THE STUDY OF THE DOCUMENT FORMATS ON INTERNET WITH SPECIAL REFERENCE TO HTML, SGML AND XML

by

Vidya Varidhi Upadhyay*
Vivekanand Jain**

ABSTRACT

The paper describes different document formats used on Internet for communication and document delivery service. Due to need of the information society various types of document formats are used by them for above purpose. An attempt has been made to examine commonly practiced formats with their characteristics in special reference to HTML, SGML, XML and their applications in Libraries.

*** Information Scientist, Banaras Hindu University, Varanasi**
**** Assistant Librarian, Banaras Hindu University, Varanasi**

0 Introduction

Internet comprises of interlinked computer network around the world facilitating communication among all computers concerned to these networks. In other words Internet is network of networks. Internet is a global phenomenon which allows us to send or receive mail or to communicate with anybody by following simple protocols. These protocols have been designed to ensure that the computers are following some common languages when sending and receiving messages. The most common protocol is TCP/IP (Transfer control protocol/Internet Protocol). The US Governments Advanced Research Projects Agency devised long-distance computer network called ARPANet now a days, the Internet has emerged as largest information service provider to the society on the globe. Now it has created the new age, known as the information age. Depending on the need and resources in hand various types of document formats are available today. In this paper an attempt has been made to briefly describe the main document formats and as we know that due to vivid and complex applications of Internet it is not possible to use only one markup language as a whole so we tried to evaluate the features of HTML, XML and SGML.

1 Various types of document formats

Leaving aside non-digital media (paper, film, magnetic tape and so on) the range of electronic formats remains huge. It includes:

1. American Standard Code for Information Exchange (ASCII) character strings, containing virtually no structural information.

2. A variety of word-processor formats such as Word and WordPerfect; rich text format (RTF) can be used as a common transfer format.
3. Bitmaps of various types, which are a bit by bit image representation and are sometimes used to store text as an image.
4. Various e-mail formats, such as Simple Mail Transfer Protocol (SMTP) and Multi-purpose Internet Mail Extensions (MIME), which can be used for delivering documents as attachments and alike.
5. Multimedia formats, including the Synchronized Multimedia Integration Language (SMIL) and Macromedia's ShockWave.
6. Audio formats, including, RealAudio, Digital Audio based Information SYstem (DAISY), AudioInter change File Format(AIFF) , Sun's AU, and Microsoft's WAV.
7. Standard Generalized Mark-up Language (SGML), of which HTML is an application; some journals have used version of SGML; for any application of SGML there needs to be a Document Type Definition (DTD) which defines the rules for marking up and thus interpreting the document.
8. Hypertext Mark-up Language (HTML) used for hypertext, including web pages, and the most commonly used DTD of SGML; Dynamic HTML (DHTML) is used to enable interaction and greater control over presentation of web pages.
9. A variety of still images formats , including the Graphic Interchange Format (GIF), Joint Photographic Expert Group (JPEG),Tagged Image File Format (TIFF) and so on.
10. eXtensible Mark-up Languages (XML), which is a simplified subset of SGML that provides support for user-defined tags and attributes (unlike HTML) where these are defined in the HTML standard); XML is of enormous importance for libraries and is described in greater detail below.
11. Video formats, including Motion Pictures Expert Group (MPEG) Audio Video Interleaved (AVI) from Microsoft and QuickTime from Apple.
12. Portable Document Format (PDF), a proprietary format from Adobe which nevertheless has a huge and growing application base, especially for downloading documents from the web; it is in effect an enhanced PostScript format, proving embedded fonts, hypertext links, index and search facilities and a number of other features.
13. TeX, used especially in the scientific, engineering and mathematical fields because of its extensive facilities for representing mathematical formulae.
14. Postscript, a method of defining the look of page representation or 'page description language'- it defines the printed page rather than a display.

2 Markup and Markup Languages

The word markup was originally used to describe annotation or other marks within a text intended to instruct a compositor or typist how a particular passage should be printed or laid out. Examples, familiar to proofreaders and others, include wavy underlining to indicate boldface, special symbols for passages to be omitted or printed in a particular font, and so forth. As the production of texts was automated, the term was extended to cover all sorts of special ``markup codes'' inserted into electronic texts to govern formatting, printing, or other processing.

Generalizing from that sense, we define markup, or (synonymously) encoding, as any means of making an explicit interpretation of a text. At a banal level, all printed texts are encoded in this sense: punctuation marks, use of capitalization, disposition of letters around the page, even the spaces between words, might all be regarded as a kind of markup, the function of which is to help the human reader determine where one word ends and another begins, or how to identify gross structural features such as headings, and syntactic units such as dependent clauses or sentences. Encoding a text for computer processing is in principle, like transcribing a manuscript from scriptio continua, a process of making explicit what is conjectural or implicit. It is a process of directing the user as to how the content of the text should be interpreted.

A markup language, may be no more than a loose set of markup conventions used together for encoding texts. A markup language must specify what markup is allowed and whereabouts, what markup is required, how markup is to be distinguished from text, and what the markup means. What the markup is intended to do. To understand and act upon the markup, additional semantic information is needed, which will differ in different situations.

3 SGML

SGML is the Standard Generalized Markup Language (ISO 8879:1985). SGML is very large, powerful and complex. It has been used in very large industrial and commercial areas. SGML is the meta-language, i.e., it is a language for describing markup language. It is useful for very big size libraries where the variety of jobs are to be done and workload is tremendous.

Essentially, SGML is a method for creating interchangeable, structured documents; with it, one can do the following:

(1) assemble a single document from many sources (such as SGML fragments, word processor files, database queries, graphics, video clips, and real-time data from sensing instruments);

(2) define a document structure using a special grammar called a Document Type Definition (DTD);

(3) add markup to show the structural units in a document; and - validate that the document follows the structure that you defined in the DTD. The official definition of SGML is in the international standard ISO 8879:1986.

4 HTML

HTML is the Hyper Text Markup Language (RFC1866) a small application of SGML used on the web. It defines a very special class of report-style documents, with section headings, paragraphs, lists, tables and illustrations, with a new informational and presentational items and some hypertext and multimedia.

HTML stands for Hypertext Markup Language. Let us take each of these words in sequence.

Hypertext is ordinary text that has been dressed up with extra features, such as formatting, images, multimedia, and links to other documents.

Markup is the process of taking ordinary text and adding extra symbols (for example, an editor’s proofreading symbols are a type of markup). Each of these used for markup in HTML is a command that tells a browser how to display the text. Markup can be very simple, or it can be very complicated. Either way, the underlying text being marked up is always present and viewable.

Language is actually a key point to remember about HTML. HTML is a computer language, related to computer programming languages (like C, C++, Java ,C# etc). HTML has its own syntax and rules for proper communication.

Markup languages are a special type of computer languages because they are solely concerned with classifying the parts of a document according to their function in other words indicating which part is the title of the document, which part is a subheading computer language, which part is the name of the author, and so on. It not really correct to speak of "programming HTML" because HTML is not a programming language. Instead HTML is a markup language that has a different goal than creating a program. Moreover HTML is neither a page-layout language nor a printing language. The only thing HTML does is classify parts of the document so that a browser can display it correctly. This allows documents to be displayed on many different kinds of platforms. (Platform is the combination of computer hardware and operating system). Although HTML has evolved to the point that it contains many layouts and commands, these functions are secondary to HTML's role to classifying the logical parts of your documents.

Advantages of HTML

(1) **Vastness:** HTML, can used to put a documents on not just on a computer screens, but also printers, fax machines, TV sets, game consoles, Braille devices, digital watches and text to speech machines.

(2) **Price:** HTML is free of cost. There are no expensive licenses to buy and no annoying upgrade to purchase.

(3) **Independence:** One need not stuck to any one vendor or any one program; One don't have to worry about bugs in particular editing program .

(4) **Flexibility:** It does not matter which computer we are using. A simple text editor is sufficient to edit raw HTML. One need not dependent on a particular piece of software.

(5) **Deeper Understanding:** One will have a better understand of the structure of the one's page.

5 XML

XML is a eXtensive Markup Language, which is within the SGML family, is designed to provide the shell or framework within which anyone can create a specialized mark-up language of their own, so that any community can create its own tags. There are no predefined tags, as there are in HTML. There are, however, a few simple rules have to obeyed, such as tags always coming in pairs so that they surround the data to which they refer .

Why XML should be use in place of HTML ?

XML allows groups of people or organizations to create their own customized markup applications for exchanging information in their domain (music, chemistry, electronics, linguistics etc.).

Information content can be richer and easier to use , because the descriptive and hypertext linking capabilities of XML are much greater than that of HTML.

XML can provide more and better facilities for browser presentation and performance, using CSS and XSL style sheets.

It removes many of the underlying complexities of SGML in favour of more flexible model, so writing programs to handle XML is much easier than doing the same for full SGML.
Valid XML files are kosher SGML, so they can be used outside the Web as well, in existing SGML environments.

Information will be more accessible and reusable, because the more flexible markup of XML can be used by any XML software instead of being restricted to specific manufacturers as has become the case with HTML.

In fact XML has been developed largely in response to the acknowledged shortcomings of HTML, which is concerned primarily with the appearance of a web page on a screen rather than with what the content actually is.HTML was never designed for the king of complex task that has now become routine on the web. Because HTML tags have to be approved by an international committee , the process of adding a new tag is time consuming and laborious, and in any case it is highly unlikely that a specialist tag – say for an application specific to libraries dealing with incunabula-would be approved.

Use of XML in Libraries

There are many reasons for the importance of this new approach to libraries, since it allows information objects to be analyzed and manipulated and permits much more sophisticated handling of metadata objects. A particular example of the advantages to users would be the ability of client software to interpret data and output it more appropriately for a blind or visually impaired user. Current HTML causes problems because the display tags, such as <H1>…</H1>, are of little use to non-visual displays.

Metadata is data associated with objects which relieves their potential users of having to have full advance knowledge of their existence (Dempsey and Heey,1997).

Metadata serves a number of purposes: it aids resource discovery (i.e. establishing the existence of an information object that may fulfill a user' needs), it assists user to evaluate the object, without necessarily having to access the object itself.

Libraries have thus long created metadata in the form of catalogue entries and have generally bought in further metadata, especially in relation to journal papers, in the form of published indexing and abstracting services. In the networked information environment metadata has become more and more important, since it holds the key not only to providing individuals with descriptions for them to browse but more importantly to the use of software to locate relevant information, to negotiate terms for its supply, to request it and to receive it .
It is clear that metadata contains typical and important information, therefore the XML is most suitable to handle now-a –days.

Advantages of XML

(1) Because XML lets one define one's own markup language, one can make full use of the extended hypertext features of XML to store or link of metadata in any format(eg ISO

11170,Dublin core, Warwick Framework, Resource Description Framework(RDF) and Platform for Internet Content Selection(PICS).

(2) There are no pre defined elements in XML, because it is an architecture, not an application, so it is not part of XML'S job to specify how or if authors should or should not implement metadata. One is free to use any suitable method from simple attributes to the embedding of entire Dublin Core/Warwick Framework metadata records.

(3) Any programming language can be used to output data from any source in XML format. There is a growing number of front-ends and back-ends for programming environments and data management environments to automate this.

(4) Since XML is not a programming language, so XML files don't 'run ' or execute 'execute'. XML is a markup specification language and XML files are data: just sit there until you run a program which displays them (like a browser) or does some work with them (like a converter which writes the data in another format, or a database which reads the data) or modifies them (like an editor).

(5) XML, is more beneficial for application-to-application information exchange rather than for application-to-individual information exchange.

Validity of XML Document :

A valid XML document describes -

(1) The structural rules that the markup attempts to follow

(2) Lists any external resources (external entities) that are part of the document

(3) Declare any internal resources (internal entities) that are used within the document

(4) Lists the types of non-XML resources(notations) used and identifies any helper applications that might be needed.

(5) Lists any non-XML resources (binaries) that are used within the document and identifies any helper applications that might be needed.

6 Conclusion

XML takes the best of the SGML family and combines it with some of the best features of the HTML, and adds a few features drawn from some of the more successful applications of both. XML takes its major framework from SGML, leaving out everything that isn't absolutely necessary. Each facility and feature was examined. XML is commonly called a subset of SGML, but in technical terms it's an application profile of SGML: Whereas HTML uses SGML and is an application of SGML, XML is just SGML on a smaller scale.

7 References

1. BROTHY (PETER). The library in the twenty-first century: new services for the information age. 2001. Library Association, London.
2. NORTH (SIMON) and HERMANS (PAUL). Teach yourself XML in twenty one days.1999. Techmedia, New Delhi.
3. SATYANARAYANA (B). Internet : its genesis, growth and benefits. Proceedings of Caliber 1997. p 124-127.
4. STEPHEN MACK (E) and PLATT (JANAN). HTML 4.0 No experience required. 1999. BPB Publications, New Delhi.
5. http://www.oasis-open.org/cover/general.html