WordPress to LaTeXwith Pandoc and J: Prerequisites (1)

Posted: 12 Feb 2012 01:33:11

There are no quick WordPress to LaTeX fixes Over the next three posts I will describe how to convert WordPress's export XML to LaTeX source code.

WordPress to LaTeX

I know that many of you are looking for a quick WordPress to LaTeX fix; unfortunately there are no quick fixes. The two formats come from different worlds and are used in different ways. Producing useful LaTeX source from WordPress export XML will require manual edits. My goal here is to minimize manual edits, produce high quality LaTeX source and to *outline* what you will have to contend with. To get an idea of what you can expect download the LaTeX compiled version of this post.

Visual and Logical composition WordPress and LaTeX are examples of the two basic approaches, *visual* and *logical*, taken by writing software. Visual systems value appearance. It matters what things look like and no effort is spared to get the right look. Logical systems value content. What's said is far more important than what it looks like. Logical systems impose order and structure and typically defer visual elements. As you might expect there is no such thing as a pure visual or logical writing system. Successful systems use both approaches to a greater or lesser degree. Composing WordPress blog posts is roughly 35% visual and 65% logical.[1] LaTeX composition is about 10% visual and 90% logical. The numbers do not line up; there is a basic mismatch here.

Many format X to LaTeX converters tackle this mismatch by attempting to maintain visual fidelity. This is a catastrophic error that renders the entire conversion useless. Here's a hint. If you're using a predominantly logical system like LaTeX you don't give a rodent's posterior about visual fidelity. This method dispenses with all but the most basic of visual elements. No attempt is made to preserve fonts, type sizes, image scale, justification, hyphenation, text color and so forth. The goal is to produce working LaTeX source that can be transformed to whatever final layout the author desires.

[1] Actually this is not bad. Page layout systems are far worse. A typical layout system might be 90% visual and 10% logical making layout systems polar opposites of LaTeX.

Prerequisite Software I use two programs to transform WordPress export XML to LaTeX the J programming language and John MacFarlane's Pandoc. Pandoc is an excellent text mark-up to mark-up converter. It wisely avoids attempting to convert entire complex documents and focuses on getting *parts* of documents right. It does a particularly good job of converting HTML to LaTeX which is a *crucial* part of this process. I use Pandoc to transform the HTML embedded in WordPress export XML `CDATA` elements to `*.tex` files and I use J to preprocess and post process Pandoc inputs and outputs and to stitch everything together into a set of LaTeX ready files.

Download Pandoc from here. I use the Windows command line version. There are Linux and Mac versions as well. Download J from here. The easiest J install is the 32 bit Windows version. Other versions require additional steps to configure and deploy. If you are already a J user there is no need to install a particular system but you will need:

1. The task library `require 'task'`
2. The utility program `wget.exe`

Both of these components are typically part of the J distribution.

Install and check prerequisites To continue download and install Pandoc and J and run the following tests; if you succeed you're system is ready for WordPress to LaTeX with Pandoc and J: LaTeX Directories (2).

Pandoc Test: Download the test file: `cdata.html` and run Pandoc from the command line:

```
pandoc -o cdata.tex cdata.html
```

`cdata.html` is an example of the HTML code you find in WordPress export XML `CDATA` elements. *Note:* required files are available on GitHub here.

J Test: Start a J session and enter the following commands:

```
require 'task'
shell  'wget --help'
site=. 'http://conceptcontrol.smugmug.com/photos/'
shell  'wget ',site,'i-mNK4RHL/0/L/i-mNK4RHL-L.png'
```

If the shell command is properly loaded and `wget.exe` is found you will see help text. The second shell command downloads an image file. Downloading post images is part of the overall conversion process.

From the blog: Analyze the Data not the Drivel
John D. Baker — revised: August 14, 2020