Beyond the PDF: Some ideas for document formats and authoring tools

This post has been sitting my work folder since Beyond the PDF in January. I'm going to post what there is of it now, as background to the Scholarly HTML hackfest that's on in Cambridge this weekend.

In the lead up to Beyond The PDF there were lots of ideas flying around about authoring tools, and associated document formats. People have mentioned Word and OpenOffice.org, ICE, Drupal, Plone, wikis, Lemon8-XML, TeX/LaTeX, XML schemas like NLM XML and dexy.it. The list goes on. These things are a bit hard to compare. It's not apples and oranges so much as tropical fruit salad. In that last sentence, there were editors, combined editor-content-management systems and document formats, at least one of which (TeX) is even Turing complete, meaning you can write computer programs in it (but not, alas, always produce web pages from it).

I thought it might be worth making a start on mapping some of this space. In this post I want to look at authoring tools, and document formats – independently of the other functions such as publishing and peer review. We need to look at both formats and tools.

I have made the case in the past for using HTML as a **format** for scholarly documents, proposing Scholarly HTML. Martin Fenner has also porposed that we “Use HTML” - I agree that we should,:

> All this would be much easier if we just used HTML. With HTML, authors, publishers and readers can all use the same document format. And they will have an endless number of tools at their hands, including of course WordPress for writing and the web browser of choice for reading. HTML in 2010 is very different from HTML in 1990. HTML5 supports new semantic elements such as <article>, microdata, embedding of video without plugins, geolocation, and offline web applications.
>
> http://blogs.plos.org/mfenner/2011/01/05/html5-or-messages-from-beyond-the-pdf/

But, the problem is that tools that create scholarly-quality HTML5, complete with article tags and so on don't exist at present. And if we were to set out to create them from scratch, I see a very long road ahead. Look at all the things that are built into a word processor like MS Word, which were added over many years of development. There's outlining and an outline view for document architecture, auto-text for inserting frequently used text, a very complicated table editor, built in vector graphic drawing, data-integration with spreadsheet functions and graphs (that's data-aware publishing right there), document navigation by headings, objects, tables etc, captions and numbering systems for embedded objects, cross referencing by document structure, by bookmark, user definable semantics via styles, revision control, footnotes and endnotes. This is all stuff that was added in response to real user feedback, at least in the first couple of decades of word processor development, after which time many of us think things got worse, not better.

I really think we need to ask a few questions:

1. Who can afford to build all that again? Not small teams like mine. As Philip Lord put it on the mailing list:

   > It's a busy application space. My worry would be that anything added into wordpress would just be the poor relation of these tools [he mentioned Google Docs, Live Writer, Subversion]

2. Given that we have authoring tools like Word processors which many author know and like, can't we find ways to create HTML (or whatever other formats are endorsed by the Beyond The PDF workshop) from word processors?
3. If we are going to invest in tools like a WordPress web-based editor it would make sense to try and make work as portable as possible so it could be used with other CMS platforms.
4. Don't forget wiki formats – it would be useful if we had one wiki format to support research, maybe

> along the lines of the apparently stalled Wiki Creole – or built on one of the major wiki formats like MediaWiki or MarkDown.

I think it's pretty clear that no one approach to authoring is going to 'win' out of this week's workshop, so what we should be looking for is a way to get the best possible interop between authoring tools and document formats.

But before we look at potential projects like Scholarly HTML editors it's worth thinking about what it is that we're authoring. That is, what's the abstract Model? This work, mentioned by Anita de Waard, who has a very cool job title, *Disruptive Technologies Director, Elsevier Labs* was new to me:

> So, in the W3C Health Care and Life Sciences group we are trying to come up with an 'ontology of rhetorical blocks' that does not only work for biology - and incorporates EPub, PRISM, BiBo, FRBR, and other bibliographic standards:http://esw.w3.org/HCLSIG/SWANSIOC/Actions/RhetoricalStructure
>
> Currently a first pass of an 'Ontology of Rhetorical Blocks' is out- http://esw.w3.org/HCLSIG/SWANSIOC/Actions/RhetoricalStructure/models/blocksontology is out, and we are working on a 'medium-grained' model - http://esw.w3.org/HCLSIG/SWANSIOC/Actions/RhetoricalStructure/alignment/mediumgrain
>
> Paolo Ciccarese, Tim Clark, Jodi Schneider and I are all working on this, and very much invite comments, contributions, and discussions at or before the San Diego meeting.

This is important background because it opens the way to layer a formal semantics-of-rhetoric on top of any format and the work that has gone into this is key input to the design of any new format.

Potential mesh of services

I have some opinions and assumptions about authoring tools, their strengths and weaknesses, based on long experience working on single source web publishing systems in the academy. I'll put down some of the assumptions here; happy to have my opinion changed if anyone wants to take that on. My comment form awaits your input.

Assumptions:

- Full schema-validating authoring tools (eg enforcing ordering constraints such as 'intro before method') are suitable only for trained editorial staff. With this class of tool, simple operations like splitting your intro into two sections can be quite complex. I have never heard of this class of tool being used successfully with a large number of loose affiliates like researchers (I've said that quite often here and I can't recall anyone turning up with an example, but as I said, the lines are open).
- A new class of cut-down editor – probably running in the browser – which only allowed you to produce sensible documents, is probably doable.
- Regarding word processors:
  - There is no magic system that can turn sows ears (any old word document) into silk purses (like, say NLM compliant XML).
  - In many disciplines, lots of authors will use their word processor to type stuff and try to paste into your system no matter what system you ask them to use.
  - Giving authors a word processing template with some set of styles and-or macros and expecting them to use it to produce structured documents doesn't work.

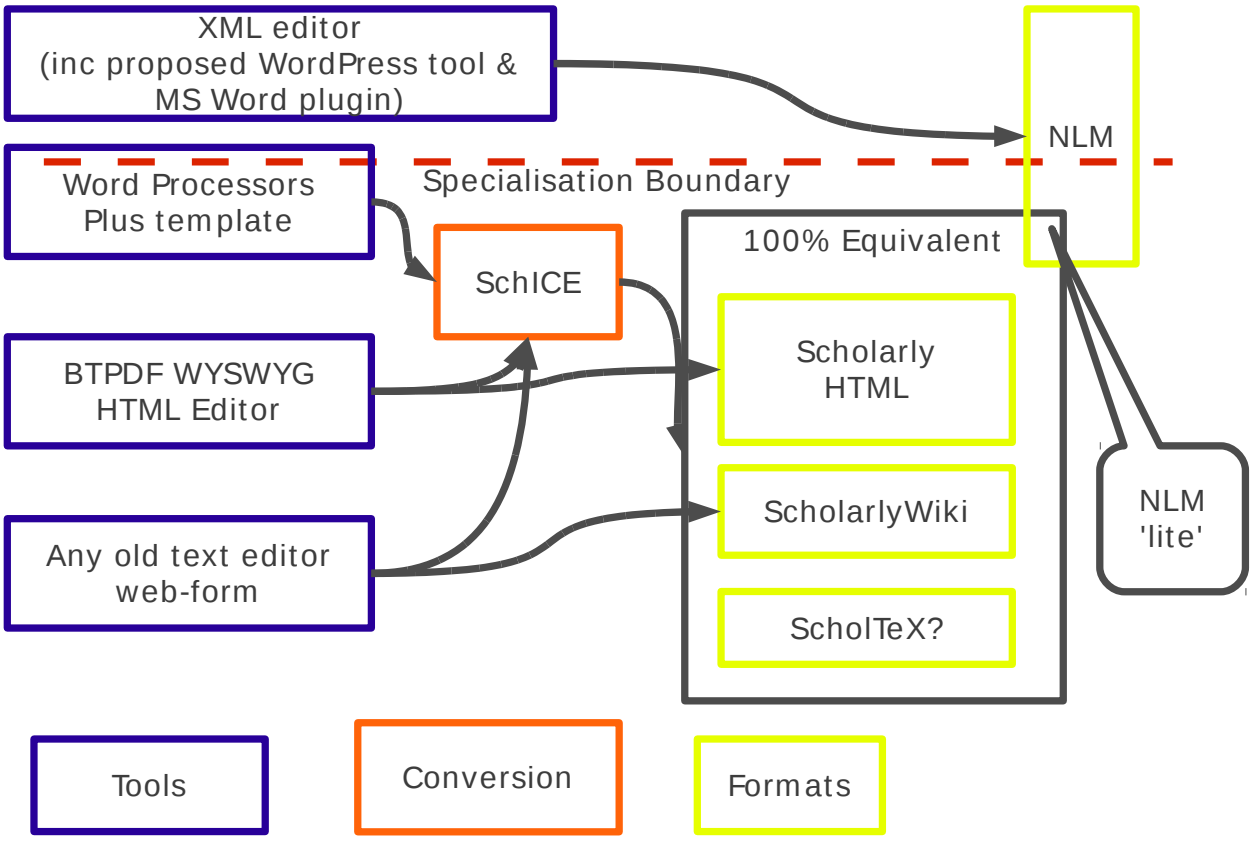

Figure 1: A map of authoring tools; 'mass market' tools below the 'specialisation boundary' line, 'pro' tools above.

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.