# Text to Hypertext Conversion with LaTeX2HTML *

Nikos Drakos,
Computer Based Learning Unit,
University of Leeds,
Leeds LS2 9JT, UK.

phone: 0532 - 334626
fax: 0532 - 334635
email: nikos@cbl.leeds.ac.uk
WWW: http://cbl.leeds.ac.uk/nikos/personal.html

April 1, 1994

**Abstract**

LaTeX2HTML is a conversion tool that allows existing documents written in LaTeX to become part of a global multimedia system. This paper presents some of the reasons for using such a system and describes the basic conversion process.

*This is an updated version of a paper which appeared in Baskerville [4]

# 1 World Wide Web - A Global Multimedia System

> Imagine a system that links all the text, data, digital sounds, graphics and video on all the world's computers into a single interlinked hypermedia "web". This is the potential of the Internet-based World Wide Web (WWW or W3) project ... [2]

The World Wide Web merges hypermedia techniques with networked document retrieval to provide a global information system of linked documents. These are traversed by "clicking" in textual or iconic active areas, or searched via query mechanisms [1]. Hypertext links may point to a different location in the same document or to another document which may be located perhaps in another continent!

Documents are not limited to containing only textual information and may include high resolution images, audio and video samples. WWW also encompasses most of the services currently available on the Internet such as Usenet news, ftp, wais, archie, etc. Access to these services as well as the invocation of arbitrary computer programs (e.g. a database access or a simulation) is completely transparent to the user who sees them all as part of some document and interacts with them in a uniform and intuitive way.

Multimedia documents are written in a language designed specifically for the World Wide Web called HTML (HyperText Markup Language) which is based on SGML (Structured Generalised Markup Language). Documents are written by information providers who just place them on the WWW using a "server" program. Then anyone with access to the Internet can use a "client" or "browser" program to access and view available documents. Clients and servers communicate via the HTTP protocol (HyperText Transfer Protocol). Apart from navigation facilities, browsers also allow full text searches, "cut and paste", text or audio annotations, personal "hotlists", saving and printing in multiple formats and others. Such browser and server programs are freely available for most popular computer configurations.

With the explosive growth of the World Wide Web (500-fold since the first graphical browsers were made available this year [3]), and a potential audience of 15 million in more than 50 countries, providing information via the WWW is becoming an extremely attractive proposition.

# 2 LaTeX to HTML Conversion: Why?

HTML is quite a simple markup language to learn and use. It allows basic formatting commands, bulleted lists, "inlined" images, and hypertext links to other documents, multimedia sources, internet services or computer programs. But despite (and because of) its simplicity it has created a few headaches for information providers:

- there are no intuitive authoring tools (yet);
- yet another hypertext language has to be learned;
- existing documents available in other formats have to be reprocessed;
- hypertext document "webs" are difficult to maintain;
- it is difficult or impossible to create highly formatted documents in HTML.

A flexible text to hypertext conversion tool can help in addressing these problems. The authoring problem simply disappears, existing documents can be reused immediately and a complex web of interlinked documents can be generated from a single source document. The automatic inclusion of formatted information such as tables or mathematical equations as inlined images also bypasses another serious problem with HTML. An additional benefit is that the paper-based version of a document can also be obtained from the same source.

The utility of a conversion tool like LaTeX2`HTML` can be seen from the variety of contexts in which it has been applied. Some examples are listed below.

- Electronic books (such as that being produced by the Computational Science Education Project[1] sponsored by the US Department of Energy and involving 35 authors or the CRS4 Active Books Library[2] in Italy from which it is possible to interact with remote programs).
- Scientific papers such as those on the MIT Transit Project[3] or this paper[4]!
- Lecture Notes, Supporting Documentation and Coursework[5]
- Online training material[6]
- General documents such as one recommending the use of LaTeX2`HTML` for the electronic submission of manuscripts to an IEEE journal[7].
- System Documentation[8] and User Manuals[9].

# 3 LaTeX to HTML conversion: How?

The basic conversion process relies on the ability to distinguish between the *structure*, the *content* and the *formatting* information in a LaTeX document.

On the basis of sectioning information, a document is broken into separate parts and an iconic navigation mechanism is constructed in HTML which reflects this structure and allows a user to "jump" between different parts. The cross-references, citations, footnotes, the table of contents and the lists of figures and tables are also translated into hypertext links. Formatting information which has equivalent "tags" in HTML (lists, quotes, paragraph breaks, type styles, etc.) is also converted appropriately.

Although in most cases the loss of some formatting information (e.g. page margins or line widths) is harmless, there are occasions where the format has meaning e.g. when dealing with tables or user defined environments. Another problem is the replication of the mathematical equations which must retain both their precise format as well as any of the predefined special mathematical symbols.

The innovative solution in such cases relies on the ability of HTML browsers to display inlined images inside the main text. Any part of a LaTeX document for which it is not obvious how it should be translated directly into HTML is

[1] http://compsci.cas.vanderbilt.edu/csep.html
[2] http://www.crs4.it/HTML/int_book/meta_page.html
[3] http://www.ai.mit.edu/projects/transit/tn-cat.html
[4] http://cbl.leeds.ac.uk/nikos/doc/maps/maps.html
[5] http://www.cm.cf.ac.uk/lecture_notes.html
[6] http://www.strath.ac.uk/CC/Courses/OnlineTraining.html
[7] http://www.research.att.com/esubmit/esubmit.html
[8] http://www.cwi.nl/cwi/people/Guido.van.Rossum/python-tut/tut.html
[9] http://olt.et.tudelft.nl/usr1/patrick/public_html/docs/wwman/wwman.html

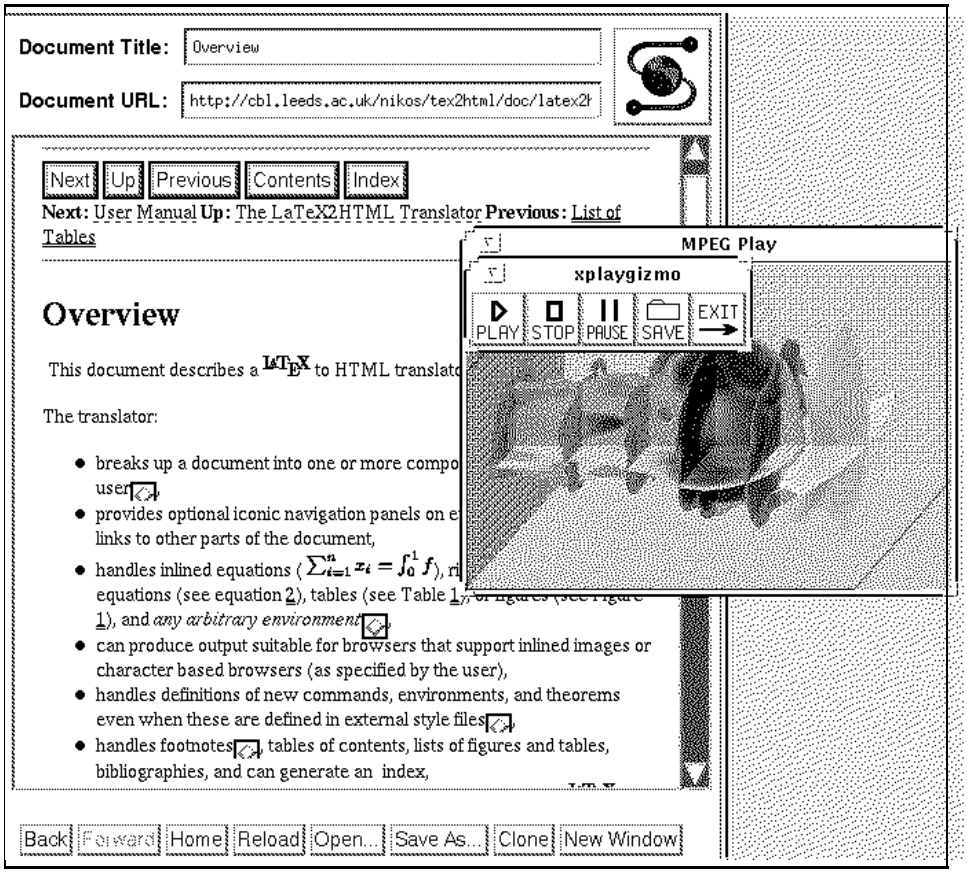


Figure 1: A converted document displayed using Mosaic with an external movie player triggered from it.

extracted from the main document and then placed on a pipeline (from latex to dvi, then postscript, then PPM and finally to GIF or XBM) which converts it into an image. Each image is then placed at the correct position in the final HTML document. Special care is taken to preserve contextual information that may affect the contents of each image (counter values, labels, references, active style files etc). An example of a converted document can be seen in Figure 1.

# 4 Hypermedia Extensions to LaTeX

Apart from the obvious hypertext links within a LaTeX document (e.g. navigation between sections, cross-references and citations) it is also possible to take full advantage of the HTML links to arbitrary multimedia sources (e.g. audio or video), electronic forms, and other remote documents or internet services.

This can be done with some new commands defined in a separate style file (`html.sty`) which are processed in a special way by the LaTeX2`HTML` translator. This style file defines commands for embedding external hypertext links, for extending the basic `\ref-\label` mechanism to operate between remote documents, and specifying that some text should only appear in the paper-based version or only in the HTML document. In most cases these commands have no effect when processed in the conventional way.

Another command allows the inclusion of arbitrary HTML markup directly in a LaTeX document. This can be used to take advantage of new HTML facilities

as soon as they become available (HTML is currently evolving towards a new specification called HTML+). A particularly good use of this feature is in the creation of interactive electronic forms from within a LaTeX document.

# 5 Concluding Remarks

Conversion tools like LaTeX2HTML provide an easy migration path from familiar concepts towards authoring complex and format-rich hypermedia documents. In this way, familiarity with a system like LaTeX makes it possible to contribute to and benefit from a rapidly expanding global hypermedia network.

# References

[1] T. Berners-Lee, R. Cailliau, J. Groff, and B. Pollerman. Worldwide web: The information universe. *Electronic Networking: Research, Application and Policy*, (1), 1992.

[2] Joe Levy. The world in a web. *The Guardian*, page 19, November 11 1993.

[3] Vern Paxson. Growth trends in wide-area TCP connections. *IEEE Network*, 1993. Available at ftp://ftp.ee.lbl.gov/WAN-TCP-growth-trends.revised.ps.Z.

[4] Nikos Drakos. Text to Hypertext conversion with LaTeX2HTML. Baskerville, December 1993.

# A Further Information

LaTeX2HTML is written in Perl and requires freely available software. More information on how to get, install and use it is available via the WWW[10] or using anonymous ftp from ftp.tex.ac.uk in pub/archive/support/latex2html.

Several computers on the Internet have public access World Wide Web clients accessible by telnet e.g.

• telnet info.cern.ch (direct connection - no username or password required)

• telnet ukanaix.cc.ukans.edu ("Lynx" requires a vt100 terminal. Log in as www.)

Information on the WWW is also available via anonymous ftp from ftp.germany.eu.net in pub/infosystems/www.

The Mosaic clients are in the directory /pub/infosystems/www/ncsa/Web.

[10] http://cbl.leeds.ac.uk/nikos/tex2html/doc/latex2html/latex2html.html