

XML and HTML
A technology overview

Introduction

HTML and XML represent two very important facets of current internet information Technology. HTML has had the spotlight for sometime due to its prevalence as the standard display language for the Web, but XML is proving to have a more far reaching effect, as a standardized data description language.

Growth of the World Wide Web has rapidly spread to all areas of government, business and consumers. All sectors are discovering the benefits of Data Sharing and linked networks. Relatively inexpensive servers and easy to code HTML has made the World Wide Web the premier method of displaying and sharing vast amounts of varied data across the entire globe. While HTML's utility is unquestioned, its limitations became very apparent when business decided to do more than display static data on storefront sites.

The internet infrastructure enabled more complex Business to Business or B2B transactions to occur via private servers over the public internet. However the one-time nature of Web Pages and delivery became hindrance as the amount and types of data between companies had grown enormous. XML emerged as a standardized way to handle non standard data.

XML provides the portability and flexibility to handle anything from complex real-time insurance and actuarial claims processing with adjusters servicing customers at a catastrophe site via wireless internet, or a Digital Medical Records System (1) encoded on a smart chip embedded under a patient's skin. Even hardware vendors are getting into the act, marketing servers designed to parse XML formatted data more efficiently.(7) XML's capabilities, easy to use tools and "Web Friendliness" among the HTML user base

and DBA's alike will eventually overtake plain old HTML in providing most all Web-enabled data interchange.

Technology Overview

These technologies for Web content delivery share much in common yet they are distinctly different. Both XML and HTML share their roots in SGML, Standardized General Markup Language. SGML was developed during the 60's and as a new way of defining data. The basic Concepts in SGML were:

- (a) The notion of separating "content and structure" encoding from specifications for [print] processing;
- (b) The notion of using names for markup elements which identified text objects "descriptively" or "generically";
- (c) The notion of using a (formal) grammar to model structural relationships between encoded text objects. (3)

SGML is more customizable (thus flexible and more "powerful") at the expense of being (much) more expensive to implement. (3) XML and HTML came about as a streamlining of SGML elements suited for specific purposes. HTML Stands for Hypertext Markup Language. HTML was designed to control document presentation, basically how data looks when presented visually. XML stands for eXtensible Markup Language. XML was designed to describe the data inside documents and to focus on what the data is. Another way of putting it is that HTML is about displaying information, while XML is about describing information.

While this might seem confusing at first, it becomes apparent when you define the purpose. Take for example a historical text.(10) If you wanted to put the contents on the Web you would use HTML to present the document, separate the headings, specify what text is bold and how it looked to the users in a browser. It would not however, tell the user or a researcher much about the document itself. That's where XML comes in. By "Tagging" various parts of the document with custom XML tags a user could tell more about the contents, such as one section of the document is always mandatory, another section always had dates, one section was original and another part was is list, while another part contained the translator's comments.

(10)

It is very important to remember that XML was not designed to do anything. (6) It has no active code, no widgets or scripting. XML was produced to arrange, store and to transmit information. This next example is a note to Bob from Lucy stored as XML:

```
<note><to>Bob</to><from>Lucy</from><heading>Hot Date</heading><body>Remember to bring the
Magnum of Champagne and Howler Monkeys!</body></note>
```

This note has all the trappings of HTML, what with everything tagged. But the tags that are being used are not part of the HTML standard. However it is pretty easy to figure out what the parts of the document are. You can see that it has a header and a message body. Also tagged are the sender and receiver information. But still, this XML document does not do anything. It is solely information wrapped in XML tags. A piece of software must be written to send, receive or display it. It will not display in your web browser properly. At least not without the HTML tags. XML tags are not predefined. You must "invent" your own tags and then create an application that parses that. The tags in the example above (like <to> and <from>) are not defined in any XML standard. These tags are "invented" by the author of the XML document. Like SGML, XML allows the author to define his own tags and his own document structure.

In contrast the tags used to mark up HTML documents and the structure of HTML documents are predefined. The author of HTML documents can only use tags that are defined in the HTML standard (like `<p>`, `<h1>`, etc.). For Example:

this page is represented as.

`<html>`

`<head>`

`<title>New Page 1</title>`

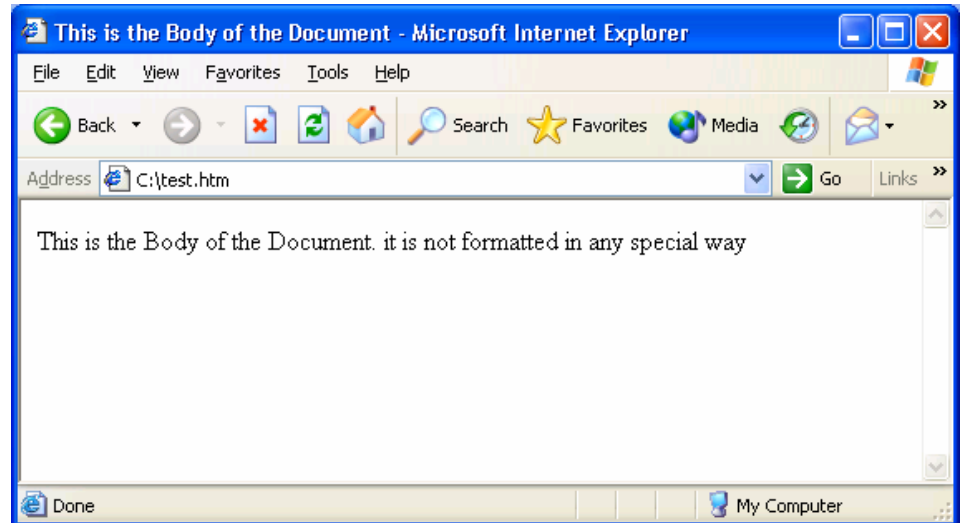
`</head>`

`<body>`

`<p>This is the Body of the Document. it is not formatted in any special way</p>`

`</body>`

`</html>`



Current Trends

The current industry trend, not surprisingly, is to develop new hardware. As more and more data-intensive applications are being developed with Web use in mind, hardware is being adapted to the new data flows.

“Research by IBM Labs shows that even small XML-based documents can increase the CPU cost of a relational database transaction by up to 10 times in the absence of a dedicated XML processing engine. The research concluded that XML parsing could have a “potentially fatal impact” on high-performance,

transaction-oriented database applications that use XML."(7) Given the trends in computing to move routing, encryption and decryption tasks to the networking layer, it's only logical to start building more efficient infrastructure components.(7) According to Frank Dzubeck, president and CEO of Communications Network Architects, "Everything is moving toward XML, it's business traffic, it's something that is no longer the exception, it is the rule,"(7) Of course like many other hot tech trends manufacturers seem to be jumping on it without any clear direction of where all the new hardware should go in the infrastructure. "No one really understands where this is going, but these vendors are making it so any system can process XML," say Tom Rhineland, an analyst with New Rowley Group. (7)

Hardware aside, XML is rapidly spreading rapidly over the net as businesses and governments use it for communication and data sharing. XML may even have a role in Homeland security. Among the failures by our intelligence services that were uncovered in the wake of the 911 investigation, were infrastructure and data sharing problems due to antiquated and incompatible systems. The 911 Commission recommended that "information be shared horizontally, across new networks that transcend individual agencies." (9)

An example can be made by considering by a situation whereby an electronic document contains classified material; however this document may also contain material that is not classified and necessary for another investigation. XML could be used in this case to "markup" the document to help screen out material that may or may not be relevant to an investigator, while maintaining the overall security of the classified portions of the documents. Parsing of the custom tags could be handled by the front end data retrieval application or in the future, by the aforementioned hardware devices. These devices could have security routing rules built-in. In this way when the user logs into the system, his credentials are already known and XML aware routers will be able to filter any requested data based on his supplied clearance level. This is

opposed to today's methods where system access clearance is not generally linked to intra-document security, but file repositories as a whole.

Additionally the 9/11 commission put forth a report from the Markle group "Creating a Trusted Network for Homeland Security" which describes "a network of interconnected databases that utilize directory services, metadata standards like XML, encrypted storage systems, search tools, rights management and authentication technologies that would give analysts and agencies unprecedented access to terror intelligence." (9)

Pros and Cons

For all the hype about XML, there are some caveats about replacing HTML wholesale. For one thing, HTML already has a big advantage; documents coded in HTML can be read in most any browser. Since XML is a document description language and not a presentation language, a separate application must be used to display and manipulate it. That said, it's safe to say that XML was not designed to replace HTML. "In future Web development it is most likely that XML will be used to describe the data, while HTML will be used to format and display the same data." (6). Indeed, the W3 Consortium has already drafted the XHTML 1.0 standard as a way of getting the best of both worlds.

"XHTML is a family of current and future document types and modules that reproduce, subset, and extend HTML. XHTML family document types are XML based, and ultimately are designed to work in conjunction with XML-based user agents." (8) This new version provides the customizable Tagging and database features of an XML Document with the standardized web browser display capabilities of HTML. The W3 consortium is really pushing this new hybrid as a replacement for HTML. More advantages provided by the new standard include backwards compatibility. "XHTML documents can be

written to operate as well or better than they did before in existing HTML 4-conforming user agents as well as in new, XHTML 1.0 conforming user agents.” (8) Other advantages cited by the W3

Consortium are:

- “XHTML documents are XML conforming. As such, they are readily viewed, edited, and validated with standard XML tools.
- XHTML documents can utilize applications (e.g. scripts and applets) that rely upon either the HTML Document Object Model or the XML Document Object Model.
- As the XHTML family evolves, documents conforming to XHTML 1.0 will be more likely to interoperate within and among various XHTML environments.
- Document developers and user agent designers are constantly discovering new ways to express their ideas through new markup. In XML, it is relatively easy to introduce new elements or additional element attributes. The XHTML family is designed to accommodate these extensions through XHTML modules and techniques for developing new XHTML-conforming modules (described in the XHTML Modularization specification). These modules will permit the combination of existing and new feature sets when developing content and when designing new user agents.

Alternate ways of accessing the Internet are constantly being introduced. The XHTML family is designed with general user agent interoperability in mind. Through a new user agent and document profiling mechanism, servers, proxies, and user agents will be able to perform best effort content transformation. Ultimately, it will be possible to develop XHTML-conforming content that is usable by any XHTML-conforming user agent. “ (8)

Technology Evaluation

XML and its integration with HTML provide many possible opportunities in improve data sharing over the web or any other network infrastructure. However the reality is that HTML and closed databases will be with us for sometime. The simplicity of HTML coupled with the bewildering array of tools accessible to users of all skill levels, gives HTML an advantage that will last for at least several years. XML is still in the province of dedicated IT types, but more and more tools like “Altova’s XML Spy” are providing the same kinds of tools sets that HTML coders have enjoyed for years. Eventually, XML will assume the importance in the day to day operation of the web that HTML enjoys today and that XML or XHTML will be the most common tool for all data manipulation and data transmission.

References

1. **ATMs for healthcare DCL news editorial** April 15th, 2004
<http://www.dclab.com/atmshealthcare.asp>
Landon Bain See Above
2. **XML Specified for Clinical Document Architecture (FDA Policy Document)**
<http://www.fda.gov/OHRMS/DOCKETS/98fr/04-2536.htm>
3. **SGML History Last modified: July 12, 2002**
<http://xml.coverpages.org/sgml.html>
4. **Comparison of SGML and XML**
World Wide Web Consortium Note 15-December-1997 James Clark
<http://www.w3.org/TR/NOTE-sgml-xml.html>
5. **XML, Java, and the future of the Web** Jon Bosak, Sun Microsystems
<http://www.ibiblio.org/pub/sun-info/standards/xml/why/xmlapps.htm>
6. **W3 schools XML Study Guide**
<http://www.w3schools.com/xml/>
7. **“Vendors to target XML traffic jam” John Fontana Network World, 05/03/04**
<http://www.nwfusion.com/news/2004/0503xmlaccel.html>

8. **XHTML™ 1.0 The Extensible HyperText Markup Language (Second Edition) Reformulation of HTML 4 in XML 1.0 W3C Recommendation 26 January 2000, revised 1 August 2**
<http://www.w3.org/TR/xhtml1/#xhtml>

9. **02:00 AM Jul. 31, 2004 Wired Magazine 9/11 Report Iffy on tech**
<http://www.wired.com/news/business/0,1367,64412,00.html>

10. **The Electronic Text Corpus of Sumerian Literature: Manual. Jarle Ebeling Graham Cunningham Jeremy Black Oriental Institute University of Oxford, 2003**
<http://www-etcs1.orient.ox.ac.uk/edition2/TheETCSLManual.html#background>