Transforming XHTML to LaTeX and BibTeX

Dan Connolly

No Institute Given

Abstract. We transform XHTML to LaTeX and BibTeX to allow technical articles to be developed using familiar XHTML authoring tools and techniques.
PRE-PUBLICATION DRAFT 1.20 of 2006-04-20T17:10:49Z. DO NOT CIRCULATE.

1 Introduction

Occasionally a web page turns the corner from a casually drafted idea to an article worthy of publication. Computer science conferences often require submissions using specific LaTeX styles; for example, the ISCW2004 submission instructions require that submitted papers be formatted in the style of the Springer publications format for Lecture Notes in Computer Science (LNCS). XSLT is a convenient notation to express a transformation from XHTML to LaTeX.

Tools to transform from LaTeX to HTML are commonplace, but there are far fewer to go the other way. A little bit of searching yielded some work[1] that was designed to undo a transformation to XHTML. It used an odd XHTML namespace and exhibited various other quirks specific to reversing that transformation, but it provided quite a boost up the LaTeX learning curve[2].

That code did not integrate with the BibTeX. In order to take advantage of automatic bibliography formatting traditionally provided by LaTeX styles, after studying the BibTeX format[3] for a bit, `xh2bibl.xsl` was born.

Together with tradtional `pdflatex` and `bibtex` tools[4] and and XSLT processor such as xsltproc[5], this transformation can turn ordinary web pages with just a bit of special markup into camera-ready PDF in specialized LaTeX styles.

1.1 A Quick Example

This article demonstrates the basic features. See:

- `Overview.pdf`
- `Overview.tex`
- `Overview.bib`

They are produced ala:

```
$ make Overview.pdf
xsltproc  --novalid --stringparam DocClass llncs \
  --stringparam Bib Overview --stringparam BibStyle splncs \
  --stringparam Status prepub  \
        -o Overview.tex xh2latex.xsl Overview.html
TEXINPUTS=.:../../../2004/LLCS: pdflatex  Overview.tex
This is pdfTeX, Version 3.14159-1.10b (Web2C 7.4.5)
...
Output written on Overview.pdf (3 pages, 62474 bytes).
Transcript written on Overview.log.
xsltproc  --novalid -o Overview.bib xh2bib.xsl Overview.html
BSTINPUTS=.:../../../2004/LLCS: bibtex  Overview
This is BibTeX, Version 0.99c (Web2C 7.4.5)
The top-level auxiliary file: Overview.aux
The style file: splncs.bst
Database file #1: Overview.bib
TEXINPUTS=.:../../../2004/LLCS: pdflatex  Overview
This is pdfTeX, Version 3.14159-1.10b (Web2C 7.4.5)
...
Output written on Overview.pdf (3 pages, 67583 bytes).
Transcript written on Overview.log.
TEXINPUTS=.:../../../2004/LLCS: pdflatex  Overview
This is pdfTeX, Version 3.14159-1.10b (Web2C 7.4.5)
...
Output written on Overview.pdf (3 pages, 67167 bytes).
Transcript written on Overview.log.
```

2 Features

The transformation `xh2latex.xsl` works in the obvious way for many idioms:

- sections headings: `h2`, `h3`, `h4`
- paragraphs: `p`
- itemized lists: `ul`, `dl`
- enumerated (numbered) lists: `ol`
- tables: `table border="1"`, `tr`, `td`
- verbatim: `pre`
- phrase markup: `em`, `code`, `tt`, `i`, `b`

Table support is limited to tables with `border="1"` and where all rows have the same number of cells. For example:

Name	Address	Phone
John Doe	123 High St.	555-1212
Jane Smith	456 Low St.	555-1234

Specialized markup is required for other idioms. An article.css stylesheet provides visual feedback for this special markup.

To use a latex package, add a link to the head of your document a la:

```
<link rel="usepackage" title="url"
  href="ftp://cam.ctan.org/tex-archive/macros/latex/contrib/misc/url.sty" />
```

The package name is taken from the title attrbute. The href attribute is not used in the LaTeX conversion.

We recommend the url.sty package, per a TeX FAQ. For example: `http://www.w3.org/People/Connolly/`.

2.1 Front Matter

The following patterns are used to extract the title page material:

- `div/@class="maketitle"`
  - title: `h1`
  - abstract: `div/@class="abstract"`
  - author: `address/a[@rel="author"]`
- keywords: `div[@class="keywords"]`
- terms: `div[@class="terms"]`

support for WWW2006 style authors, following ACM style, is in progress.

2.2 Cross references and footnotes

The `a[@class="ref"]` pattern is transformed to the LaTeX \ref{`label`} idiom, assuming the reference takes the form `href="#label"`.

The footnote pattern is `*[@class="footnote"]`.

2.3 Figures

The `div[@class="figure"]` pattern is transformed to a figure environment; any `div/@id` is used as a figure label. The file pattern is `object/@data`. *Figures are currently assumed to be PDF; the* `object/@height` *attribute is copied over.* The caption pattern is `p[@class="caption"]`. *@@need to test this.* Be sure to include the `epsfig` package a la:

```
<link rel="usepackage" title="epsfig" />
```

2.4 Citations and Bibliography

An `a` element starting with an open square bracket `[` is interpreted as a citation reference. The `href` is assumed to be a local link ala `#tag`.

The pattern `dl/@class="bib"` is used to find the bibliography. Each item marked up ala...

```
<dt class="misc">[<a name="tetex">tetex</a>]</dt>
<dd>
<span class="author">Thomas Esser</span>
<cite><a
href="http://www.tug.org/tex-archive/help/Catalogue/entries/tetex.html"
>The TeX distribution for Unix/Linux</a></cite>
February <span class="year">2003</span>
</dd>
```

or

```
<dt class="misc" id="tetex">[tetex]</dt>
...
```

Note the placement of the bibtex item type `misc` and the tag `tetex` and keep in mind that `bibtex` ignores works in the bibliography that are not cited from the body.

The `xh2bibl.xsl` transformation turns this markup into BibTeX format. `xh2latex.xsl` transforms the entire bibliography `dl` to a `\bibliography{...}` reference.

capitalization of titles seems to get mangled. I'm not sure if that's a feature of certain bibliography styles or what.

2.5 Bugs/Caveats/Misfeatures

- Composed characters and such in the bibliography are handled with a sort of kludge, e.g. `K<span title='\"o'></span>bler`
- The `samp` element is used to pass LaTeX math markup thru, e.g. `<samp>\Delta</samp>`

3 Makefile support

Formatting a LaTeX document is done in several passes. One typical manual shows:

```
ucsub>  latex MyDoc.tex
ucsub>  bibtex MyDoc
ucsub>  latex MyDoc.tex
ucsub>  latex MyDoc.tex
```

The follwing excerpt from `html2latex.mak` shows some rules to accomplish this using make:

```
.html.tex:
$(XSLTPROC) --novalid $(HLPARAMS) \
-o $@ xh2latex.xsl $<

.html.bib:
$(XSLTPROC) --novalid -o $@ xh2bib.xsl $<

.tex.aux:
TEXINPUTS=$(TEXINPUTS) $(PDFLATEX) $<

.tex.bbl:
BSTINPUTS=$(BSTINPUTS) $(BIBTEX) $*


.aux.pdf:
TEXINPUTS=$(TEXINPUTS) $(PDFLATEX) $*
TEXINPUTS=$(TEXINPUTS) $(PDFLATEX) $*
```

Sources:

- `xh2latex.xsl`
- `xh2bib.xsl`
- `article.css`

References

1. Gurari, E.M.: XSLT from XHTML+MathML to LATEX. `http://www.cse.ohio-state.edu/~gurari/docs/mml-00/xhm2latex.html` (2000)
2. Mann, S.: Beginner's LaTeX Tutorial. `http://www.csclub.uwaterloo.ca/u/sjbmann/tutorial.html` (1994)
3. Rugaber, S.: The Citation project . `http://www.cc.gatech.edu/classes/RWL/Projects/citation/` (1998)
4. Esser, T.: The TeX distribution for Unix/Linux. `http://www.tug.org/tex-archive/help/Catalogue/entries/tetex.html` (2003)
5. Veillard, D.: The xsltproc tool. `http://xmlsoft.org/XSLT/xsltproc2.html` (2003)