Extrapolating TEX4ht

http://tug.org/tex4ht

TEX4ht Project

This manual is for TEX4ht.

Table of Contents

1 Introduction

TEX4ht is a TEX package created and developed by Eitan M. Gurari, who was Associate Professor of Computer Science at Ohio State University until his premature death on June 22, 2009. Our continuing work on his software is dedicated to his memory.

TEX4ht translates documents written in TEX or any of its common variants (LATEX, ConTEXt, etc.) into other markup formats, such as HTML, XML, SGML, etc., optionally using MathML or other formats, with nearly endless possibilities for customization. The home page of the project is http://tug.org/tex4ht. The software is released under the LATEX Project Public License, version 1.3 or later.

The present document is currently focused on maintenance of TEX4ht itself, which includes hundreds of TEX packages, hypertext fonts, C and Java programs, DTDs, usually all wrapped in a (homegrown) literate programming style. For user documentation, please see the resources on the home page. Perhaps this manual will be more extensive one day.

TEX4ht is currently maintained by CV Radhkrishnan and Karl Berry (the "TEX4ht Project"); we would be very grateful for additional volunteers. The development site, mailing lists, etc., are also linked from http://tug.org/tex4ht.

2 Implementation: How TEX4ht works

TEX4ht has a three-step approach to the translation process:

2.1 Preprocessing with `ht*` to DVI

`foo.tex` is processed with the appropriate script (`htex`, `htlatex`, `htcontext`, `...`) which will load `tex4ht.sty` and other relevant packages to create `foo.dvi` by calling the `tex` compiler with appropriate format. TEX4ht adopts a different pattern of package loading. It loads `tex4ht.sty` at the beginning of the document, stops after a while, then allows loading all the packages which the author wants with `\usepackage` function. Once it reaches the `\begin{document}` hook, which means that all extra package loading has been completed, `tex4ht` loads itself for the second time. This time, since it has the information about all additional packages loaded, it will call the relevant `.4ht` macro packages to assist the main `tex4ht.sty`.

For instance, if the author has used `biblatex.sty`, `tex4ht` will call `biblatex.4ht` or if `amsmath.sty` was used, `amsmath.4ht` will be input, and so on. Eitan wrote a `*.4ht` for nearly all of the most often used LATEX packages.

Then the source `foo.tex` is processed in the usual manner to create `foo.dvi`. With TEX4ht, we always need `.dvi` output since `.pdf` output is not useful for conversion. This is the first stage in the translation process.

2.2 Processing with `tex4ht`

The second stage is to call the `tex4ht` binary to post-process `foo.dvi`. This is the real meat of the process where ASCII characters of element and attribute names, attribute values, etc., which are output in `\special`s in the `.dvi`, are extracted. Also, it does the substitution of characters in textual strings in the typeset version.

As you may be aware, the `.dvi` file has font and position information of all characters of all strings in the document. Suppose the `.dvi` has a character γ. When rendered to a particular media, the character is taken from the 13th position of the font by name, `cmmi`. When extracting text from the `.dvi`, instead of taking the glyph from `cmmi.pfb`, `tex4ht` takes the character from the 13th position in the corresponding hypertext font, `cmmi.htf` (`htf` denoting hypertext font, multitudes of which were again created by Eitan).

A *hypertext font* is an ASCII file, created by hand in a text editor, with each line defining a character of the font. The first line corresponds to character code 0, the second to character code 1, etc. In `cmmi.htf` for example, the first 13 lines look something like this:

```
cmmi 0 127
'&#x0393;' ''  Gamma     0 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
'&#x0394;' ''  Delta     1 % cmmi.htf (unicode)            2003-03-27 %
'&#x0398;' ''  Theta     2 % Copyright (C) 2000--2003 Michel Goossens %
'&#x039B;' ''  Lambda    3 %                            Eitan M. Gurari %
'&#x039E;' ''  Xi        4 %                                            %
'&#x03A0;' ''  Pi        5 % This file can redistributed and/or         %
'&#x03A3;' ''  Sigma     6 % modified under the terms of the LaTeX      %
```

```
'&#x03A5;' ''  Upsilon   7 % Project Public License Distributed from  %
'&#x03A6;' ''  Phi       8 % CTAN archives in directory               %
'&#x03A8;' ''  Psi       9 % macros/latex/base/lppl.txt; either       %
'&#x03A9;' ''  Omega    10 % version 1 of the License, or (at your    %
'&#x03B1;' ''  alpha    11 % option) any later version.               %
'&#x03B2;' ''  beta     12 %    However, you are allowed to modify    %
'&#x03B3;' ''  gamma    13 % this file without changing its name, if  %
...
```

The character code given in the 13th position of `cmmi.htf` is `&#x03B3;`, which is the Unicode entity for lower case gamma (γ). `tex4ht` will happily substitute this code in place of the typeset gamma character in the `dvi` during post-processing the `.dvi`. Hence, the converted document will have appropriate entities (or whatever we want) in place of the TEX-font-specific `.dvi` references. You can add prefixes or suffixes to the entities or character codes. eg., `<mi>&#x03B3;</mi>` (MathML code for γ).

2.3 Post-processing

The third and final stage is to post-process the translated document further which may involve:

- parse the document with appropriate parser.
- create `.png` or other images of math formulae and equations, if requested (for the sake of browsers which do not support MathML).
- write out `.css` files for proper rendering in a browser.
- perform system dependent tasks like copying to target directories or `ftp` to different destinations, etc.
- During post-processing, one can output the translated document as several chunks, such as one file for each section, instead of having a single long document. We use this feature to write out many files to overcome various I/O limitations of TEX.

3 Literate sources

Following are the literate source files which comprise TEX4ht. Some modifications to specific files are described below. We have globally updated the license information.

Specific processing instructions are provided as remarks at the top of each source file. All packages, C and Java sources, fonts, DTD's, etc., are generated from the literate sources by running TEX, LATEX or any of the many TEX4ht scripts such as `ht`, `htlatex`, ...

1. `tex4ht-4ht.tex`
2. `tex4ht-auto-script.tex`
3. `tex4ht-bibtex2.tex`
4. `tex4ht-c.tex`
5. `tex4ht-cond4ht.tex`
6. `tex4ht-cpright.tex`
7. `tex4ht-dir.tex`
8. `tex4ht-docbook-xtpipes.tex`
9. `tex4ht-docbook.tex`
10. `tex4ht-env.tex`
11. `tex4ht-fonts-4hf.tex`
12. `tex4ht-fonts-cjk-utf8.tex`
13. `tex4ht-fonts-cjk.tex`
14. `tex4ht-fonts-modern.tex`
15. `tex4ht-fonts-noncjk.tex`
16. `tex4ht-htcmd.tex`
17. `tex4ht-html-speech-xtpipes.tex`
18. `tex4ht-html-speech.tex`
19. `tex4ht-html0.tex`
20. `tex4ht-html32.tex`
21. `tex4ht-html4.tex`
22. `tex4ht-info-html4.tex`
23. `tex4ht-info-javahelp.tex`
24. `tex4ht-info-mml.tex`
25. `tex4ht-info-ooffice.tex`
26. `tex4ht-info-svg.tex`
27. `tex4ht-info.tex`
28. `tex4ht-javahelp-xtpipes.tex`
29. `tex4ht-javahelp.tex`
30. `tex4ht-jsmath.tex`
31. `tex4ht-jsml-xtpipes.tex`
32. `tex4ht-jsml.tex`

33. tex4ht-mathltx.tex
34. tex4ht-mathml.tex
35. tex4ht-mathplayer.tex
36. tex4ht-mkht.tex
37. tex4ht-moz.tex
38. tex4ht-oo-xtpipes.tex
39. tex4ht-ooffice.tex
40. tex4ht-ooimpress.tex
41. tex4ht-options.tex
42. tex4ht-sty.tex
43. tex4ht-svg.tex
44. tex4ht-t4ht.tex
45. tex4ht-tei.tex
46. tex4ht-unicode.tex
47. tex4ht-word.tex
48. tex4ht-xhtml-xtpipes.tex
49. tex4ht-xhtmml-xtpipes.tex
50. xtpipes.tex

3.1 tex4ht-4ht.tex

This is the (extremely large) literate source for all the `.4ht` files in the TEX4ht bundle. Run the following command to generate all `.4ht` files:

```
ht tex tex4ht-4ht
```

Nicholas Cole posted a bug report on the `texhax` mailing list regarding an undefined control sequence error of `\blx@resetpuncthook` and `\blx@csq@ifkernmark`. The reason was that these macros were not initialized. So, we added the following lines at the beginning of `\<config biblatex\>`:

```
\let\blx@resetpuncthook\@empty
\let\blx@csq@ifkernmark\@empty
```

Christoph Haug reported that `\bib@field@entrykey` creates an undefined control sequence error if `\printbibliography` is invoked. Another of with uninitialized macros, solved by adding:

```
\let\bib@field@keyentry\@empty
```

Also, Christoph said that there were a few spurious spaces after the opening parenthesis of year in an author-year citation and few other places. All were fixed.

3.2 tex4ht-cpright.tex

The standard copyright statement was changed to the following:

```
\<TeX4ht copyright\><<<
%
```

```
% This work may be distributed and/or modified under the
% conditions of the LaTeX Project Public License, either
% version 1.3c of this license or (at your option) any
% later version. The latest version of this license is in
%   http://www.latex-project.org/lppl.txt
% and version 1.3c or later is part of all distributions
% of LaTeX version 2005/12/01 or later.
%
% This work has the LPPL maintenance status "maintained".
%
% The Current Maintainer of this work
% is the TeX4ht Project <http://tug.org/tex4ht>.
%
% If you modify this program, changing the
% version identification would be appreciated.
>>>
```

Filename, author name and date are inserted at the top of this statement.

3.3 tex4ht-dir.tex

Defines the path of your `tex4ht` package files. The default provided by Eitan was:

```
\def\HOME{/home/4/gurari/tex4ht.dir/}
\def\DTDS{/home/4/gurari/dtd.dir/}
```

We switched these to use '.' instead of his hardcoded path.

3.4 tex4ht-fonts-4ht.tex

This file generates all the `*.4hf`—hypertext font files—of the TEX4ht bundle. The file has 101806 lines! We had to increase TEX's memory and make new format for \latex to run this file. Here are the new values:

`strings=494909`
`pool_size=1180334` (string characters)
`main_memory=7999999` (words of memory)
`multiletter control sequences=15000+50000`

Also, these needed values are the default in TEX Live 2009:

`font_mem_size=3000000` (words of font info)
`hyph_size=8191` (hyphenation exceptions)

3.5 tex4ht-mkht.tex

CVR made significant changes on September 13, 2009:

- All the backslash characters in the path names (conventional directory Separators under Windows) have been changed to forward slash. This is per the suggestion of Akira Kakuto, primary Windows developer for TEX Live.
- `\version` has been redefined.

- New functions, `\ScriptFileName` and `\AddExtn` have been defined to add file names of the script at the top of each script or batch file. These were not provided in the versions written by Eitan, but now needed for best license practices.
- `\AddExtn` will add `.bat` if and only if the script is a batch file.
- A new function `\<Mycopyrightnotice\>` has been defined to add the usual copyright information (see Section 3.2 [tex4ht-cpright.tex], page 5) to each script when written out.
- The `\Rem` macro used in `\<Mycopyrightnotice\>` expands to the `#` character in Unix scripts and `Rem` in Windows batch files.