8. Markup Languages

Markup language is the combination of text and information (about text s structure and presentation style) about the text.

A language that has codes for indicating layout and styling of a document (such as boldface, italics, paragraphs, placement of graphics, etc.) within a text file is called a Markup language. Widely used markup languages include SGML (Standard General Markup Language) and HTML (Hypertext Markup Language.

In other words, when we format text to be printed (or displayed on a computer), we need to distinguish between the text itself and the instructions for printing the text. The markup is the instructions for the text.

Markup can also indicate information about the text. For example, many students in school highlight certain phrases in their text books. This indicates that the highlighted text is more important than the surrounding text. The highlight color is markup.

Markup

Text added to the data of a document to convey information about the document; for example: tags, processing instructions, and hyperlinks.

Markup language-

- *A text-formatting language designed to transform raw text into structured documents, by inserting procedural and descriptive markup into the raw text .*
- *A language designed to describe or transform in space or time data, text, or objects into structured data, text, or objects, for example: SGML, HTML, and VRML.*

History

The concept of Markup was introduced by William W. Tunnicliffe, in 1967, as Generic Code but working implementation was actually done in 1980 by Brian Reid. However, Charles Goldfarb is known as father of Markup Languages due to his contributions in development of IBM GML and SGML. SGML is first widely used Markup System developed by International Organization for Standardization committee. Todays XML, a W3C recommondation from 1998, is subset of SGML.

Characteristics of Computer-Use Markup Languages

The important characteristics of a Markup language are:

- It should be easy to recognize, and to distinguish between data and markup. It should be easy to look at a piece of data and say this is data or this is markup .
- There should be a minimum of variants in data formats.

  Each character or other piece of data, and each mark, should have, ideally, one encoding. As well, similar things should be encoded similarly, reducing the number of different cases computer programs have to deal with even when there are a variety of different things being encoded.
- Markup and character recognition should be context-free.

  This is another take on minimizing variations. For data, this means that the same character, for example, should be encoded the same way no matter where it occurs.
- There should be nothing in the data and markup that is not required for the transmission.

  Most importantly, this means that computer-to-computer data should not have extra white space or comments within it neither serves the needs of the computer.
- Binary encoding is preferable, but not necessary.

  There is no need for a computer-to-computer transmission to be clear text be directly readable by human beings. Where such reading is required, some kind of computer-to-human translation of the data can usually be done without a lot of difficulty.
- Redundancy is a mixed blessing.

  Duplicated information is often just a waste of time. Data is serialized for transmission.

Classes of markup languages

- **Presentational markup**

  Presentational markup try to guess documents structure from cues in the encoding. For example, in a text file, the title of a document might be preceded by several newlines and/or spaces, thus suggesting leading spacing and centering. Word-processing and desktop publishing products sometimes use this conventions.

- **Procedural markup**

  Procedural markup is also focused on the presentation of text, but is visible to the user editing the text file, and is expected to be interpreted by software in the order in which it appears. For example- To format a title, a succession of formatting directives would be inserted into the file immediately before the titles text, instructing software to switch into centered display mode, then enlarge and embolden the typeface. The title text would be followed by directives to reverse these effects; in more advanced systems macros or a stack model make this less tedious.

- **Descriptive Markup**

  It labels the fragments of text without necessarily mandating any particular display or other processing semantics. The fragments of text are labeled as to what they are as opposed to how they should be displayed , software may be written to process these fragments in useful ways not anticipated by the designers of the languages. It is also called as Generic Markup. Descriptive markup systems structure documents into trees. Thus document can be treated as databases. Descriptive markup also facilitates the simpler task of reformatting a document as needed, because the format specification is not intertwined with the content.

Different Types Of Markup Languages

1. General purpose markup languages

   - **GML**- Generalized Markup Language marked up the text document with tags to define paragraph, list, tables, headers(to define important and less important sections). It is used for IBM text formatter, SCRIPT which is the main component of IBM s Document Composition Facility (DCF). It frees document creators from specific document formatting concerns such as font specification, line spacing, and page layout required by Script.

   - **SGML**- Standard Generalized Markup Language is a meta language, which is used to define markup languages for documents. Such as by changing the SGML Declaration one does not even need to use angle brackets . SGML is a descendant of GML and predecessor of XML.

   - **XML** Extensible Markup Language is a simplified subset of SGML, used for creating special-purpose markup languages, capable of describing many different kinds of data, provide sharing of data across different systems/ Internet. XML based languages (RDF/XML, RSS, MathML, XHTML, SVG) are defined in a formal way, allowing programs to modify and validate documents in these languages without prior knowledge of their form.

   - XML describes data and to focus on what data is whereas HTML designed to display data and to focus on how data looks. XML tags are not predefined like in HTML; you must have to define your own tags. XML describe a tree based structure to information in simple text. The fundamental unit in XML is the character, as defined by the Universal Character Set. Characters are combined in certain allowable combinations to form an XML document. The document consists of one or more entities, each of which is typically some portion of the documents characters, encoded as a series of bits and stored in a text file.

2. Document markup languages

- HTML- HyperText Markup Language (HTML) is a markup language designed for the creation of web pages with hyper text and other information to be displayed in a web browser. HTML describes certain text as headings, paragraphs, table, lists and so on and also describe some degree of appearance and semantics of a document. It is Inventated by Tim Berners-Lee and further developed by the IETF with a simplified SGML syntax, HTML is now an international standard (ISO/IEC 15445:2000).

- **MathML**- Mathematical Markup Language represent mathematical symbols and formulae for integrating them into Web documents. MathML provides *information about presentation* and the *meaning of formula components*. Because the meaning of the equation is preserved separate from the presentation, how the content is communicated can be left up to the user. It is an application of XML and also recommended by W3C.

- **WML** Wireless Markup Language implement the WAP (Wireless Application Protocol) specification and is the primary content format for devices. WAP gateway acesses WML pages from a web server and translates WML pages into a format which is wel-suited for mobiles. It is situated between mobile devices and the World Wide Web and passing pages from one to the other much like a proxy. Wireless Markup Language is a not like HTML in that it provides navigational support, data input, hyperlinks, text and image presentation, and forms. A WML document is known as a deck . Data in the deck is structured into one or more cards (pages) each of which represents a single interaction with the user. WML is based on XML.

- **XHTML** The Extensible HyperText Markup Language has the same expressive possibilities as HTML. HTML and XHTML, both are subset of SGML but HTML is very felxible in nature where as XHTML is more restrictive. XHTML documents allow for automated processing to be performed using a standard XML library unlike HTML, which requires a relatively complex, lenient, and generally custom parser (though an SGML parser library could possibly be used). XHTML can be thought of as the intersection of HTML and XML in many respects, since it is a reformulation of HTML in XML. XHTML is an application of XML and W3C Recommendation.

- **XHTML Basic**- XHTML Basic is a subset of XHTML, including a minimal set of XHTML modules for document structure, images, forms, basic tables, and object support. XHTML Basic is suitable for mobile phones, PDAs, pagers, and settop boxes. It will replace WML and C-HTML because of One large advantage XHTML Basic has over WML and C-HTML is that XHTML Basic pages can be rendered differently in web browsers and on handhelds, without the need for two different versions of the same page.

3. Content syndication markup languages

- **RSS**- RSS is a lightweight XML format designed for sharing headlines and other Web content. It is a kind of mini database containing headlines and descriptions of whats new on your site, is a natural for layering on additional services. In addition to displaying your news on other sites and headline viewers, RSS data can flow into other products and services like PDAs, cell phones, email ticklers and even voice updates. The RSS is variously used to refer to the following standards:

- Rich Site Summary (RSS 0.91)
- RDF Site Summary (RSS 0.9 and 1.0)
- Really Simple Syndication (RSS 2.0)

Each RSS text file contains both static information about your site, plus dynamic information about your new stories, all surrounded by matching start and end tags. RSS in particular, delivers this information as an XML file called an RSS feed, webfeed, RSS stream, or RSS channel. Web feeds provide web content or summaries of web content together with links to the full versions of the content, and other metadata. In addition to facilitating syndication, web feeds allow a website s frequent readers to track updates on the site using an aggregator. RSS is popularly used in news websites, weblogs and podcasting.

- **SyncML**- SyncML (Synchronization Markup Language) is a platform-independent open industry standard for universal synchronization of remote data and personal information across multiple networks, platforms, and devices. Because it supports multiple transport protocols (including HTTP, Wireless Session Protocol, OBEX (Bluetooth, IrDA), Simple Mail Transfer Protocol, pure TCP/IP networks,), SyncML easily passes data across the vast array of networks (both wired and wireless) and networked devices.

  SyncML is a method to synchronize contact and calendar information between a handheld device and a computer, but mostly used for remote synchronization of mobile devices. It can work over various types of connections, including Wireless Internet, Bluetooth, and infrared. The new version of the specification includes support for push email, providing a standard protocol alternative to proprietary solution like BlackBerry.

4. Lightweight markup languages

A lightweight markup language has simpler syntax and used in applications where there is relatively little bandwidth and so conciseness is important.

- **Markdown**- Markdown provides readability and publishability of both its input and output forms, taking many cues from existing conventions for marking up plain text in email. Markdown converts its marked-up text input to valid, well-formed XHTML and replaces left-pointing angle brackets (<) and ampersands with their corresponding character entity references. It was originally implemented in Perl but now being used in many of programming languages including PHP, Python, Ruby and Java.

- Simple Declarative Language - The Simple Declarative Language is a cross-platform language used for defining basic data structures such as lists, maps, and trees of typed data in a compact, easy to read representation. A simple API allows one to read, write and access all the data structures using a single class. For property files, configuration files, logs and simple serialization requirements. SDL is designed to be an alternative to XML and properties files.

5. User interface markup languages

An user interface markup language is used to define user interfaces.

- **HTML**- (already described)
- **MXML** - MXML is an XML markup language used mainly to declaratively lay-out the interface of applications, and can also be used in conjunction with ActionScript to allow developers to implement complex business logic. Common practices are employed, such as the use of curly braces ({) to force the computer to evaluate an expression, and dot notation to drill-down through an object.
- **XForms** - XForms is an XML format for the specification of user interfaces, specifically web forms. XForms was designed to be the next generation of HTML / XHTML forms, but is generic enough that it can also be used in a standalone manner to describe any user interface, and even perform simple and common data manipulation tasks.

- **Macromedia_Flex** - Flex allow Web application developers to quickly and easily build Rich Internet Applications. Macromedia Flex is an application server. It is a J2EE application or JSP tag library that compiles Flex Mark-Up Language (MXML)

6. Vector graphics markup languages

A vector graphics markup language describes an image in terms of lines, curves, and other vector graphics primitives at a higher level than a bitmap.

The list of vector graphics markup languages includes-

- **Scalable Vector Graphics**- Scalable Vector Graphics (SVG) is an XML markup language for describing two-dimensional vector graphics, both static and animated (either declarative or scripted). It is an open standard created by the World Wide Web Consortium.
- **Encapsulated PostScript**- EPS file contains description of the rectangle which contain information about the image. Applications can use this information to lay out the page, even if they are unable to directly render the PostScript inside.
- **VRML**- VRML (Virtual Reality Modeling Language- designed by W3C) is a standard file format for representing 3-dimensional (3D) interactive vector graphics. VRML is a text file format where vertices and edges for a 3D polygon can be specified along with the surface color, image-mapped textures, shininess, transparency, and so on. URLs can be associated with graphical components so that a web browser might fetch a web-page or a new VRML file from the Internet when the user clicks on the specific graphical component. Animations, sounds, lighting, and other aspects of the virtual world can interact with the user or may be triggered by external events such as timers. A special Script Node allows the addition of program code (e.g., written in Java or JavaScript (ECMAScript)) to a VRML file.

2D vector graphics-

2D computer graphics is the computer-based generation of digital images mostly from two-dimensional models (such as 2D geometric models, text, and digital images) .

- **SVG**- Scalable Vector Graphics is a markup language for graphics proposed by the W3C that can support rich, graphical user interface for web and mobile applications. SVG is not a user interface language, it is a standard that includes support for vector/raster graphics, animation, interaction with the DOM and CSS, embedded media, events and scriptability. When these features are used in combination, rich user interfaces are possible.
- **XAML** - XAML is not just an XML-based user interface markup language, but an application markup language, as the program logic and styles are also embedded in the XAML document. Functionally, it can be seen as a combination of XUL, SVG, CSS, and JavaScript into a single XML schema. Some people are critical of this design, as many standards (such as those already listed) exist for doing these things. However, it is expected to be developed with a visual tool where developers do not even need to understand the underlying markups.

3D vector graphics

- 3D computer graphics represent a geometric data stored in the computer for the purposes of performing calculations and rendering 2D images.
- **3DXML**- 3DXML is an XML schema developed to share 3D data between users.
- **IPA**- The International Phonetic Alphabet (IPA) is a system of phonetic notation devised by linguists to accurately and uniquely represent each of the wide variety of sounds (phones or phonemes) used in spoken human language. It is intended as a notational standard for the phonemic and phonetic representation of all spoken languages.

7. Web service markup languages

- **SOAP** - SOAP is a protocol for exchanging XML-based messages over a computer network, normally using HTTP. There are several different types of messaging patterns in SOAP, but by far the most common is the Remote Procedure Call (RPC) pattern, where one network node (the *client*) sends a request message to another node (the *server*), and the server immediately sends a response message to the client. Indeed, SOAP is the successor of XML RPC.

- **UDDI** - UDDI (Universal Description, Discovery, and Integration) A platform-independent, XML-based registry for businesses worldwide to list themselves on the Internet. UDDI is an open industry initiative enabling businesses to publish service listings and discover each other and define how the services or software applications interact over the Internet.

- **WSDL** - The Web Services Description Language (WSDL) is an XML format published for describing Web services.

- **XML-RPC** -XML-RPC is a remote procedure call protocol which uses XML to encode its calls and HTTP as a transport mechanism. It is a very simple protocol, defining only a handful of data types and commands, and the entire description can be printed on two pages of paper. This is in stark contrast to most RPC systems, where the standards documents often run into the thousands of pages and require considerable software support in order to be used.

8. Unclassified:

- **SMIL** (Synchronized Multimedia Integration Language) **SMIL** (**Synchronized Multimedia Integration Language**- describes multimedia presentations using XML (Extensible Markup Language). It defines timing markup, layout markup, animations, visual transitions, and media embedding, among other things. A SMIL document is similar in structure to an HTML document in that they are typically divided between a <head> section and a <body> section. The <head> section contains layout and metadata information. The <body> section contains the timing information, and is generally comprised of combinations of two main tags: parallel (<par>) and sequential (<seq>). SMIL refers to media objects by URLs, allowing them to be shared between presentations and stored on different servers for load balancing. The language can also associate different media objects with different bandwidths.

- **VoiceXML** VoiceXML (VXML- W3C s standard) is a XML format for specifying interactive voice dialogues between a human and a computer. It is fully analogous to HTML, and brings the same advantages of web application development and deployment to voice applications that HTML brings to visual applications. Just as HTML documents are interpreted by a visual web browser, VoiceXML documents are interpreted by a voice browser.

  VoiceXML has tags that instruct the voice browser to provide speech synthesis, automatic speech recognition, dialog management, and soundfile playback.

References:

www.w3.org