An Information Provider's Guide to HTML

Nathan Torkington

September 16, 1993

Introduction

This document is an introduction to the HyperText Markup Language (HTML). It is intended to be a gentle primer for information providers who want to know the background, purpose and functionality of the language. It is not intended to be a first introduction to the Web, nor a definitive guide to the markup language.

This document is available on the Web, as well as being posted fortnightly to the Usenet newsgroup **comp.infosystems.www**. It is available as LaTeX, plain ASCII, DVI and Postscript files via anonymous ftp. For instructions on retrieving the latest version of this document, consult the last section, called "How to obtain this document".

This document was last revised on *Wed Sep 15 16:32:24 NZT 1993* by *Nathan Torkington.*

Table of Contents

7. Conversion from HTML
8. Editing HTML
9. What to Avoid
10. See Also
11. How to obtain this document

HTML and SGML

HTML is a way of marking up documents that conforms to the ISO standard 8879: "The Standard Generalized Markup Language (SGML)". It provides a way of encoding document structure with a minimum of presentation information. SGML provides a standard way of describing what the markup looks like.

The description of the markup is called a "Document Type Definition", or DTD. The DTD for HTML is available on the Web as http://info.cern.ch/hypertext/WWW/MarkUp/HTML.dtd.html.

The Appearance of HTML

HTML looks like plain text with tags attached. The tags are enclosed in angle brackets (<...>) and the names of the tags reflect the structure of the document. For instance, there are tags to enclose headings (<H1>This is a heading at level 1</H1>), the title of a document (<TITLE>The Title</TITLE>), lists (<OL> for an Ordered List), and so on.

Despite the efforts to encode only meaning in HTML, authors have requested some tags that define presentation. For instance, <B>text</B> places the word "text" in a bold font, if this is meaningful to the program interpreting the HTML (it might be a useful tag for a browser, but not for an automatic indexer).

The Future of HTML

The HTML language is very nearly locked as a standard. Design has already begun on another language, HTML+, which encodes more structure than HTML. Plans are also underway for style sheets, which give authors the ability to provide specific hints to browsers (along the lines of "12 pt Roman for this", "centre that").

Any future language(s) will not replace HTML in the sense that HTML will no longer be supported. They will exist side-by-side with HTML.

Conversion to HTML

[A complete list will be forthcoming from Rich Brandwein when he gets server access]

Plain ASCII can be turned into HTML by enclosing it in <PRE> ... </PRE> tags. This doesn't take advantage of the structure and presentation capabilities of HTML, however, and is only recommended for "quick and dirty" tasks.

LaTeX documents can be converted to HTML with the program tex2html by Nikos Drakos (nikos@cbl.leeds.ac.uk). See the entry on "tex2html" in the section "See Also" for information on obtaining this program.

RTF documents can be converted to HTML with either the program rtf2html (see the entry on "rtf2html" in the section "See Also" for information on obtaining this program) or ....

setext documents can be converted to HTML with the Perl script setext.pl that is part of the Plexus package (see the entries on "Perl", "setext" and "Plexus" in the section "See Also" for information on obtaining these products).

Conversion from HTML

HTML can be turned into formatted ASCII by the CERN LineMode Browser (see the entry on "LineMode Browser" in the section "See Also" for information on obtaining the source for this program). The emacs browser also has this ability (see the entries on "EMACS" and "w3.el" in the section "See Also" for information on obtaining the source for this program).

HTML can be turned into LaTeX with either the CERN html2latex.sed script (see the entry on "html2latex.sed" in the section "See Also" for information on obtaining this script) or Nathan Torkington's html2latex C program (based on the XMosaic v0.11 parser). See the entry on "html2latex" for information on obtaining the source for this program.

Editing HTML

There are currently no WYSIWYG (What You See Is What You Get) editors for HTML besides the one in TkWWW (see the entry on "TkWWW" in the section "See Also" for information on obtaining TkWWW).

There is EMACS mode for editing HTML, written by Marc Andreeson (marca@ncsa.uiuc.edu). See the entry on "html-mode.el" in the section "See Also" for information on obtaining it.

jeff.grover@gtri.gatech.edu (Jeffrey L. Grover) has written a set of WordPerfect for Windows macros, but they are still in alpha-test and aren't for release yet.

If you have access to an RTF editor, you can edit with your favourite text editor and add in any hypertext links after converting with rtf2html.

I, personally, write with a text editor and add the tags by hand.

What to Avoid

Avoid taking advantage of the way that one browser interprets tags, because other browsers may not display the document in the same way. Attempt to check the way your document appears on many different browsers. Use style-sheets when they become available.

Avoid writing HTML that doesn't conform to the standard. You can verify your HTML document by running the sgmls program over the HTML DTD and your document (see the entry on "sgmls" in the section "See Also" for information on obtaining sgmls).

See Also

EMACS EMACS is the GNU text editor. The source code is available via anonymous FTP from prep.ai.mit.edu — the file FTP in the directory /pub/gnu/GNUinfo explains the arrangements for obtaining copies.

html-mode.html This EMACS major mode for editing HTML is available via anonymous FTP from ftp.ncsa.uiuc.edu in the directory /Web/elisp as html-mode.el.

html2latex Nathan Torkington's conversion program from HTML to LaTeX is available via anonymous FTP from ftp.ncsa.uiuc.edu in the directory /Web/xmosaic-contrib/ as html2latex-XXX.tar.Z, where XXX is a version number.

html2latex.sed Available on the Web as http://info.cern.ch/hypertext/WWW/Tools/TeX/html2latex.sed – see also http://info.cern.ch/hypertext/WWW/Tools/TeX/Makefile and http://info.cern.ch/hypertext/WWW/Tools/TeX/sub1.sed.

LineMode Browser The PC-NFS version is available via anonymous FTP from info.cern.ch in the directory /pub/www/bin/pc-nfs/wwwpcnfs.zip. The source-code (which compiles under Unix and VMS) is available via anonymous FTP from info.cern.ch in the directory /pub/www/src/ as WWWLineMode_XXX.tar.Z, where XXX is a version number.

Perl Perl is an interpreted language, especially good for text handling. It is available for anonymous FTP from ftp.uu.net in the directory /pub/languages/perl/ as perl.tar.gz.

Plexus Plexus is [where?].

rtf2html The source code is available via anonymous FTP from oac.hsc.uth.tmc.edu in the directory /public/unix/WWW/ as rtf2html.tar.

setext setext stands for Structure Enhanced Text, and is a markup system that provides a way to format ASCII documents with visually unobtrusive anchors to parts of it above the paragraph level. More information is available via anonymous FTP from garbo.uwasa.fi in the directory /mac/tidbits/setext/

sgmls sgmls is a validating SGML parser. It is available via anonymous FTP from ftp.th-darmstadt.de in the directory /pub/text/sgml/sgmls/.

tex2html The source to this is available on the Web as http://cbl.leeds.ac.uk/nikos/tex2html/latex2html.tar, and documentation is available on the Web as http://cbl.leeds.ac.uk/nikos/tex2html/doc/latex2html/latex2html.html

TkWWW TkWWW is available via anonymous FTP from any X11 site in the contrib/ directory — TkWWW uses the tcl/tk language and graphics libraries.

w3.el William M. Perry's World Wide Web browser for EMACS is available via anonymous FTP from cs.indiana.edu in the directory /pub/elisp/w3/. The author's e-mail address is wmperry@cs.indiana.edu.

How to obtain this document

The latest version of this document is always available on the Web as **http://www.vuw.ac.nz/non-local/gnat/www-html.html**, and the most recently posted ASCII version will be available via anonymous FTP from rtfm.mit.edu in the directory /pub/usenet/news.answers/www as **html-guide**. The ASCII, LaTeX, DVI, and PostScript versions will be available via anonymous FTP from wuarchive.wustl.edu in the directory /doc/misc/www/.

This document is part of a series: "World Wide Web Primer", "An Information Provider's Guide to HTML", and "An Information Provider's Guide to Web Servers". The other documents in the series are available from the archives above.

Please send feedback to the author, Nathan Torkington, at the e-mail address Nathan.Torkington@vuw.ac.nz — all discussion will be treated as public domain and may be used in future versions of this document.