EPub for word processing users

[This is a repost of http://jiscpub.blogs.edina.ac.uk/2011/04/05/epub-for-word-processing-users/ - if you have comments please make them over on there on the jiscPUB blog.]

Author: Peter Sefton <pt@ptsefton.com>

Date: Time-stamp: <2011-04-05 15:38:38>

Description: First informal report on progress with Workpackage 3. Looks at tools for turning word processing documents such as Microsoft Word documents into epub.

Summary

Last week I started on the JISCPub project on workpackage 3. My role on this project is to be the “Wordprocessing Tool Expert”.

I started by considering the first use case, seeing how a PhD thesis could be converted to EPUB for mobile use. I happen to have a very worthy test file in the form of Danny Kingsley's Creative Commons Licensed: “The effect of scholarly communication practices on engagement with open access: Australian study of three disciplines” from 2008.

I am working three kinds of deliverable:

•Blog reports. This is the first.

•Demonstration software – I have produced a simple demo that can covert word documents into EPUB files – you try it out, assuming you can work out how, or read on for more background about the service, how raw it is, excuses about its somewhat agricultural interface and the fact that it might not work very well yet.

•Demonstration files showing the results of various processes. I have set up a dropbox.com folder for the project team, where I will be keeping records of the experiments I'm doing with various bits of software. This will be available as a data set at the end of the project. One demo is this post, formatted as an EPUB file – using the tool I discuss below. I tried to upload it to WordPress but it does not meet security requirements, you can grab the post in EPUB format from here. Comments welcome – this is a test process only.

Initial impressions

I started this investigation by assuming that I had a thesis I wanted to publish as an ebook, via EPUB and doing various Google searches for stuff like “How to make an EPUB”. Theo Andrew is going to talk to some real users about this soon – so we'll find out a lot more about what kinds of assumptions are valid to make about our users, graduate students and academics.

There are some sites which review software, notably Jedisaber's, which has reviews of many software packages related to EPUB and ebooks in general, and a how-to on making an EPUB from scratch, by hand. I have been trying out various bits of software against the uses cases for this project – without a great deal of success – and I will document this behind the scenes and save input and output documents as part of the data set for this project.

As it says in the plan for this workpackage the most obvious bit of software to try out is Calibre which is a mature open source ebook management application available for all the major platforms. Calibre is a very feature-rich application with a somewhat quirky interface which doesn't do the one thing I wanted for my first experiment. It doesn't convert Microsoft Word .doc files into `.epub` format. Yes, you can deal with Word docs by doing a 'Save as HTML' but that's not the ideal process for casual users. Calibre does do open

document format (`.odt`) files, though, so I tried that with Danny's thesis, using open office to save it as `.odt`, after some minor tweaking in OpenOffice. I found that:

•On large files like the thesis it takes around 100-200 minutes to convert the document, in the background, using the GUI on my modest desktop PC.

•Some graphics are not supported, for example embedded vector drawings. I don't think there's a good solution for this that does not involve firing up a word processor to render some things – I will come back to this in a future post on the current (terrible) state of the art in Word processor to HTML conversion, and what could be done about it.

(I am wondering if the speed of Calibre is the reason I never heard back from the odt2ebook.com site, which offered free conversions; there was supposed to be an email notification but it has been a few days.)

Demonstration software

After about half a day of investigation, It became clear that there was a big crater in the ebook software landscape. There is no obvious way to make an ebook from a Word document. Yes there are a some word processing packages which will do it, but nothing simple or online. Liza Daly – the ebook expert on this project confirmed this so I set out to at least try to fill that gap. My starting point was the ICE software I established at the University of Southern Queensland, where I worked until very recently Over the years this has evolved into a very capable word processing format converter for the web. It is mainly designed to work with documents with a known input format, using its own set of styles, but it can also deal with generic documents with standard headings. It already had some EPUB export but only for collections of documents, not for a single word processing file via a simple web service.

The idea was to use ICE to generate HTML then write code to break it up and store it a zip file, EPUB style. ICE does a reasonable job of converting the test thesis to HTML by looking for standard heading styles (Heading 1 .. n) and guessing that 'Quote' is a style for quotations. There are no widely used standards for word processing styles, so some heuristics are required; the current ICE code does not do a lot of this, but it could be extended.

What I did was:

•Looked at building on work already in ICE and other USQ software that makes EPUB files. I got something running, but in the process of developing it, I discovered the command line features in Calibre were probably a better option.

•Set up a possibly temporary code repository at Google Code and checked in the latest trunk of ICE.

•Added simple code to the ICE conversion service to call Calibre on the ICE-generated HTML. The command currently looks like this.

```
ebook-convert "/tmp/ICE-u6hayU.ice/Kingsley-Formatted PhD 12May09.htm"
"/tmp/ICE-u6hayU.ice/Kingsley-Formatted PhD 12May09.epub"  --chapter
"//*[name()='h1']"  --level1-toc "//*[name()='h1']" --level2-toc
"//*[name()='h2']" --title "The effect of scholarly communication
practices on engagement " --authors "" --publisher "" --pubdate ""
```

•Using Word docs there are some problems, like the way the title is truncated above, and other metadata is missing. I'll come back to metadata in a future post making sure that works are correctly described and labelled is an important dimension of scholarship. I see metadata as a gateway feature (as in gateway drug) for improving the semantic content of documents in general and I think once people see the power of embedding metadata semantics in their documents – so they don't have to retype stuff all the time – they'll be ready to deal with citations and rhetorical structure, and domain-specific semantics.

•Set up a test/demo server in Amazon's cloud services. Here's Danny's thesis as rendered by the service. The input document was a word document, the output is a EPUB file. It takes about ten minutes to create.

- Try it out with a couple of sample files (these might change during the project as we get more ambitious) by uploading the documents into the service or pasting the URL into the form.
  - A word document using a sample thesis template from the University of Edinburgh: http://dl.dropbox.com/u/24994372/thesis-test-document.doc
  - An ICE-version of the same thing in OpenDocument form using the ICE style-set with a few more examples of formatting, such as bullets and pre-formatted text. http://dl.dropbox.com/u/24994372/test-thesis-ice-styles.odt

The test service uses ICE to convert .doc and .odt documents to HTML – you can feed it generic documents using Heading styles and it will do its best. ICE deals with all sorts of images and maths etc., it does a better job than most processors because it uses the OpenOffice.org word processor to create images using its rendering engine – most tools can't deal with some kinds of content because they can't render them. Then the service calls Calibre's HTML to EPUB conversion tools. To start with I coded this to treat `<h1>` elements in the HTML as the major sections in the document. This contrasts with the way Calibre attempts to generate HTML directly from the `.odt` file, which means that lots of graphics won't be able to be rendered at all.

Status:

- Alpha code only.
- Might break.
- Slow (circa 5 minutes for a big document, though, not 100-ish)
- I will endeavour to keep this up and running for the life of this project.

Later in the project I plan to work with Theo and his focus group users to see whether the ICE approach for styles is viable for thesis production and for the various other use cases. Other questions include:

1. Should long documents such as theses be managed as multiple parts or single documents?
2. What's the demand for being able to integrate other kinds of resources (spreadsheets, images, other data like chemistry) and are there viable ways to embed this stuff in EPUB documents? The question here is whether EPUB is suitable for the 'research object of the future' where publications are not just documents, but need to be embedded in or carry-around more of their context.
3. Should we be thinking about EPUB converters as part of repository deposit processes?

Known issues with the ICE/EPUB generator:

- Sometimes he first section of the document appears in the TOC as 'unnamed'. I don't want to resort to hacks like adding a heading called 'frontmatter' to the document, but I might.
- I have been focussing initial testing on the Firefox addon for EPUB: https://addons.mozilla.org/en-us/firefox/addon/epubreader/ more testing and validation required (this goes for all the tools I have been looking at.)

Where next?

Potential next steps include:

- Improving and glamorising the test conversion service.
- Creating new calibre plugin to use ICE as an HTML generator – this could be used to add Word support, and potentially improve Open Document support.

Lessons learned

Even if you start from good quality HTML, writing an automated tool for creating EPUB is likely to be quite a big job. Calibre already does it, has lots of configuration options and seems to be a good starting point for a conversion tool – I'm interested in comments about this though. A couple of days of my own coding didn't get me close to what I could do with Calibre in about an hour learning what command-line options to use (there were several hours spent refining the initial options, though). And it's not just me, others of the staff at USQ have spent a fair bit of time getting basic EPUB support working; the one thing that makes me think it might be worth tackling is the glacial speed of Calibre – but that might just be because Calibre is doing all sorts of important things and a new converter could be just as slow.

[This is a repost of http://jiscpub.blogs.edina.ac.uk/2011/04/05/epub-for-word-processing-users/ - if you have comments please make them over on there on the jiscPUB blog.]

This post was written in OpenOffice.org, using templates and tools provided by the Integrated Content Environment project and published to WordPress.