Beyond the PDF proposed session : mainly on authoring tools

Peter Sefton

Australian Digital Futures Institute, University of Southern Queensland

I'm posting this as a kind of extended abstract of my proposed presentation at the Beyond the PDF workshop. I want to demonstrate some of the services we have build in my group but more importantly I'd like to discuss what researchers and publishers would like to see, and find out how they react to some of the things we've done and the reasons we've done them, and how these might help out with a post-workshop project.

Abstract

This presentation touches a number of the workshop topics. It will take the form of a demonstration, of systems developed at the Australian Digital Futures Institute for scholarly workflows. It is intended to spark discussion about issues and challenges in taking Scholarship beyond the PDF. It covers the following tasks: (a) managing draft documents and local data sets together, (b) formatting draft documents with as much semantics as possible, using a word processor, so they can become part of a rich human and machine readable web of research practice and (c) pre-publication collaboration with immediate collaborators and via the web using annotations using systems that have potential for use post-publication as well.

These tasks are demonstrated against a background of four guiding principle that inform the work. To make **all resources part of a repository** as soon as they are created or acquired. To **provide a web view** of all resources as early as possible in their production process. To provide a hub from which **resources can be pushed to other services** – journal review processes, blogs, repositories. And to make **interoperable, reusable services, not monolithic systems**.

Introduction

In this post I will run through a tour of some of the software we have built in my team. First, the credits. The work I'm talking about here has had substantial contributions from these people over the last six or so years:

- Bron Chandler
- Daniel de Byl
- Duncan Dickinson
- Pamela Glossop
- Oliver Lucido
- Greg Pendlebury
- Ron Ward
- Cynthia Wong
- Jason Zejfert

Proposal

User community?

The community we're targeting is pretty much any researcher, with an emphasis on those who use word processors, and particularly the ones who are not being helped out already with workflow tools for their disciplines, but the systems I am talking about extremely extensible, and can easily be configured to render XML, or LaTeX or whatever.

User tasks

The tasks I'll cover here include:

1. Managing draft documents and local data sets together.

2. Formatting draft documents with as much semantics as possible, using a word processor.

3. Pre-publication collaboration with immediate collaborators and via the web using annotations.

Current solution & issues

The current solution for many teams is an ad hoc process involving mailing around draft documents. There may or may not be an institutional document or content management service. Many teams use heterogeneous tool sets – for example not every author has access to the latest version of MS Word at all times. Research data management is also very often ad-hoc and it's uncommon for researchers to have the means to link documents to data sets and to manage those relationships through to publication and beyond – although we are starting to see spectacular progress in some areas with projects like My Experiment and My Tardis.

Proposed improvements

I will go through some of the improvements to ad-hoc workflows used by researchers in the areas I outlined above, using our tool The Fascinator. This tool can be configured in lots of ways – here I'm going to look at it on the desktop and running as a team intranet server. The Fascinator Desktop is a Java application that can be installed on a user's computer. Like consumer tools such as Picasa for images or iTunes and Winamp for music, It indexes files and provides a faceted browseable view of them. The team edition is the same thing installed over a file-share. We are trying both to see what works.

There are four key aims:

- To make **all resources part of a repository** as soon as they are created or acquired.

- To **provide a web view** of all resources as early as possible in their production process. A major motivation for this is that the web provides a platform for doing a new kind of research where data are available and linked to publications.

- To provide a hub from which **resources can be pushed to other services** – journal review processes, blogs, repositories etc (if we go Beyond the PDF then these things might all become part of one repository-journal-blog-thing).

- To make **interoperable, reusable services, not monolithic systems**, so the stuff I will demo here for formatting documents, annotations, adding semantics is all designed to be used not just in our systems but in others as well.

I'll demonstrate what this looks like using the sample files from Beyond the PDF.

Improvement 1: Use styles and other techniques to make documents as useful as possible

The workshop organizer provided an early draft of an article to work with in Word format and some other sample files – I gather this is known as the Bourne Corpus.

The document is a good example of a typical word draft of a research article. On the plus side, it uses styles - the title is in Heading 1, and the major headings are in Heading 2. So far so good – but there are a few issues, notably some paragraphs which are in Heading 1 style, but formatted to look like text, and couple of headings in a style that looks like it is intended for captions.

This is why we have this aim: to **provide a web view of all resources as early as possible**. If the authors were working in a system that showed them an HTML version of their document as they went, then formatting problems would show up straight away. To illustrate this, take a look at this screenshot that shows the document after it has been converted to HTML by the Integrated Content Environment (ICE) service that is used by The Fascinator[1]. A couple of body paragraphs are showing up in the TOC because they have the wrong style applied. Given that articles like the sample are going to end up in PubMed in HTML – why not be able to preview that right from the start?

- Abstract
- Introduction
    - The network construction and analysis of molecular interactions provide a po
      system and to reveal hidden relationships between drugs, genes, proteins, a
      would facilitate improving polypharmacology and rational drug design2. In re
      developed to predict the drug-target network based on ligand chemistry3-5,
      combination of chemical features of drug molecular and sequence features of
      computational evaluations have proven that these methods are extremely va
      However, these methods are biased to annotated drug-target pairs, which m
      as G-protein coupled receptors (GPCR). They only cover a small portion of hu
      severer in pathogens than in human. For example, among 3,999 encoded pro
      nine proteins (cmaA1, cyp51, embA, embB, embC, folK, InhA, katG and rpoC
      target network that is reconstructed from most of existing drug targets is mo
      druggable targets and predicting off-target profile of new compounds on a pr
      target data bring up questions whether or not the topology of the drug-targe
- Results
    - A drug binding site database
    - Drugome-TB: a reliable and unbiased protein-drug interaction network
    - Drugome-TB is a scale-free and modular network
    - Highly connected proteins are potential druggable targets
    - DISCUSSION
    - Concurrent vs linear drug discovery process
    - Conventional drug discovery and development proceeds as a linear process fr
      optimization, to preclinical and clinical trial. It is estimated that above 90% o
      development mainly due to poor drug efficacy or safety54. If the information

The point? With an automatic web preview an author can easily see here where they have a structural issue and correct it. Using the templates we provide for ICE, that would be a matter of locating the offending

---

1   I have done some processing on this document – which was mainly just search-and-replace to turn the styles used by the authors into ones supported by ICE; Heading 1 → Title, Main text → p etc. ICE uses a specific set of styles which has been refined over many years to cover most generic structural needs – if there is some interest in using ICE-like techniques then the style names up are up for negotiation with a standards committee. I also used OpenOffice.org rather than Word because I'm running Linux on this computer, but I note that it does not really seem to be improving lately – there has been a fair bit of regression with recent upgrades, it's almost getting so it's worth the pain of firing up Word in a virtual Windows machine.

paragraph and clicking the leftmost "P" (for left-aligned paragraph button) on this toolbar:
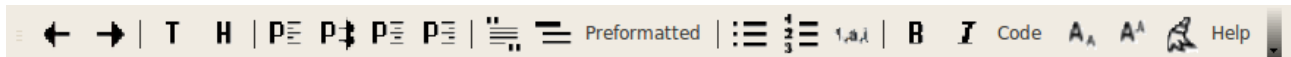
This toolbar is something we maintain for both OpenOffice.org Writer and Microsoft Word. Clicking buttons on the toolbar doesn't just format the document, it applies styles, whether you like it or not. Styles are important as they can be used to structure the document into sections, and potentially to add semantics (if you can train the authors). You can get the toolbar from the ICE site. If you want to try out your documents you can convert them to HTML, PDF, Dublin Core, etc here. Try it with the paper, which I have loaded on to the workshop website (in ODT format for now).

This approach is quite different from others:

- Microsoft have produced a Word (2007 plus) Add-in that allows authors to create rich XML to the NLM DTD. This is much more specialised than our tool.

- The PKP team have a product called Lemon8XML which tries to create rich documents without using style information at all.

I am very interested in working with others on ways to target particular formats from mainstream authoring tools including Google Docs and web-based editor like those found in content management systems, and particularly interested in doing so in as interoperable way as possible. A discussion on this is one of my big hopes for the workshop.

Now, to talk about the semantics of the document. One of the most fundamental parts of scholarship Beyond The PDF is having good metadata. So, I used a simple technique (Sefton et al. 2009) to label the abstract of the document – I styled the abstract using a style named `p-meta-abstract`. An author would not need to apply this, it could be built in to an article template.

The ICE service can extract this metadata and provide it in Dublin Core:

```
<dc:description xmlns:dc="http://purl.org/dc/elements/1.1/">Proteome-
wide analysis of protein-drug interaction network on multiple
scales, ... the methodology is ready for other pathogen genomes.
</dc:description>
```

But as I said in a recent post this use of styles is a bit fragile and hard to manage, for anything more complicated than an abstract, so we are working on other ways to embed document semantics. In that post I wrote about how to encode metadata about authorship inline using hyperlinks. My team has done some more work on this now. To demonstrate on the sample paper this week I:

- Made myself a bookmarklet, AuthorIze. What this does is generate a URI that says someone is an author. (This is an early prototype – it does not work on URLs which contain parameters, at the moment).

- Went to the Thomson Researcher ID site and looked for IDs for the authors. Only the last one, Bourne seems to have one. I could have just linked his name to his page at Researcher ID but that's just a link, it does not express authorship.

- I clicked my new bookmarklet, which gave me this machine parseable URL: http://ontologize.me/?tl_p=http://purl.org/dc/terms/creator&triplink=http://purl.org/triplink/v/0.1&tl_o=http://www.researcherid.com/rid/C-2073-2008

Then I used that ungainly URL to link Phil Bourne's name in the author list. This process is not completely horrible – it's pretty standard linking, but it could be made much better with plugins for authoring tools.

Result?

The ICE conversion service now recognises that http://www.researcherid.com/rid/C-2073-2008 AKA Philip Bourne is an author:

```
 <dc:creator xmlns:dc="http://purl.org/dc/elements/1.1/">Philip E.
Bourne &lt;http://www.researcherid.com/rid/C-2073-
```

```
2008&gt;</dc:creator>
```

And the ICE service has added RDFa (v 1.1) to the page which means that it should be possible to auto-discover the relationship.

```
<span rel="dc:creator" resource="http://www.researcherid.com/rid/C-
2073-2008"><span property="foaf:name"
resource="http://www.researcherid.com/rid/C-2073-2008"><a
href="http://ontologize.me/?
tl_p=http://purl.org/dc/terms/creator&amp;tl_o=http://www.researcherid
.com/rid/C-2073-
2008&amp;triplink=http://purl.org/triplink/v/0.1">Philip E.
Bourne</a></span></span>
```

But wait – there's more! We can do the same trick with document semantics. I have made a similar bookmarklet for mentions of a document subject-matter inline. This idea is to enable markup similar to that used by Microsoft Word Ontology Add-in, and to a text mining tool called Whatizit both of which automatically scan a document for terms and match them to formal ontologies. In the version of the MS plugin we looked at a couple of years ago, though, it was simply linking the ontological term to the text but not saying what the relationship was – in other words, imparting exactly as much information as a simple link to the ontological term.

I asked on the list for an example where linking a term to an ontology might be useful for disambiguation, in the sample paper, and I got this from Tudor Groza explaining that katG means two different things:

> On Mon, Dec 6, 2010 at 6:50 PM, Tudor Groza <email-removed> wrote:

> \> Hi Peter,

> \> Not sure if this will help ... you could look at the term _katG_ on page 4

> \> in the FinalPaper.pdf (8 lines from the end of the page), which can act as a

> \> vaccine (http://purl.obolibrary.org/obo/VO_0012369 - DNA vaccines encoding

> \> KatG antigen ...), or in that particular context as a protein

> \> (http://purl.obolibrary.org/obo/PRO_000023043 - A protein that is a

> \> translation product of the katG gene or a 1:1 ortholog thereof.)

> \> Regards,

> \> Tudor

So in the demo document I marked up the term katG with this link that says it that it is the subject of the document: http://ontologize.me/?
tl_p=http://purl.org/dc/terms/subject&triplink=http://purl.org/triplink/v/0.1&tl_o=http://pir.georgetown.edu/cgi-bin/pro/entry_pro?id=PRO_000023043.

As with identifying authors, it should be possible to embed lookup widgets in the authoring environment.

The point of this is that this information is now embedded in the document, and my desktop repository can extract it:

METADATA

Title

Proteome-Wide Polypharmacology Drug-
Target Space of Mycobacterium tuberculosis

Creator

Philip E. Bourne

Description

Proteome-wide analysis of protein-drug
interaction network on multiple scales, from
atomic details of molecular interaction to
global topology and systematic behaviour
may provide practical solutions to... *(more)*

I have talked about how our services convert documents to web pages, but it is not only word processing documents we deal with – there is an extensible collection of conversion plugins that work with different formats. I'll have a quick look at a few here.

Here's the data spreadsheet from the sample files has been rendered for the web automatically by the Fascinator, which called the ICE conversion service:

Overview

Mtb protein info

Drug site info

SMAP results

Homology model info

Homology results

Sheet 1: *Mtb protein info*

Gene	Rv number	Protein name	PDB codes
aac	Rv0262c	aminoglycoside 2'-N-acetyltransferase AAC (AAC(2')-IC)	1M44, 1M4D, 1M4G, 1M4I
accD5	Rv3280	propionyl-CoA carboxylase beta chain	2A7S, 2BZR
acpA	Rv0033	acyl carrier protein AcpA	2CGQ

The ICE conversion service just processes the spreadsheet as a table, but for known kinds of data we could add a plugin that did useful things with the data – in particular we could work out conventions for linking from the paper to the spreadsheet. So the opportunity here would be to host services for scientific workflows in the desktop or team repository environment rather than in a word processing document (Mesirov 2010).

Our conversion services also deal with a wide range of common media types:

• PowerPoint is rendered as a series of images (lots more work to be done on this to break decks into slides and let people re-purpose them and re-publish new slide shows). Here's one of the sample files rendered for the web.

- Images: the system generates thumbnails – this is highly configurable.

- Video: we have worked on scripts to convert common video formats to web-ready files at multiple sizes, and for the key platforms including HTML 5 and iOS devices. I'll come back to images and video below in the section on annotations.

- Scientific: we've done proof of concept work on adding visualisations for stuff like Chemical Markup Language, including a project with Peter Murray Rust's group at Cambridge to handle the packaging and delivery of web-ready chemical theses (Sefton & Downing 2010). There is a demo on the ICE site. The way this works is via simple link in the original document to the CML – when the ICE service sees the link it inserts the viewer. This idea could be generalized for all sorts of visualization and workflow tools very cheaply.

- Check out the current supported format conversions at our site.

Improvement 2: Think of all resources including data and documents, as being on the web and in a managed repository from birth

I have looked at how our platform works to provide web-ready views of all kinds of files. One of the key challenges, though is to work out how to provide that web view of things that are not, well, on the web yet. Our current approach – and this is very much work in progress – is to look at providing services that 'watch' a file system or content management system and convert materials as they appear.

We're looking at this from a number of different 'altitudes'. First up I will talk about work we're doing from the 10,000m level, with the University of Newcastle and other collaborators on an institutional approach to managing research metadata. The drivers are compliance requirements that data be managed so that research is reproducible and a commitment from the Australian government to make as much research data available for re-use as possible. Here we are building a system which will 'watch' what is happening at an institutional level – grant funded project opening and closing, and data sets being deposited on the university's research storage facility. This is an institutional Repository approach to managing research data. This work is being documented on a blog; at this stage we have a pilot system up and are refining the metadata model with the stakeholders.

Back to the ground-level view, here's how the sample corups from the workshop looks in The Fascinator Desktop software running on my laptop. You can see a demo version of the same software on the web. Note that we're interested in supporting all kinds of research, including the humanities and our main collaboration is with a historian, Leonie Jones (Dickinson & Sefton 2009).

We have already seen how the system provides web (and PDF) views of all these materials. One of the other important things is being able to relate items to each other and package them. Using The Fascinator I can group objects together, order them and arrange them into a hierarchy. Here's a screenshot of a package I put together with the sample files – the table of contents for the package is at the left and the document is on the right in HTML, with a link to a PDF version.

This packaging is done using a lightweight toolkit we call Paquete – it's designed to do two things:

1. To make implementing packaging systems in web applications easy, via re-usable interface code.

2. To provide an pure-web way to read packages – see our demo. The hope here is that HTML 5 browsers will be able to save packages like this using 'Save as App' – giving us a way, at last, to package rich sets of material, including potentially workflow-engines in a way that can be moved around.

So I have organised these materials. Now what?

The point of this is to be able to send that package off elsewhere.

For courseware we export as IMS content packages (from ICE). For mostly-text materials, ePub and for research outputs, we would expect to push the content to a journal or a repository or a mashup of both. We've done some work on SWORD deposit to repositories, including the ICE-THEORM project (Sefton & Downing 2010) but no researchers have shown much interest apart from ourselves, and we're an R&D group, with a bit more emphasis on the D than the R – so no SWORD support at the moment, but it's almost done and ready to turn on if anyone cares.

Anyway, to demonstrate what's possible, I have made a package of the paper, the presentations and the data spreadsheet and do two things with it.

First of all I made the trio of files into an ePub eBook. Epub provides a single file that can contain anything – it must have a basic HTML version of the content, but could also have the source files. Here's a screenshot using a Firefox plugin to read it – please try it out for yourself and let me know if it works for you:
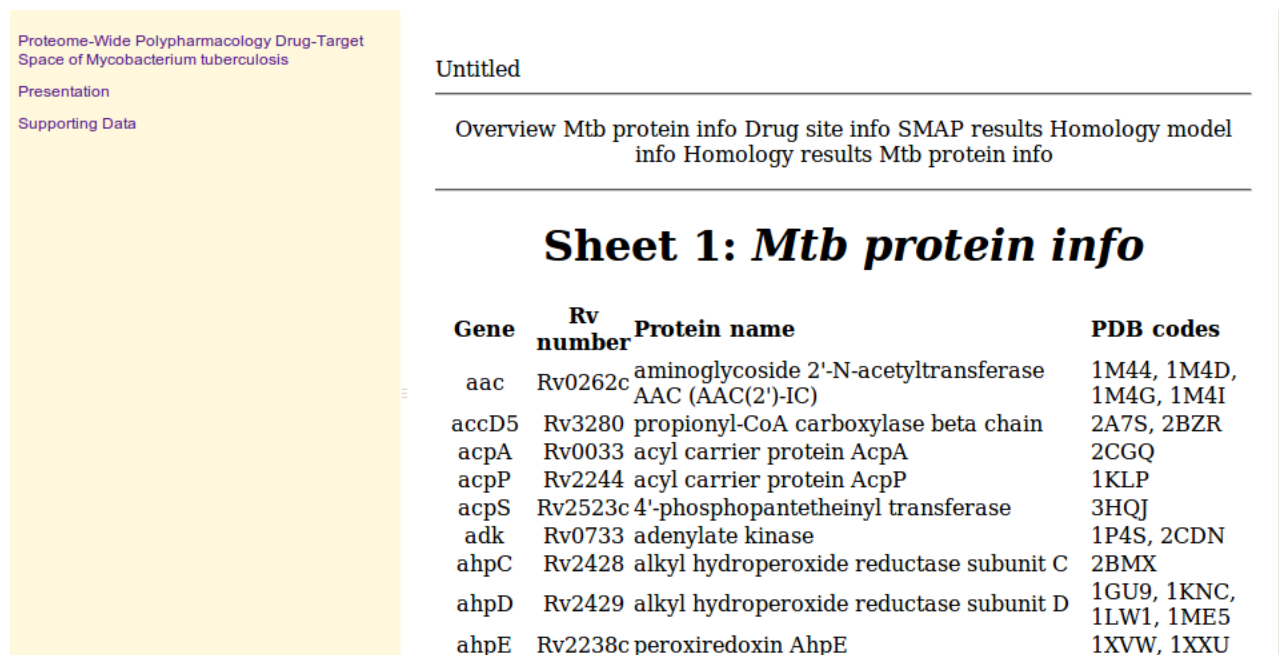
(To make the ePub I used ICE – the ePub packaging in The Fascinator is not quite ready but it is under development)

The other thing I did was push the same content to a blog, from The Fascinator, in this case my test blog for the Anotar plugin in WordPress. For blogging, the software flattens the parts of the package into one long post (without a TOC-even - possibly not the optimal behaviour, feedback wanted). The test post is up here complete with the draft paper, supporting data spreadsheet and a PowerPoint all in one post – OK so that's a bit of a stunt, but **what would be useful to you**?

(I'm looking forward to discussing the use of WordPress with Martin Fenner who has been working on WordPress as an authoring tool. One of the frustrations I have with WordPress and most other content management systems is that they are not designed to deal with stuff like compound documents (in WP, images don't belong to a document explicitly for example, they live in an uploads directory) and multiple renditions of the same thing, or data sets. It's a constant tension between wanting to work with ubiquitous successful systems and wanting to engineer new ones that are better architected for research workflows[2].)

And, guess what? That test post is annotatable. Go and see if there are any encoding errors in the text and point them out (I bet there are). Which brings us to the next section.

Improvement 3: Provide rich annotation services cheaply and in an interoperable way

One of the other things we want to do with all these tools is to make everything annotatable. Our focus is on making it easy to add annotations to a web application, not on providing general web-wide annotation services. We are very keen to build on existing work on annotation software, ontologies and protocols where they fit with our goal of developer and user friendly services.

---

2 I'm all for working with WordPress, I even got teased about it by another USQ staffer.

We have created a toolkit for annotation called Anotar. You can read about it on our wiki. The idea of Anotar is that it client-side browser code is portable, and you can either use a standard back-end on your server or store annotations in a host application. But, in order to meet our goals of having easy to deploy web software, we found that using RDF to talk between a web client and an annotation server made coding too difficult in the browser, so we started work on a JSON-based protocol for talking to the server. It is definitely in-scope to interoperate with services that use the Open Annotation ontology (Hunter et al. 2010) from the Anotar server-side.

My test blog is running an implementation of Anotar for WordPress (only a very early version ATM – I need to update it) it re-uses the Anotar code for the client but stores annotations in the WP comment database.

Sarah L. Kinnings[1], Li Xie[2], Kingston H. Fung[3], Richard M. Jackson[1], Lei Xie[4],*, Philip E. Bourne[2,4],*
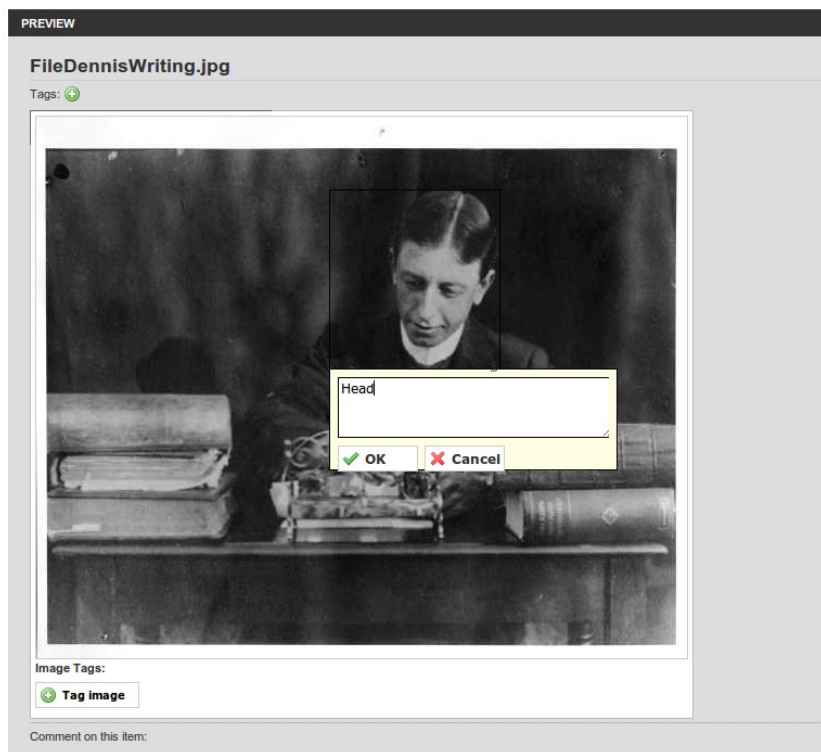
[Collapse 1 comments...]

[Reply]
Comment by: *admin*   less than a minute ago
    I couldn't find Researcher IDs for most of the authors.

The same client code is used in The Fascinator – so I can comment there too in a similar way. We have not done this yet, but it would be possible to bring together annotations made in the repository with those made on WordPress by linking up the back-ends. Also supported in Anotar and The Fascinator are:

Tags – both *ad hoc* and using ontologies.

Image region annotation

PREVIEW

FileDennisWriting.jpg

Tags:

Head

✓ OK   ✗ Cancel

Image Tags:

⊕ Tag image

Comment on this item:

Video time-span annotation.

Whatever YOU want: rating systems, peer-review voting, thesis examination, markup of chemical structures. What DO you want?

The paragraph-level annotation uses a neat system worked out by Ron Ward. It works by calculating an ID for each paragraph-type element based on its textual content, using a hash function. If a paragraph changes, as it does in an authoring-oriented system, then there is no longer a match, but Anotar stores the original paragraph text with the annotation, so you can still see the obsolete paragraph and its annotation at the bottom of the document. This can be done client-side in the browser, rather than having to process

documents on a server, making it easy to add to web systems.

Why it is better*

* This heading is from the conference organisers.

The tools I have touched on above – The Fascinator web-ready repository and the ICE content conversion service – have been built to address several gaps in the academic tool-suite. So in a sense, the things we're doing here are better than nothing – because without some of this work **nothing** is all there is for most academics:

1. Lack of a good way to turn word processing documents into HTML (or XML), and lack of standard style-name sets to assist in that process.

2. Lack of a way to embed useful semantics in documents being authored in non-specialist tools like Word or WordPress.

3. Lack of repository-services that reach back upstream to where the users are doing their research and authoring (there are lots of other groups working in this space).

4. Lack of software components that are easy for developers to use to build annotation

5. Lack of ways to packaged multiple resources into a single file that can be moved around as easily as a PDF.

Proposed discussions for Beyond the PDF

I'm not only trying to push our particular software platform here, although we are seeking adopters to help make the toolkit more sustainable and I think it could make a good foundation for a repository-cum-journal application to support a praxis of scholarship 'Beyond The PDF'.

Our particular application aside, though, at the workshop I want to have a discussion about some of the underlying principles I've touched on in the above tour.

1. How important is interop?

    I think the answer is **very**[3]. This includes making sure that user-communities can choose their own tools, and within those communities, can work together. For the stuff that I'm interested in this means supporting at least Microsoft Word, OpenOffice.org and probably Google Docs for word processor-based authoring, and for dissemination, making sure we can deposit to Eprints, Dspace, Fedora et al, and interact with Drupal, SharePoint, WordPress and the gang.

2. What to package where? This is a complex trade off between interoperability, usability, future proofing, cost to develop, cost to deploy and risk.

    Options include:

    *   Word processor files as containers.

    *   Open web services.

    *   Zip packages (Epub, IMS packages, Plain old ZIP with manifests such as Bagit, OAI-ORE, METS)

---

3 Others take a different view and decide for various reasons to build tools that make interop at the user level hard, such as the Chem4Word project with Cambridge and Microsoft. Sure, there is a theoretical chance that someone could build interoperable tools for document authors to bounce a chem-enabled Word docx between Word, and say Apple Pages, but given the technology choices I would argue that it's almost certain none will eventuate – it's Word 2007 plus, or nothing – but it need not have been that way.

- "PDF plus" - jamming extra semantics into PDF without waiting for a new standard or trying our best with the current state of the art.

- Web-native conventions such as HTML 5 apps.

  In one sense this doesn't matter a lot, as there are transforms between them can be automated. But I would bet on one of the web-based approaches (HTML 5 and EPub, with IMS content packages or common cartridge or whatever they call it these days as an outside chance).

3. Can we come up with some best-practice guidelines for tool developers to decide where to build their tool? That is, does make sense as a web service or a word processor plugin, or an new PDF viewer or all of the above?

Dickinson, D. & Sefton, P., 2009. Creating an eResearch desktop for the Humanities. In eResearch Australasia 2009. Sydney. Available at: http://eprints.usq.edu.au/6090/ [Accessed December 9, 2009].

Hunter, J. et al., 2010. The Open Annotation Collaboration: A Data Model to Support Sharing and Interoperability of Scholarly Annotations. In *Digital Humanities 2010*. pp. 175-177. Available at: http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/pdf/book-final.pdf#page=201.

Mesirov, J.P., 2010. Accessible reproducible research. *Science*, 327(5964), p.415.

Sefton, P. et al., 2009. Embedding Metadata and Other Semantics in Word Processing Documents. *International Journal of Digital Curation*, 4(2). Available at: http://www.ijdc.net/index.php/ijdc/article/view/121 [Accessed October 22, 2009].

Sefton, P. & Downing, J., 2010. ICE-Theorem - End to end semantically aware eResearch infrastructure for theses. *Journal of Digital Information*, 11(1). Available at: http://journals.tdl.org/jodi/article/viewArticle/754 [Accessed March 24, 2010].