1 Basic

“Standard Generalized Markup Language” (SGML) is a language for defining markup languages. SGML is defined by the International Standards Organization Document ISO 8879 [1986].

The ISO document is not freely available. A copy of it is found in the book:

“Hypertext Markup Language” (HTML), the basic language of the World Wide Web, is a markup language under SGML.

“Extensible Markup Language” (XML) is a limited form of SGML that is currently under heavy promotion by the World Wide Web Consortium (W3C). It is sometimes perceived as “extended HTML”. XML has been designed to be usable on browsing platforms, while full-fledged SGML is usually more suitable for authoring platforms. In fact, XML has for most purposes become the only form of SGML that is suitable for public sharing. Many SGML languages\(^1\) that are realistically suitable for authors admit rapid automatic translation to nearly equivalent XML languages. (Note that an XML language need not contain the HTML tag set nor have any relation to HTML, and HTML is not an XML language although it may be automatically converted to a language under XML.)

2 Classical HTML is not an XML Language

Classical HTML refers to the markup language behind World Wide Web locations from the beginning of the Web at CERN, Geneva, until very recently. The versions of W3C HTML numbered from 2.0 through 4.01 are all languages under SGML that do not fall within XML.

Three simple reasons why HTML is not an XML language are:

1. In HTML most paragraphs are marked up using an open tag "<P>" at the beginning of the paragraph without needing a close tag "</P>" at the end, while there must be a close tag for every open tag in XML.

2. In HTML tag names are not case-sensitive, while in XML tag names are case-sensitive. (A new standard way of converting HTML into an XML language will specify that tag names all be lower case.)

3. In HTML some attribute values need not be placed inside quotation marks, while in XML all attribute values must be quoted.

Early in the year 2000 a new evolution of HTML referred to as XHTML\(^2\) — but bearing the formal document type name “html” (lower case characters only) — acquired the status of W3C Recommendation. XHTML, version 1.0, is an XML language that has the same tag set as HTML 4.01\(^3\). Apart from technical details XHTML 1.0 is almost the same language as HTML 4.01. Because of the technical differences, however, a computer does not need the full weight of an SGML processor to interpret XHTML. This advantage is offset by the fact that it is slightly more difficult for authors to create XHTML than to create classical HTML.

3 The Nature of SGML

While SGML may be described as a language for creating markup languages with a shared

\(^1\)The phrase “SGML language” used here, as well as the parallel phrase “XML language”, is formally not correct usage. What is called here an SGML (respectively, XML) language is formally known as an SGML (respectively, XML) application. Every XML application may also be regarded as an SGML application. There is an identifying correspondence between applications in this sense and document types.

\(^2\)URI: http://www.w3.org/TR/2000/REC-xhtml1-20000126

\(^3\)URI: http://www.w3.org/TR/1999/REC-html401-19991224
syntax, more realistically and more abstractly, an SGML language (formally, an SGML application) is a template for processing. For this reason when an SGML document (formally, an SGML instance) is written, the author is, in fact, setting its text as organized data.

The abstract character of languages under the SGML umbrella makes it possible to use the family to describe computer programs. The Extensible Style Language (XSL) described below is an example of such an SGML application that is, in fact, an XML application.

4 Styling and Translating XML documents

In principle, an author may create a personal XML language. To do so the author must be prepared to provide, in addition, (1) companion “style sheets” or (2) companion translators.

If one uses a language under XML or SGML, one must understand what companion style sheets or translators will be used with that language.

A style sheet is a document that is created to provide directions for a processing program, perhaps a printing formatter or a web browser, on the formatting or rendering of a document that is prepared in a markup language.

While a translator may be any program, typically a translator is a package of small programs (sometimes called functions) for processing a document under an XML language to some other language, which might be \( \TeX \), HTML, another XML, ... under a general framework for processing XML or SGML. There are free frameworks for writing such programs in various languages. Most of these frameworks require pre-processing parsers, and free parsers are also available.

Near-term plans for the development of the World Wide Web anticipate major web browsing programs having the capability to provide finely-tuned rendering of XML documents that are accompanied by a style sheet. Style sheet support for HTML documents is currently available.

Limited rendering of XML documents on the World Wide Web is based on “Cascading Style Sheets” (CSS), which has been in use for customized rendering guidance with HTML browsing programs.

A future standard style language for XML documents in World Wide Web browsing programs is called “Extensible Style Language” (XSL). XSL is a restricted form of “Document Style Semantics and Specification Language” (DSSSL) that is written with XML syntax. The specification for XSL was still under draft at W3C on March 1, 2000, while a variant called “XSL Transformation Language” (XSLT), which is now widely used for defining translations of XML languages to other languages (whether XML or not), became a W3C recommendation in late 1999.

While XSLT-based translation offers more precision than is available with CSS-guided formatting, in the overall world of XML processing one should expect formatting based on either CSS or XSL style sheets to be a limited type of formatting. One should expect to obtain the finest typesetting results by going beyond the narrow class of XML translation programs that admit expression in a style sheet language.

\(^4\)URI: http://www.w3.org/Style/CSS/
\(^5\)URI: http://www.w3.org/Style/XSL/
\(^6\)URI: http://www.jclark.com/dsssl/
\(^7\)URI: http://www.w3.org/TR/xslt
A relatively new simple example of SGML processing may be found in the system manual under SunOS, version 5.7. Observant users of University at Albany SunStations may have noticed that as of the summer of 1999 most of the system manual in the central "/usr/man" area now exists in source form under an SGML document type for the manual rather than, as formerly, in the nroff typesetting language. (This is temporarily hampering the operation of the classical X11 program xman for the affected portions of the system manual; text rendering is not affected.) See the manual page for "solbook" and browse "/usr/lib/sml" and "/usr/share/lib/sml" for more information.

A document created carefully today under a well designed XML or SGML language should admit automatic conversion to future formats once an SGML or XML translator for such conversion has been created.

5 Example Languages under XML and SGML

1. CALS is a language under SGML that is widely used in the U.S. Department of Defense.

2. “DocBook” is a public language under SGML that may be used by authors. A fall 1999 book, Norman Walsh, DocBook: The Definitive Guide is available online and in bookstores. Walsh maintains a web site http://nwalsh.com/ with a great deal of information about related topics, including an excellent tutorial on XSL. (Campus UNIX Network only: A copy of the DocBook DTD is available on the local network.)

3. The TEI Consortium has emerged from the Text Encoding Initiative at The University of Illinois at Chicago as custodian of the TEI language definition. TEI is another public language that may be used by authors. Its modular design has led to the creation of the TEI Pizza Chef web site at Oxford.

A copy of the current TEI Guidelines in HTML, which includes A Gentle Introduction to XML is available for local browsing on the Sun network from the file system location /usr/share/local/xml/tei/P4X/index.html.

4. HTML is a language under SGML.

5. XHTML (formerly HTML-Voyager) is a language under XML, recommended by the World Wide Web Consortium (W3C), that is designed to be equivalent to HTML. It is intended to be the base for extending HTML to a language under XML. See:

http://www.w3.org/TR/xhtml1/.

6. MathML, Mathematical Markup Language is a client platform language under XML that is intended to add mathematical functionality to the world wide web. See:

http://www.w3.org/TR/mml1/.

---

8URI: http://www.oasis-open.org/docbook/
9URI: http://www.docbook.org/tgd/html/
10URI: file:///usr/share/local/xml/docbook/dtd/
11URI: http://www.tei-c.org/
12URI: http://www.uic.edu/orgs/tei/
13URI: http://www.hcu.ox.ac.uk/TEI/newpizza.html
14URI: http://www.tei-c.org/P4X/
15URI: http://www.tei-c.org/P4X/SG.html
16URI: file:///usr/share/local/xml/tei/P4X/index.html
http://www.w3.org/Math/.

The W3C Recommendation for MathML, version 2, points to a document type definition at W3C for the implementation of a namespace-based extension of XHTML that includes MathML.

7. The W3C working draft on Scalable Vector Graphics (SVG) format proposes an XML language for online graphics. This draft may be found along with other related information at

http://www.w3.org/Graphics/SVG/.

8. XGMML\(^\text{18}\), eXtensible Graph Markup and Modeling Language, developed recently in New York’s Capital District at RPI\(^\text{19}\), is an XML application based on GML which is used for graph description. See also

http://xml.coverpages.org/xgmml.html.

9. Any programming assembly language in which each line consists of an operation code followed by parameters is equivalent to an XML language.

10. The device independent typesetting file format (DVI) associated with the typesetting language \(\text{TeX}\) (and with the program groff) is equivalent to an XML language.

6 References

The World Wide Web Consortium is the driving force behind XML. See:

http://www.w3.org/XML/.

A 1998 book on XML is:

Charles F. Goldfarb and Paul Prescod,

A very comprehensive catalogue of information about SGML and XML may be found on the web at

http://xml.coverpages.org/.

An interesting and useful web site with ties to Sun MicroSystems, one of the principal sponsors of XML, is

http://metalab.unc.edu/xml/.

\(^{17}\) URI: http://www.w3.org/TR/REC-xml-names/

\(^{18}\) URI: http://www.cs.rpi.edu/~puninj/XGMML/

\(^{19}\) URI: http://www.rpi.org/
An early survey *Comparison of SGML and XML* is available from W3C.

Monitoring the UseNet newsgroups `news:comp.text.sgml` and `news:comp.text.xml` is an excellent way to have a window on current discussion.

One may also seek answers to questions in the newsgroups when the answers cannot be obtained locally through the HelpDesk at `mailto:helpdesk@csc.albany.edu`. However, one should first make sure that the question is appropriate to the specific topic of the newsgroup. For example, most questions about creating web pages do not belong in these two newsgroups.

Information about the topic of “mathematics and SGML” may be found at the (local) URL

```
http://math.albany.edu:8800/hm/sgml/about.html.
```

## 7 Software Available Locally

The University at Albany UNIX Network has several basic, general purpose, freely available tools for working with SGML and XML including:

1. The “open source” evolute, called "onsgml", of James Clark’s SGML parser "nsgmls", which is an application under the OpenSP C++ library.

   The public location for OpenSP is the OpenJade Project\(^\text{21}\) at SourceForge. The public location for information about SP is:

   ```
   ```

   Note: "onsgmls", when properly called, may be used to check the structural correctness of an HTML document. At the University at Albany the command "validhtml" is an interface to "onsgmls" for this method of HTML validation.

2. The well-known XSLT engine `xsltproc`, which is included in the University’s present version of the Sun Microsystems operating system.

3. Script interfaces to various Java-based tools of James Clark for handling XML including:

   - `dtdinst` a utility to generate an XML instance that models an XML document type definition given in DTD form.
   - `jcxt` the engine called “xt” for transformations specified in the XSLT language.
   - `jing` a utility to validate an XML instance against a document type definition specified in the form of either a Relax-NG schema\(^\text{22}\) or a W3C schema\(^\text{23}\).
   - `trang` a utility for translations between various types of XML document type definitions

4. David Megginson’s general purpose SGML-to-anything processor, "sgmlspl", which is an application under his Perl-5 library "SGMLSPM".

   Local documentation on "SGMLSPM/sgmlspl" may be found at:
The public location for information about "SGMLSPM/sgmlspl" for many years was

http://home.sprynet.com/sprynet/dmeggin/. 

That appears to have been superseded by

http://www.megginson.com/Software/;

and SGMLSPM/sgmlspl is also available at CPAN\textsuperscript{24}.

8 Miscellaneous

8.1 XML and Electronic Data Interchange (EDI)

XML offers a standard framework for the general interchange of many kinds of data. The usefulness of XML-EDI lies in the inherent adaptability to this end of the many new tools for handling XML. There is a substantial amount of material on this topic in the book by Goldfarb and Prescod cited above. See the web site:

http://www.geocities.com/WallStreet/Floor/5815/.

The World Wide Web Consortium (W3C) has basic information about how one might proceed to model a database in XML at the site:

http://www.w3.org/XML/.

8.2 Library Metadata

The Open Archives Initiative (http://www.openarchives.org/) has developed a protocol for interoperable handling of library metadata across the network based on records prepared under special purpose XML document types that are defined using the new notion of XML schema\textsuperscript{25}.

8.3 How This Document Was Prepared

This document was prepared in Generalized Extensible \LaTeX-like Markup (GELLMU), which is the author’s user markup interface for SGML languages. Presently the system, still under development, may be used to create both standard \LaTeX\textsuperscript{26} and HTML\textsuperscript{27} versions from a single \LaTeX-like source\textsuperscript{28}, a text file. The program \texttt{latex} may be used to prepare a high quality typeset
version\textsuperscript{29} in DVI format\textsuperscript{30} suitable for printing on this system using the program \texttt{dvips}, and a variant of \texttt{latex} known as \texttt{pdflatex} may be used to prepare a different typeset version\textsuperscript{31} in PDF format and an alternate form of processing to HTML will produce XHTML\textsuperscript{32} extended by MathML. For more information on this system see

\url{http://www.albany.edu/~hammond/gellmu}.

\footnotesize
\textsuperscript{29}URI: general.dvi
\textsuperscript{30}Donald Knuth's Device Independent Format (DVI)
\textsuperscript{31}URI: general.pdf
\textsuperscript{32}URI: general.xhtml