ICE to DocBook? Yes, but I wouldn't bother.

Following my last post on the demise of Google Wave and the future of scholarly word processing, Anthony Hornby of Charles Darwin University asked:

> @ptsefton Met Cameron Loudon here @CDU, he mentioned ICE. Could any parts of that system allow a simple path from docs to DocBook XML?

DocBook, XML, and the cost of 'doing it right'

For those of you who don't know.

> **DocBook** is a semantic markup language for technical documentation. It was originally intended for writing technical documents related to computer hardware and software but it can be used for any other sort of documentation.
>
> As a semantic language, DocBook enables its users to create document content in a presentation-neutral form that captures the logical structure of the content; that content can then be published in a variety of formats, including HTML, XHTML, EPUB, PDF, man pages and HTML Help, without requiring users to make any changes to the source.
>
> http://en.wikipedia.org/wiki/DocBook

Which sounds like a good idea for documents – getting all those formats for free, right?

It is, but you have to take into account the cost of creating the documents, inducing the authors to capture the semantics, and providing tools for authors that they will actually use. (And keep in mind that a lot of those output formats are actually HTML based: HTML, XTHML, ePub HTML help, more on that later).

Back in 2004 I used the analogy that XML is like a hill – and presented that idea at Open Publish. The message was that if you manage to climb up high enough then generating output formats like HTML is a downhill slide. (Or a semantically rich document is like a tightly wound spring, ready to unleash its stored energy).

> *The XML as hill metaphor, together with the author as cyclist metaphor* to talk about the importance of picking the right tool, and the right template/schema for your authors. You can get them to pedal up hill (ie encode potential energy into their documents) if they get a small reward; a downhill coast. That is, they will follow a standard if they can see the point, such as a web-site of their course building before their eyes. They will not cooperate if you ask them to climb too steep a hill for no apparent reason.

As I noted in my last post, when USQ tried to get academics to climb a steep hill with the GOOD system, they simply wouldn't do it (and nor would I have). We ended up with a system with a much flatter gradient, ICE.

Anyway, back in the present, to Anthony I said:

> @ahornby SHort answer is yes. Longer answer (that HTML is all you need) coming via blog soon
> 1:14 PM Aug 7th via web in reply to ahornby

Now, the promised long answer.

The facts

ICE templates can be used to make DocBook documents. Dr Ian Barnes did the work for this back when he was with the Australian Partnership for Digital Repositories, APSR for a project called "Digital Scholar's Workbench". The final report gives some more detail, including future plans for development that didn't happen.

Ian wrote fiendishly complicated XSLT stylesheets that could turn open office documents (and by extension Word documents loaded via Open Office) into DocBook – assuming the use of a set of styles, which is where ICE came in.

There have been other systems that attempt to use word processor styles to create DocBook documents, notably a software application of legend from O'Reilly called Aardvaark which I wrote about here[1]. You can find other half-hearted attempts littered around the web.

My opinion

So having answered the question "can you" with a yes, I want to say, "**but don't**".

If you have a set of word processing documents that have enough structure in the form of Headings etc to even consider something like DocBook as a format then you can use ICE or an equivalent system (if such a thing exists) to:

1. **Make PDFs** of them and push them out on the web by using the word processor to make the PDF – no loss of formatting as you try to deal with edge cases that might not fit into DocBook. (One of the things that Ian had trouble dealing with is that lots of word processing documents use a mixture of numbered and not-numbered headings. That's a no-no in DocBook.)
2. **Turn them into HTML** and weave them into the fabric of the **web**.
3. Make **other formats as required**: eg ePub for mobile use, IMS packages for educational use. (Both of these are HTML derivatives).
4. **Some kind of package format** for deposit in a repository with HTML, PDF and the rest. (This is work in progress, but DocBook won't help).

Given that both Word and OpenOffice.org now use well-specified XML based formats bundled in Zip files, I would assume that the range of formats mentioned above takes care of preservation. What else is DocBook going to give you?

The one thing I can see is that DocBook can be turned into PDF using standard, free, tools. That's reasonable, you might want to use it as part of a rendering system. But don't forget you'll be dealing with all the things that might be embedded in those word processing documents, charts, diagrams, complex tables with very important formatting like double borders in the right places, and best of all, **maths**. Sure, DocBook can do mathematical typesetting, but you have to get all the maths out of the word processing docs and into DocBook you have to get it **right**. This is hard and there are no right answers, only compromises. My team at ADFI continue to work on these compromises for the Integrated Content Environment, ICE.

[1] That post is why one of the most popular search strings that brings people to my blog is "long tongue" and variants thereof. Oops I did it again, frustrating the people who really want a longer tongue or want to look at pictures of other people with long tongues, maybe both. If I ever get sick of discussions about metadata or whether WordPress is a good tool for writing a PhD I'm definitely going to write a guide to longer tongues. Who knows, maybe O'Reilly will publish it with an aadrvark on the cover. I am an O'Reilly author you know. I have a photocopy of a cheque signed by Tim himself here in my office for FOUR HUNDRED AND 0/100 DOLLARS. They didn't make me write in DocBook, though, for my XML.com articles, they wanted HTML. Just saying.

You get what you pay for / get back the effort you put in

But wait! I can hear you saying, DocBook is much richer than HTML!

So it is. But in my considerable experience, the sweet-spot for using a word processor is around the use of headings to create a document outline, some sanity in lists using our very carefully designed set of styles, plus blockquotes, preformat text. That's about the extent of what you can collect from large groups of authors who are not obsessed with document structure and semantics. Go after those simple things and you will get usable documents. Go after something that attempts to layer a few hundred elements of DTD over a flat word processor document and you'll get nothing, as in nobody will comply and you will have to add all the structure to the DocBook later. That's fine if you have the business case for such activities. I know I don't.

If you want to do more than simple basic structure then a word processor is not the right tool, you might as well use an XML editor.

But then you won't get a large general user base using an XML editor.

At this point we have come full circle so I'll leave it at this. Word processors are good for capturing simple structure and semantics and that turns out to be all you need in many cases. If you do want more then HTML has a way to do semantic extensions. It's called the class attribute. (And we're exploring ways that more semantics might be layered into Word processing documents using a kind of word processor equivalent to RDFa).

But I **still** want DocBook!

Like I just said, the structure you get from an ICE-like word processor document after it has been converted into DocBook is **exactly** the same level of complexity as the HTML we generate from ICE: Headings, lists , quotes, preformat.

For the final word on why you don't need DocBook, I refer you to Mark Pilgrim who authored a wonderful book in DocBook but abandoned the DocBook format for plain-old HTML in the latest edition.

Summary

So, the main points I have made are:

- Yes you can make a kind of cut-price DocBook from ICE, or any reasonably structured word processing file, but it won't be getting the best out of DocBook.
- In situations where source documents come from word processors I don't see the point in adding another format to the tool-chain unless you (a) really need to produce PDF in a variety of formats from the source text (rather than just automating a 'Save as PDF' or (b) you are going to add value by using DocBook or similar later in the editorial process and you have the resources.

  But note that there are other options for generating PDF/Print from HTML including HTML to PDF XSL-FO stylesheets or using (non-free) Prince. I predict that this will be a continued growth area.

But, If Anthony Hornby really wants to have DocBook, and I suspect he does, then what could he do?

Me, I'd talk to Ian Barnes, who is now residing in Scotland and **is available to consult and/or program** – he's the localhworld expert on ICE-to-DocBook and has explored many of the options for rendering to PDF (including from HTML).

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