# Converting files to HTML

# *2HTML

**HTML is the most important format on the Web but a lot of data is still created or is available in other formats, such as Office documents, tables, PDF or ASCII files. Hans Georg Esser looks at the available conversion methods**

When considering the conversion of a diverse range of document types into HTML format, it's worth examining not only how these conversions work, but also how good they are. Most Office packages (under both Linux and Windows) have an HTML export option, but the results vary widely and are often unsatisfactory.

## Microsoft Word

There are several different possibilities when converting Word documents into HTML. Firstly, Word (2000) itself offers its won conversion function under File/ Save as Web Page. HTML files created this way can be viewed in all Web browsers but they're not particularly suitable for further editing due to the continual use of variously defined styles in the text. An example of this is the simple listing of individual items, which instead of being shown as:

```
<li>Text</li>
```

is presented as a line of the form:

```
<li class=MsoNormal style='mso-list:l0 ➥
level1 lfo1;tab-stops:list 36.0pt'>Text</li>.
```

This solution may be practical if you merely seek a quick means of placing a Word file on the Net, however it does require having both Windows and Word installed. For those of you who want to edit HTML files further, or don't have access to Word, there are some alternatives.

One of these is the program word2x. You can find the current version (0.005) on the Web at *http://word2x.alcom.co.uk/*. In our test, a Word 8 document could not be converted (the output was empty). The reason for this is that a special conversion tool, called *wv* (earlier referred to as *mswordview*), exists for the current Word 8 format. It can be found on the Web at *http://www.wvware.com/*. Once the wv tools are activated, the command

```
wvHtml test.doc test.html
```

begins file conversion. Unfortunately, the results of the conversion are even more disappointing than when using Word directly. The simple listed item given in the example above takes the following form:

```
<li><p><div align="left" style="padding: ➥
0.00mm 0.00mm 0.00mm 0.00mm;"> <p style=➥
"text-indent: 0.00mm; text-align: left; ➥
line-height: 4.166667mm; color: black; ➥
background-color: white;">
Text
</p></div></li>
```

In addition to this, headings do not conform with the HTML standards of <h1>, <h2> and so on. The *wv* tools offer conversions into other formats besides Word, such as LaTeX, PostScript, PDF and more. However, even the LaTeX format generated by wvLaTeX could not be converted into useful HTML with *latex2html* (see below).

## StarOffice

In a similar manner to Word, the text module of StarOffice also offers its own HTML export function. This produces quite useful HTML, and an export function also exists for converting files to Word

## LaTeX

The free text typesetting system LaTeX (*http://www.latex-project.org/*) has many friends, particularly amongst scientists, due to the fact that it enables the simple creation of complex formulas for seminars or theses. Thanks to LyX, you can create documents using a text-processing environment, which means you don't need to learn the LaTeX syntax (*http://www.lyx.org/*). LaTex uses its own mark-up commands, which have a certain similarity with HTML in terms of their structure. An automatic conversion from LaTeX files to HTML is not therefore a giant leap.

A program that does this is latex2html (*http://www-texdev.mpce.mq.edu.au/l2h/docs/manual/*). This has an added function which is particularly useful for larger documents – you can select whether to create a single HTML file (option -split 0), or whether each chapter, paragraph etc. should be outputted to a separate file. The HTML code produced is very clean, and tables of contents, footnotes and cross-references are correctly converted.

format. To export an HTML file, simply select *File/ Save as* and then select the file format *HTML (StarOffice Writer)*.

## Tables

After text files, the most important Office documents are the products of the various spreadsheet applications such as Excel and StarCalc. As we did for text programs, we will first look at the export functions.

## Microsoft Excel

The first program we put to the test was Microsoft's Excel 2000. For the purpose of this test we created a simple table with the columns and column totals, which was then saved as a HTML file. The result looked tidy enough in a Web browser and the formatting of bold and colour were preserved. The only downside was that the sum formula was lost (i.e. the result values were saved). As with Word, the outputted HTML file was very large – here producing a 8,194 byte file from 800 bytes of information – with most of its bulk arising from style declarations. The formulas were retained for the subsequent re-import into Excel: In this case the sum fields had the structure:

```
<td class=xl29 align=right x:num="41.96 " ➘
x:fmla="=SUM(D5:D9)">41.96</td>
```

which is useful for other Excel users. One error during the conversion did come to our attention: While the number columns in Excel were right-aligned (format ###,##), they became left-aligned in the HTML file.

## PDF as a picture?

The converter pdf2html takes the path of least resistance. Instead of analysing the data in a PDF document (PDF: Portable Document Format, a standard of Adobe) and converting this to HTML, it simply converts the individual PDF pages to .png pictures and produces HTML pages from these. This is fast and simple but it doesn't permit any analysis of the data by the web page viewers. The program is found under *http://atrey.karlin.mff.cuni.cz/~clock/twibright/pdf2html/*.

The similarly named tool pdftohtml (*http://www.ra.informatik.uni-stuttgart.de/~gosho/pdftohtml/*)takes another path. Here, the PDF data is analysed and is converted into an HTMl text file. pdftohtml also detects and converts links in PDF files. Pictures are likewise extracted from the file and built into the appropriate place in the HTML file. The visual and layout quality is not amazing (any formatting information is ignored), but at least the created file can serve as a starting point for subsequent fine-tuning – the only problem with this is that the tool has a strange habit of only putting one word in each line in the HTML code.

## StarCalc

Our next candidate was StarCalc, the spreadsheet analysis tool from the StarOffice package. Converting the same table produced a somewhat smaller HTML file (3,572 bytes). Here, a table entry had the form:

```
<TD WIDTH=86 HEIGHT=17 ALIGN=RIGHT ➘
SDVAL="41.96" SDNUM="1031;"><B><FONT ➘
COLOR="#0000FF">41.96</FONT></B></TD>
```

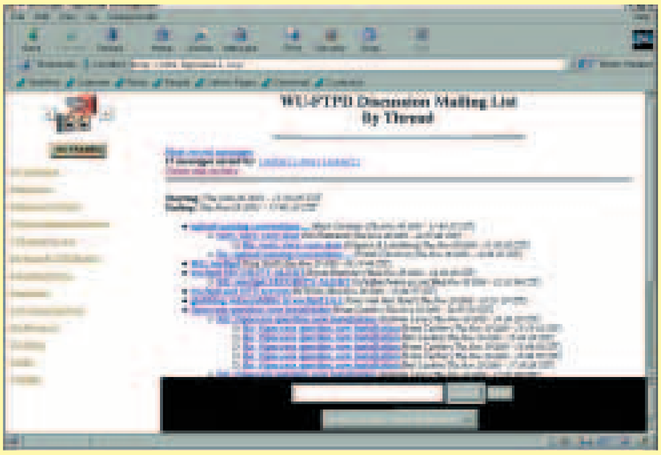
The style specifications were done without and as a

## Email

Why would anyone want to convert emails into HTML, one may ask? This is an interesting option that enables mailing lists administrators to make postings public through a Web page, for example. The relevant package Hypermail, *http://www.hypermail.org/*, permits the conversion to an Mbox compatible mail file (like those produced by Netscape Mail, Kmail, mutt and elm). To create a new directory with the mailbox generated HTML files, simply use the command

```
mkdir /tmp/hypermail
hypermail -m ~/Mail/Incoming -d /tmp/hypermail
```

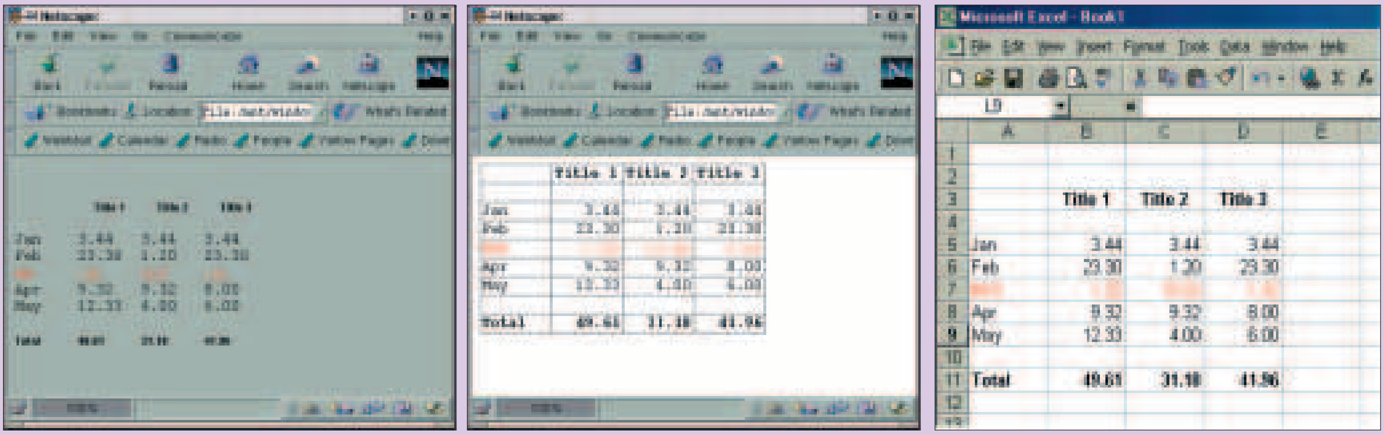
The program produces a separate HTML file as well as additional overview files for sorting (according to thread, date, topic, author and attachments) for each individual mail, and puts these in the indicated directory. The sorting data is streamed to a separate subdirectory, which enables fast access to all attachments – the HTML file of a mail with an attachment no longer contains the attachment's information; instead it contains a link to the file in the mail's directory. This is also extremely practical for purposes of archive keeping. An example of the thread overview is shown above.



WU-FTPD Discussion Mailing List
By Thread

**Hypermail produces an eye-pleasing overview page, in which the threads can be seen**

Figure 1

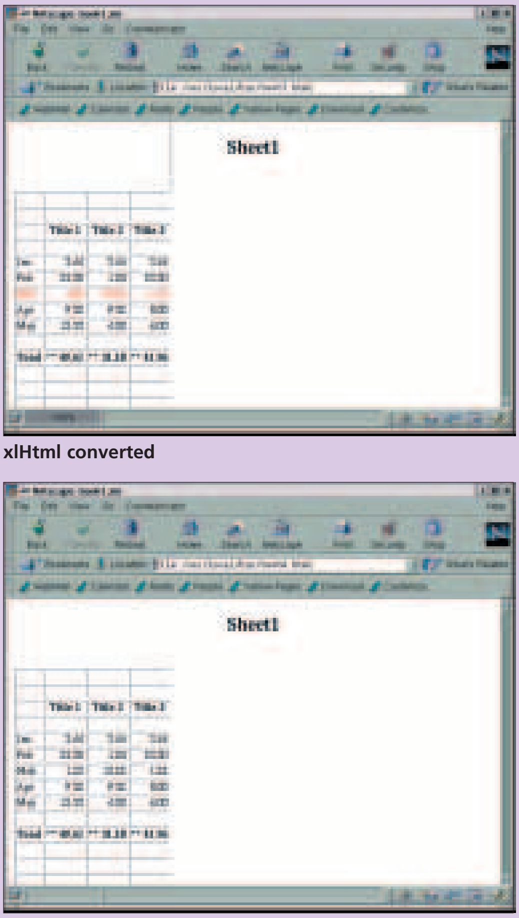


|  | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 |  |  |  |  |  |
| 2 |  |  |  |  |  |
| 3 |  | Title 1 | Title 2 | Title 3 |  |
| 4 |  |  |  |  |  |
| 5 | Jan | 3.44 | 3.44 | 3.44 |  |
| 6 | Feb | 23.30 | 1.20 | 29.30 |  |
| 7 |  |  |  |  |  |
| 8 | Apr | 9.32 | 9.32 | 8.00 |  |
| 9 | May | 12.33 | 4.00 | 6.00 |  |
| 10 |  |  |  |  |  |
| 11 | Total | 49.61 | 31.18 | 41.96 |  |
| 12 |  |  |  |  |  |

Excel export

StarOffice export

Excel original



Sheet1

xlHtml converted

Sheet1

xlHtlm -nc converted

consequence, the format options for each field need to be indicated separately. The alignment of the number fields was, incidentally, correctly converted. StarOffice did not store the formulas, however, and with the re-import into StarCalc the formulas were thus lost.

## xlHtml

Anyone who receives a table via email, and who has neither StarOffice nor Excel handy, will be pleased to know there is a tool that enables conversion to HTML without needing to start an Office package. This service is offered by xlHtml (*http://www.xlhtml.org/*), the current version being 0.2.8. The program can be translated and installed in the normal way with ./configure; make; make install and with a cue, produces the form

```
xlHtml test.xls > test.html
```

This table only proves useful on closer inspection. Most fields appear in white writing on a white background and are therefore illegible. Only after changing the colour can something be seen. Table content is however correct.

The sums were correctly calculated, although a warning was given that they would possibly be incorrect – this warning can be suppressed using the parameter -fw ("suppress formula warnings"). The outputted HTML file, with a size of 2,705 bytes, is even smaller than that produced by StarOffice. This is due to minimum formatting as seen here with the HTML data of a table field:

```
<TD><FONT COLOR = ➜
"00FF00"><B>41.96</B></FONT></TD>
```

The syntax of the FONT attribute is incorrect. If correct, it would read: *font color="#00FF00"*. As the colours were not correctly detected (blue writing became green, black became white), you may as well ignore the colours completely – xlHtml provides the option -nc ("no colours ") to do this and the result is now pure, simple HTML, which is very suitable for subsequent editing:

```
<TD><B>41.96</B></TD>
```

There is another option, which is also useful. -te ("trim edges") gets rid of empty columns and lines in the upper left hand corner of the document. xlHtml matches the source of an Excel file with remarkable accuracy. This was also shown by the fact that a pure Excel document was converted free of errors whereas an Excel document produced by StarCalc supplied only zeros in the total fields.

In Figure 1 you can see the different converted files as displayed in Netscape.

## ASCII files

ASCII files occasionally crop up, usually in the form of longer HOWTOs or Read Me files. These contain no mark-up information at all about the document's structure. There are nevertheless tools, which attempt the conversion to HTML. txt2html (*http://www.aigeek.com/txt2html/*) is a program that offers the possibility of indicating something about the structure of the text file on the basis of templates.

t2t (*http://216.254.0.2/~dogbert/t2t/*) takes on the task of analysing tables in ASCII files. Many databases and spreadsheet programs offer the opportunity to export files into "Tab Delimited ASCII" format. – Such tables are easily processed by *t2t*.

## PostScript

Many documents can be found in the PostScript format and as such, there are also HTML converters for this class of document. One example is the Pscript Document Publishing System *http://www.ubka.uni-karlsruhe.de/~guenter/pscript/*. One should not expect too much of the conversion however: items such as header and footer lines, or page numbers are simply added in with the produced HTML. As with the PDF converter, formatting is also ignored.

## Is anything better than nothing?

The results of these free conversion tools are not always convincing, and in many cases, a manual rework with a HTML editor is necessary. What is crucial however, is that the data can be converted into HTML in some form or other. A re-edit is surely easier than manually transferring the data and formatting the HTML from scratch. Good luck in generating your own Web content.