Workshop on
Multimedia and Internet Technologies
26th to 28th February, 2001
DRTC, Bangalore

Paper: AB

XML and Electronic Publishing

Ayesha Taranum

Informatics India Ltd. Bangalore - 560 003

Abstract

XMI, is a technology recommendation of the W3C (World Wide Web Consortium). XML is concerned with the electronic representation of the structure and content of information. It is a simplified subset of an ISO standard known as SGML (Standard Generalized Markup Language). Extensible Markup Language (EVIL,) is the key to creating markups that can be used by any number of applications beyond the Web browser. XML is fast gaining popularity for primary data storage and transfer mechanism for web-based applications. It is being used extensively in Electronic publishing, the web and for Electronic Commerce. This paper in tends to broadly deal with the various features of EVIL, its applications to libraries, both traditional and digital, and challenges for creating and managing content.

1 Introduction to Electronic Publishing

Electronic publishing, which was considered the wave of the future, is here and hence we can say that the future is here. This concept is already making waves at various levels, making its presence felt at the world of education with institution's websites, electronic journals, digital library initiatives, archives, and lots of creative activities at the academic and non-academic levels. Faculty and researchers are exploring and exploiting the capability for innovation offered by rapidly changing information technologies. The concept of distance learning will also be highly dependent on the availability of electronic content for its success and electronic publishing will be a rich source for distance learning.

The term "electronic publishing" is one of those popular buzzwords that cover a multitude of things, which are in fact disparate in certain significant ways. "Electronic" - one where the text is read on, and/or printed from, the end user's computer rather than as print on paper. Electronic publishing includes not only on-line publishing but also CD-ROM and related technologies such as CDI. Electronic Publishing is an increasingly important means of publishing with significant differences from traditional ink on dead trees approach. This is surely an area in which the information profession needs to be proactive.

Electronic publishing, defined broadly, encompasses the creation and access to digital publications, the ability to deliver print-on-demand service, the support of electronic commerce, and the guarantee of a well-maintained, permanent digital archive or some other form of document preservation.

At the same time, economic factors, notably the inexorable double-digit increase in the cost of scientific journal literature, have encouraged librarians and administrators to view electronic publishing as a cost-effective alternative to traditional paper publications. The shift to a world in which the primary mode of scholarly exchange and communication of research findings is digital will profoundly impact libraries, the university, and scholarly publishers. Although even experts find it impossible to predict the future, there is a strong belief that electronic publishing promises both intellectual and economic benefits for higher education. If publication patterns change from the present model, in which faculty publish the majority of their research in commercial journals that university libraries purchase for millions of dollars, to a model in which scholars and researchers disseminate their findings and analysis at low-cost over the Internet, substantial savings will accrue, first to the libraries, and then elsewhere in the university.

2 Changing Publishing Models

The increasing usage of the Internet is the major cause for the rise of Web-based publishing by the academic community. Over the past few years, there have been significant changes, a major change is the technology necessary for electronic publishing and multimedia presentations -both hardware and software - has gotten dramatically faster, better, and cheaper. Challenges faced are:

2.1 Constantly changing Publishing technology

With every passing day we see the birth of a new publishing format. The evolution of a new format requires re-working of the entire document. With the arrival of new formats, the older formats become obsolete, and data caught in these obsolete formats is called as `legacy data format', which ends in costly conversion to the new format. The last few years have seen quite a few formats like: Windows 3.1 Help, Windows 95 Help, Microsoft Multimedia Viewer, RTF 2, RTF 6, RTF 7, PDF, Postscript, Maker Interchange Format, TeX, Lotus Notes 3 and 4, Folio Views 3 and 4, HTML 2,3 and 4. Dynamic HTML, HTMLHelp, etc. etc.

2.2 New Paradigms

Hypertext is becoming an increasingly important part of electronic publishing. The problem with hypertext lies in creating and managing large collections of hypertext links, tailoring them to the users needs. The success and failure of HTML stem from the same cause. HTML describes the way that a page looks, which is its greatest strength. It proves it frontiers in the number of tags it uses. It allowed non-experts to mount pages, which would be accessible even if invalid.

2.3 Volume of publications

The floppy disk has given way to the 60ONM CD which will soon give way to the 16GB DVD etc. The sheer volume of textual information that can be packed onto these digital delivery media is new territory for the publishing sector. Managing, searching, developing user interfaces for such vast volumes of information is hard.

2.4 Users Needs

The user is the ultimate person to be catered to, and with the growing days the users are increasing their demands that the published products be tailored to their individual needs. Different users want different views of the same information, different document orders, formatting, search functionality etc. etc.

3 What is XML

XML is a technology recommendation of the World Wide Web Consortium (W3C) Perhaps the most important web organization, the W3C was founded in October 1994 "to lead the World Wide Web to its full potential by developing common protocols that promote its evaluation and ensure its interoperability" (W3Q. An example of one of these protocols is Hyper Text Markup Language (HTML) It is this simple markup language that is the basis of web publishing, and its success is synonymous with the Web itself. XML is concerned with the electronic representation of the structure and content of information. It is a simplified subset of an ISO standard known as S^ (Standard Generalized Markup Language). Current XM1-related working drafts include eXtensible Stylesheet Language (XSL) XML Linking Language (XLink), and XML Pointer Language (XPointer).

XML is a meta-language-a language for creating languages. An XML derived language is a grammar consisting of named information elements organized into a hierarchical and/or recursive structure. The order and occurrences of information elements can be controlled by a grammar specification known as a DTD (Document Type Definition).

XMI, customizes SGML in a number of significant ways. First, a specific choice of syntax characters was made so that everyone using XML will use the same concrete syntax. For example all start tags must begin with "<" and end with ">". Second, a new empty-element tag may be used to indicate that this is an empty element and that an end tag is not expected. This new empty-element tag is like a start tag with a slash character just before the closing greater-than angle bracket. Third, tag omission is not allowed as it is in SGMIL. This means that each non-empty element will have a both a start tag and an end tag

The SGML Editorial Board has set the design goals for XML as:

- XML shall be straightforwardly usable over the Internet.
- XML shall support a wide variety of applications.
- XML shall be compatible with SGML.
- It shall be easy to write programs, which process XMIL documents.
- The number of optional features in XML is to be kept to the absolute
- minimum, ideally zero.
- XML documents should be human-legible and reasonably clear.
- The XML design should be prepared quickly.
- The design of XML shall be formal and concise.
- XML documents shall be easy to create.
- Terseness in XML markup is of minimal importance.

4 XML and electronic publishing

Information flexibility is essential as Web pages begin to undergo the transition from the HTML-based format at present to XML, or Extensible Markup Language. Like HTML, XML uses tags to define text characteristics, but the tags can also contain a lot of information about the text being tagged. You can think of the tags as style sheets providing additional supporting content that can be turned off and on - or added, changed, or deleted - as needed.

XML has a number of important facets that help address the electronic publishing problems as discussed in the section XML is an open, vendor neutral, formally defined standard. XML encourages the removal of presentation information from document data. Presentation information is layered onto XML data at point of presentation using style sheets.

XML provides simple but flexible, standards-based format, which captures the information content of a document separately from any styling and display instructions. This helps the document creator to focus on its content with display decision being taken later. Visual display possibilities and web publishing functionality far exceed those provided by HTML.

XML is independent of the application that created it. Thus if the vendor/developer of editor or database or browser disappears, access to my data does not disappear with them. XML data will never become legacy data. It will always be possible to programmatically extract the content and structure of an XML document.

XML documents are essentially hierarchical databases. Information in them can be programmatically located, harvested and re-used over and over again. Contrast this with the typical single-use lifestyle of a WYSIWYG document.

With XML it becomes feasible to target multiple output formats from a single XML source document. Moreover, it becomes feasible to do it in a completely automated fashion. The revolutionary aspect of XML is the modularization of information. Information presents itself as a self-describing unit that can do not inhibit processing, storing or display. Topical subject qualifiers (e.g., attributes) are placed at the appropriate level of granularity.

4.1 Creating a XML Document

Constructing a XML document is not a difficult task once you familiarize yourself with the components, which you can use. The main components of an XML document are **ELEMENTS.** An example of an XML element is as illustrated:

<ArticleTitle>XML and web publishing</ArticleTitle>

Elements are made up of an opening tag, which contains the element name, the content, and a closing tag, which again contains the element name. Every XMIL document has to have what is called a 'root element'. This is a pair of tags, which enclose and describe the whole document. It is usually straightforward to pick the name of a root element. For example if we were constructing this whole paper as an XMIL document, the root element would probably take the name 'Article'. In the above example of an element the content used is called 'character data'. This term describes information that contains no markup. As well as just character data, an element can contain other element(s), a mixture of other element(s) and character data, or nothing at all. An example of an element containing another element is as follows:

<Author><FirstName>Ayesha</FirstName></Author>

An example of an element containing another element and character data is as follows:

<Author>Ayesha<Lastname>Taranum</Lastname></Author>

What is important to note from this example is that when elements contain other elements, your information will start to have a hierarchical structure XML, documents have one 'root' element in which other elements may appear, in which other elements may appear and so on. This is known as the 'logical tree structure' of XML, document. An example for an element without content is:

<Company name= "Informatics India Ltd."/>

The example has a trailing slash only because it is an empty element.

Another important component of XML is **ATTRIBUTES,** attributes are properties of elements. In the example above, 'Company' is the element and the attribute 'name' is a property of that element with the value Informatics India Ltd.'. Elements with attributes, then, take the form:

<ElementName attribute-name="attribute-value">Content</ElementName>

An element can have more than one attribute, for example 'telephone-number' could have been another attribute of the element 'Company'. You are also allowed to give a different element an attribute of the same name, for example an element 'Product' could also have the attribute 'name', but to save yourself any confusion when it comes to defining your attributes in a DTD it is best to try and avoid this in the same document. For example, you could give the element 'Product' the attribute 'title', which would be just as fitting.

So the question is: when should we use attributes and when should we use elements? The answer, that is up to us, but we should bear in mind that, whereas elements can contain other elements, attributes cannot contain any other attributes. One final thing to note when you are defining attributes is that the attribute values must be surrounded by quotation marks but that these can be either single or double. If you use double quotes, you can then use single quotes within the value and it will not be a problem. Similarly, if you use single, then you can use double within.

Entities allow us to insert information into a particular place, or various other places, of an XML document. If we put an 'entity-reference' in a certain place, then when the XML file is being processed, it will replace that reference with the 'entity-content'. The entity-content could be a word or phrase, or even an entire XML document. It is in a DTD, though, where we must define what will replace an entity reference.

5 Conclusions

The Benefits of XML are

- Metadata is data about data. For example if we are looking for a specific book, the first thing that we would try to find is the library-record for that book, which would tell us where we might find it. That record is a good example of metadata. It contains details about the book such as title, author, publisher, and ISBN number. XML is another example of metadata. Each set of tags describes the information that it contains. And such metadata can prove very useful; therefore metadata is added value to the information content itself This means that XML has the potential to tell us a lot more information about a document than that a library-record will be able to tell us about a book.

- Whereas HTML tags describe the appearance of information, XML tags describe the meaning of information.

- Currently, web-browsers will read even broken HTML code. The same will not be true of XML

XML is a clear and well-formulated markup language. Its benefits outweigh the drawbacks. It is those benefits that will be key in persuading people to adopt it. In that sense we can say that XML will indeed play an important role in the future of web markup. XML is fast appearing to be the future of electronic publishing. The shortcomings can be overcome as standards stabilize.

6 References

1. CALIRE (Warwick) and ELLIOT (Pritchard). 'Hyped' text markup language. XML and the future of web markup. *ASLIB Proceedings,* 52(5), May 2000.
2. SCAN (McGrath). A Python Based Production System for High Volume Electronic Publishing. *Digitome Electronic Publishing.* *http://www.digitome.com/*

3. FRANCO (Mastroddi). Electronic publishing trends and advances, The Impact of Electronic Publishing on the Academic Community. *In* An Intennational Workshop organized by the A cademia Europea and the Wenner-Gren Foundation Wenner GrenCenter, Stockholm, 16-20April 1997. *http://tiepac.portlandpress.co.uk/books/online/tiepac/session4/chl.htm*

4. ELIA (T. Ben-Ari) Electronic publishing: past, present, and future. *Bioscience,* March, 1999.

5. STUART (Campbell). XML Perception to Practice. *http://www.freepint.co.uk/issues/250500.htm*

6. ELLIOT (Pritchard). XML the future of web markup, *Dissertation* M.Sc. in Information Management, University of Sheffield - Department of Information Studies, 1998/1999.

7. ANDENSON (M. J.). What is XML, and what does it mean for you? *Austin Business Journal,* 1999. *http://www.amcity.com/austin/stories/1999/05/31/focus8.html*

8. ANN (Apps) and ROSS (MacIntyre). XML Using an Evolving Standard in Electronic Publishing.

9. The XMIL industry portal. *http://www.xml.org*

10. OASIS. *http://www.oasis-open.org/*

11. XML Specifications. *http://www.w3.org/TR/1998/REC-xml-19980210*

12. Electronic Publishing. *http://www.huridocs.org/elecpubl.htm*

13. Web Developers Virtual library. *http://www.stars.com/Authoring/Languages/XML/*

14. *http//:www.idgnet.com/*