unfortunately, these same rules could rank the document shown in Figure 2 as more relevant.

Whereas the page that actually deals with the author might reasonably mention Mark Twain once or twice, a business might repeat the name often, and thus be interpreted as more relevant. The problem here is that search tools operate without context. One could attempt to improve the search results by adding terms like "literature" or "writer", but unless the author of a page saw fit to include such terms, this strategy will have little impact on the quality of the pages retrieved.

Meta-tags could be a tremendous aid to Web search tools. Included in the header of the document, one can ascribe keywords to describe the content of the document through use of the "keyword" attribute. This allows Web authors to create the context that is sorely lacking in most HTML documents:

```
<HTML>

<HEAD>

<META NAME= "KEYWORD" CONTENT = "American Literature, Authors, Mark Twain, Works">

<TITLE> Mark Twain


</TITLE>

</HEAD>
```

Use of meta-content could help improve search results using search engines that incorporate meta-content into their ranking algorithm; unfortunately, only Infoseek and Hotbot currently factor meta-content into their formula (Sullivan, 1999). This approach is further limited, as McDermott notes, because many Web authors do not include meta-content in their Web pages (McDermott, 1999); also, this approach has failed to create standard terms that could be reused and shared within an entire industry.

So, how will a World Wide Web of XML documents improve searching? One important element lies in XML's ability to structure data as well as documents. HTML provides a full set of generic tools for structuring documents, such as tables and headings, which have no semantic properties. As St. Laurent notes, "Document structures help humans read documents, but they do little to help computers find the critical pieces of data they need without human assistance". On the other hand, XML data structures "reflect the content directly, with little concern for where they appear in the overall document structure" (St Laurent, 1998). Since XML allows for the creation of tag sets that are content savvy, an XML data structure can serve as a road map to information, as with the example shown in Figure 3.

A search engine would no longer be looking solely at the information inside the tags, but at the tags themselves. The tags in Figure 3 create a logical hierarchy that would situate search terms within the necessary context, and thus eliminate the irrelevant type of hits so common to keyword searching a Web of full-text documents. Thus an XML savvy search engine, which would rank markup hits much higher than content hits, would be able to distinguish quite easily between Mark Twain the author and the Mark Twain Insurance Company.

XML is able to accomplish these structuring tasks through the language's ability to associate an XML document with a document type definition (DTD). The DTD is where the structured tags are actually declared and attributed to certain values. DTDs can be included at the beginning of the XML document, or maintained as an external file. The external DTD, author.dtd, is declared in Figure 4.

DTDs are optional in the XML 1.0 specification. Documents that have an associated DTD are said to

be "valid" XML. Documents that conform to the rules of XML, but do not have a DTD are said to be "well- formed". Well-formed XML would be an attractive option for Web authors wishing to convert HTML to XML, without going to the time and expense of creating a DTD. However, DTDs provide the key to much of the promise of XML, since they give the end user an efficient means of associating search terms with what is in effect a list of key terms. In Figure 4, an ELEMENT has been declared for each item pulled from Figure 3. ELEMENT AUTHOR is actually a group element composed of the name, nationality, period, genre, and work, and sub-elements are then defined further through the list. The "+" symbol following GENRE and WORK is similar to the truncation with which librarians are familiar, declaring these group elements as having one or more genre types or that the author has more than one published work. Notice also the "|" symbol found between PERIOD and GENRE which also relates to the Boolean logic used by those who search databases. In this element group, the DTD will associate the XML document with either the period or genre in which it was written, depending on whether both elements appear in the document. PCDATA stands for Parsed Character DATA and basically implies that the information found within the element will be character data as opposed to numerical data (Bourret, 1998).

While standard tag sets are derived, one can envision the potential for precise searching. As seen in Figure 3, and implicitly through the element declarations in Figure 4, online users could limit searches by author, title, the time period in which a work was written, and, perhaps serendipitously, by Library of Congress classification schemes. Considering the plethora of different types of files found on the Web, XML even allows for searching by document extension.

A global access feature of XML is that it supports the International Standards Organization 10646 UNICODE standard (UTF-8 in Figure 2), which means that XML can decipher and categorize a variety of international languages without relying on plug-ins to accomplish translation tasks. As an end result, an XML enabled browser could display the search results of a query alphabetically, chronologically, by subject sets, or even by file type and language (Boeri, 1998; Floyd, 1998; Piven,1998). Basically, XML allows users to search for and manipulate data found on Web sites to suit their own needs.

XML could also help eliminate another bane to Web searching, namely the dead link. Links are generally hard-coded into HTML documents, and if the document being linked to is moved or eliminated, the familiar 404 "not found" message appears. The latest proposed XML linking specification, XML Linking, provides several attributes that improve on HTML's simple <HREF> tag, including linking to multiple sites with a single link, embedding linked material, and two-directional links. More importantly, it allows for the creation of extended links, making easier link updating and management possible. The problem of outdated links is currently addressed by relying on a uniform resource name (URN) server which redirects outdated links to a uniform resource identifier (URI); usually a parent URL that should remain consistent for a site even though a file within the site may be defunct. At the present, OCLC maintains the PURL service which redirects browsers to Web servers that maintain persistent naming schemes, although this process becomes complicated when dealing with inconsistent server addressing (Weibel and Jul, 1995). Eve Maler, the co-author of the XLink specification that preceded XML Linking, states that extended links provide an additional solution. Extended links can "reside outside any of the resources that are pointed to, and thus when a target resource changes in some fashion, you have an easier opportunity to update the link" (Maler, 1998). In

other words, if the URL of a resource changed, updating a single central link would keep all documents pointing to that resource up to date.

The future of XML on the Web

XML is still in its early stages, and several things need to occur before XML begins to fulfill its potential as a dominant markup language. First, tools need to be developed that will allow Web authors to easily create XML pages, as well as convert HTML to XML. Also, because XML does not work with the layout of pages, solutions need to be developed concerning the two most likely candidates to handle display functions with XML, the extensible style language (XSL), and cascading style sheets (CSS). Much as XML and HTML will likely coexist in the future, CSS and XSL will also serve different needs: "XSL is intended for complex formatting where the content of the document might be displayed in multiple places; ... CSS is intended for dynamic formatting of online documents in multiple media" (Lilley and Quint, 1999). XSL is currently in its second working draft, and there is yet no final recommendation for XSL 1.0, so any final solutions must wait until this work is complete. In addition, search engines need to be developed that will search XML effectively. Infoseek has already introduced a prototype, and it will be interesting to see how quickly the other major search engines follow suit (Oakes, 1998). And last, industry standards need to evolve that will require all documents to follow a common tag set. Once established, these standards will make it easier not only for people to add new data, by eliminating the need to come up with the DTDs, but also permit what will become field- specific searching on the Web. Many of these are already beyond their nascent stages, such as the chemical markup language (CML) and the mathematical markup language (MathML). Other uses are also currently being developed, including IBM's SpeechML, which will "use XML to deliver speech capabilities to Web browsers"; similar is Motorola's VoxML, which will open the Web to telephones (Walsh, 1999). Industry seems to be embracing XML, as Cover lists over 100 XML applications that are currently in the works (Cover, 1999). So, whereas searching the Web will benefit from the introduction of XML, one can see that XML will open up new capabilities as yet unimagined on the Web.

Conclusion

In 1997, members of the IEEE Computer Society argued the shift to "XML markup is a critical phase in the struggle to transform the Web from a universal information space into a knowledge network". Although subsets of XML are already being utilized being specific sectors of academe, it will be interesting to see how quickly the library community embraces XML. We as librarians, however, are in a unique position to take advantage of the opportunities the language provides for diffusing data found on the Web. But most importantly, XML's ability to structure Web documents in ways that have been relatively non-existent in the past will greatly improve the methods employed for searching the Web in the future.

```
<HTML>
<HEAD>
<TITLE>
Mark Twain
</TITLE>
</HEAD>
<BODY>
<H1>Mark Twain</H1>
Nationality: American<P>
Period: American<P>
Genre: Fiction<P>
Summary: Mark Twain was the pen name of Samuel Clemens, an American humorist
who lived from 1835-1910.
Works:
<UL>
<LI>Adventures of Huckleberry Finn - 1884
<LI>A Connecticut Yankee in King Arthur's Court – 1889
</UL>
</BODY>
<HTML>
```

Figure 1 *Possible document result for "Mark Twain" search on the Web*

```
<HTML>
<HEAD>
<TITLE>
Mark Twain Insurance Company
</TITLE>
</HEAD>
<BODY>
The Mark Twain Insurance has been in business since 1956. During that time, the folks at
Mark Twain have ....
Call Mark Twain Insurance today.
</BODY>
</HTML>
```

Figure 2 *Another possible document result for "Mark Twain" search on the Web*

```
<?xml version="1.0" encoding="UTF-8"?>
<DOCTYPE AUTHOR SYSTEM “author.dtd”>
<AUTHOR>
        <NAME>Mark Twain</NAME>
        <NATIONALITY>American</NATIONALITY>
        <PERIOD>19th Century</PERIOD>
        <GENRE>Fiction</GENRE>
        <WORK>
                <TITLE>Adventures of Huckleberry Finn</TITLE>
                <YEARPUBLISHED>1884</YEARPUBLISHED>
        </WORK>
        <WORK>
                <TITLE> A Connecticut Yankee in
            King Arthur's Court </TITLE>
                <YEARPUBLISHED>1889</YEARPUBLISHED>
        </WORK>
</AUTHOR>
```

Figure 3 *Example of XML data structure serving as road map to information*

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE simple [
<!ELEMENT AUTHOR (NAME,NATIONALITY,PERIOD | GENRE+,WORK+)>
<!ELEMENT NAME (#PCDATA)>
<!ELEMENT NATIONALITY (#PCDATA)>
<!ELEMENT PERIOD (#PCDATA)>
<!ELEMENT GENRE (#PCDATA)>
<!ELEMENT WORK (TITLE,YEARPUBLISHED)>
        <!ELEMENT TITLE (#PCDATA)>
        <!ELEMENT YEARPUBLISHED (#PCDATA)>
]>
```

Figure 4 *External DTD (document type definition)*

References

1. Boeri, R. (1998), "Intranet searching: a light at the end of the tunnel", EMedia Professional, Vol. 11 No. 6, pp. 63-6.

2. Bourret, R. (1998), "Declaring elements and attributes in an XML DTD", The Database Research Group at Die Technische Universitat von Darmstadt, 3 March 1999. Available: http://www.informatik.tu-darmstadt.de/DVS1/staff/bourret/xml/xmldtd.html

3. Cover, R. (1999), "Extensible markup language", March. Available: http://www.oasis-open.org/cover/xml.html#applications

4. Floyd, M. (1998), "Extreme markup", Web Techniques, pp. 38-41.

5. Gottesman, B. (1998), "Why XML matters", PC Magazine, Vol. 17 No. 17, p. 215.

6. Lilley, C. and Quint, V. (1999), "Extensible style sheet language", March. Available: http://www.w3.org/Style/XSL/

7. McDermott, I. (1999), "Searchers net treasure in Monterey", Searcher, Vol. 7 No. 1, pp. 23-8.

8. Maler, E. (1998), "Re: XML and broken links", March. Available: http://www.oasis- open.org/cover/maler980331.html

9. Oakes, C. (1998), "Infoseek goes bilingual", January. Available: http://www.wired.com/news/news/trchnology/story/16221.html

10. Piven, J. (1998), "XML stakes out Web future", Computer Technology Review, pp. 43-5.

11. St Laurent, S. (1998), XML: A Primer, MIS Press, Foster City, CA.

12. Sperberg-McQueen, C. (1998), "XML and the future of digital libraries", Journal of Academic Librarianship, Vol. 24 No. 4, pp. 314-17.

13. Stanek, W. (1998), "Structuring data with XML", PC Magazine, Vol. 17 No. 10, p. 229.

14. Sullivan, D. (1999), "Search engine features comparison chart", March. Available: http://www.searchenginewatch.com/webmasters/features.html

15. Walsh, J. (1999), "IBM, W3C advance XML standards", Info World, February, p. 16.

16. Weibel, S. and Jul, E. (1995), "PURLs to improve access to Internet", OCLC Newsletter, July, p. 19. Available (3 March 1999): http://purl.oclc.org/OCLC/PURL/ SUMMARY

17. "XML editor bullish on spec's future" (1998), PC Week, Vol. 15 No. 39, p. 30