Beyond the PDF Trip Report

A couple of weeks ago I was in San Diego, USA for the Beyond the PDF meeting. Other parts of the country were a bit snowed under, but in San Diego it was like a Brisbane winter, with students at UCSD getting around in shorts and ugg boots. The trip was paid for by my employer, USQ, although I was handed a reimbursement form for some of the expenses so we may be able to claim some of it back.

In this post I'll look at the goals of the meeting what I contributed, what I thought of it and what might happen next.

The Goal

From the website:

> The goal of the workshop was not to produce a white paper! Rather it was to identify a set of requirements, and a group of willing participants to develop a mandate, open source code and a set of deliverables to be used by scholars to accelerate data and knowledge sharing and discovery . Our starting point, and the only prerequisite to participating, was the belief that we need to move Beyond the PDF (meant to capture a common philosophy, not necessarily to be taken literally).
>
> In a heady moment we might also describe our efforts as the desire to contribute to the development of a free and open digital printing press for the 21st century. A platform, when utilized, moves us beyond a static and disparate data and knowledge representation to a rich integrated content which grows and changes the more we learn. A system (content plus platform) from which a scholar can interact and once evaluated shows improved understanding and interest.
>
> https://sites.google.com/site/beyondthepdf

I said on the list before the conference what my aims were:

> 1. What at the end of three days do you want to have achieved by attending the workshop?
>
> I'd like to see (a) some progress in defining at least one re-engineered scholarly process for data-supported research publication, which will meet the current needs of researchers for recognition via traditional metrics and have a clear business model (in the broad sense) as well as show the way to the future of data-integrated, reproducible research, and (b) properly resourced projects to begin implementing at least one such model.
>
> 2. How do you see the workshop contributing to your goals in contributing to scholarly communication in the next 1-3 years?
>
> I hope the workshop can assist us with prioritizing and planning the various strands of our open source work on authoring tools, annotation and multi-format repositories. I note that one of the most useful signals for such prioritization would be funding that meets the goals in 1.

I think we made some progress on (1), and there is some hope for (2) – we have a demonstration system that my team is putting together now, and we can see if that can be used as a platform for funding applications, and as a way to generate feature requests from the Beyond the PDF community.

I put in an long abstract for a presentation which covered how I think the suite of systems built by my team can contribute to open, data-supported research and publication process. The presentation slots were only 15 minutes, though, with just ten for the presentation part so I tried to look more at some of the issues and try to start some discussion, and assume that people would read the presentation and get a sense of the scope of the

tools we have built. Not sure that approach really worked. I think I might have been better off doing a demo – as I got a much better reaction from the demo I did to the writing tools group late in the conference than I did to my talk. You can watch on YouTube if you really want to, but I recommend reading over listening to me rush through the stuff I wanted to say, going 'ah, um' a lot in workshop thinking on feet mode.

I have attached my presentation to this post using our the ReDBox research data management system that we'll be using as a demonstrator – the PowerPoint has been ironed out into a series of pictures, imagine the same for spreadsheets and other data in the research artefact of the future.

I tried to hit important themes rather than specifics. One point I wanted to make was that we can't assume that if you build it, they will come. That is, we can make what we think are nice commodious systems for researchers but *nice* is not enough. I used the analogy of a chook house (that's chicken coop to you non-antipodeans) – I showed a picture of one my family built for our chooks to live in, so they don't get eaten by someone's dog and so the eggs get laid in the right place. But this summer we have had a trio of young pullets who went a bit feral. We didn't have time to teach them to live in the house with the old girls before we went on holidays, and they ended up sleeping in trees, and laying eggs under the house, and lately the trampoline, and maybe other places. This is soluble, we (think) we know how to manage the flock, and teach them to live happily in their chook house, but the point is that the chooks didn't use the purpose built chook facility just because we built it. The repository crowd will know what I'm talking about.

In a similar vein Phil Bourne, the workshop organiser kept reminding us to consider reward mechanisms, it's no use building 'better' systems that don't suit the researchers, we need to give them reason to use them. In the case of scholarly communications, though, this is BIG change and there are very entrenched players and processes so change is slow and hard. I know from reading and hanging around with Peter Murray-Rust just how hard this change is proving to be in chemistry, but there are success stories like the Public Library of Science where Open Access publishing has started to catch on. There was some talk of build Beyond the PDF stuff on the PLOS platform.

The workshop was a lot to take in, like most conferences, but this one was very focussed. Still, trying to follow both the in-room discussion and try to assimilate all the new and challenging stuff and the twitter-fall was very challenging. I can't hope to capture all of it.

Outcomes

The site says:

> Short talks (see program, webcasts, twitterfeed and twitter archive) were followed by discussion and the major issues identified leading to a La Jolla Manifesto (under review) and a set of deliverables to move us towards common goals. One major goal is to use our tools to have some positive impact on the understanding and treatment of spinal muscular atrophy (SMA).

The workshop has spawned a series of working parties looking at different issues.

The two I have paid the most attention to, and the two which have content on the web site at the breakout session page are:

- Research objects

  > The central focal point of this discussion was focussed on **nanopublications** as a computable representation of scientific assertions with added provenance information. We consider '**Research objects**' (attempting to follow on from the definition provided in the MyExperiment link above) to act as holders of several data artifacts (that may or may not have semantic structure; including PDF files, excel spreadsheets, text files, *etc.*) *as well as any nanopublications that we might add to the data as annotations.* Note that this action of annotating non-semantic elements of a research object with nanopublications provides us with an incremental way of adding our additional content to existing data.

There were some people at the workshop who are very interested in being able to represent scientific claims as formal RDF statements (nanopublications), the idea being that we will have machine readable, and perhaps machine understandable science (it was all science – hardly a mention of the humanities that I heard). My opinion is that formal semantics are almost certainly useful at the level of statements like "Peter Sefton wrote this document" and may prove to be useful in some of the hard sciences. I keep arguing that one of the places we should start to build semantic web or nanopublication experience in metadata where we have known, identifiable (but as yet not consistently identified) parties (researchers, publishers etc) and identifiable versionable products (papers, data sets, pictures etc) that can be related to each other with formal statements. These are universal to all research and scholarship. If we can start to reason just at the level of who wrote what, and what data sets belong where then that's a good start. But that's a big if – metadata is hard enough without adding RDF to the mix.

I want to start with metadata, then maybe we can move on to dealing with how to get machines to reason about statements like "Mosquitoes transmit maleria"; where the arguments get pretty complicated pretty quickly.

- Breakout Group: Writing [I was in this one]

  The writing group decided that we should aim for developing a common tool (or set of tools), rather than to work on several independent projects, or just a set of standards, protocols, etc. The requirements for the tool we want to develop are:

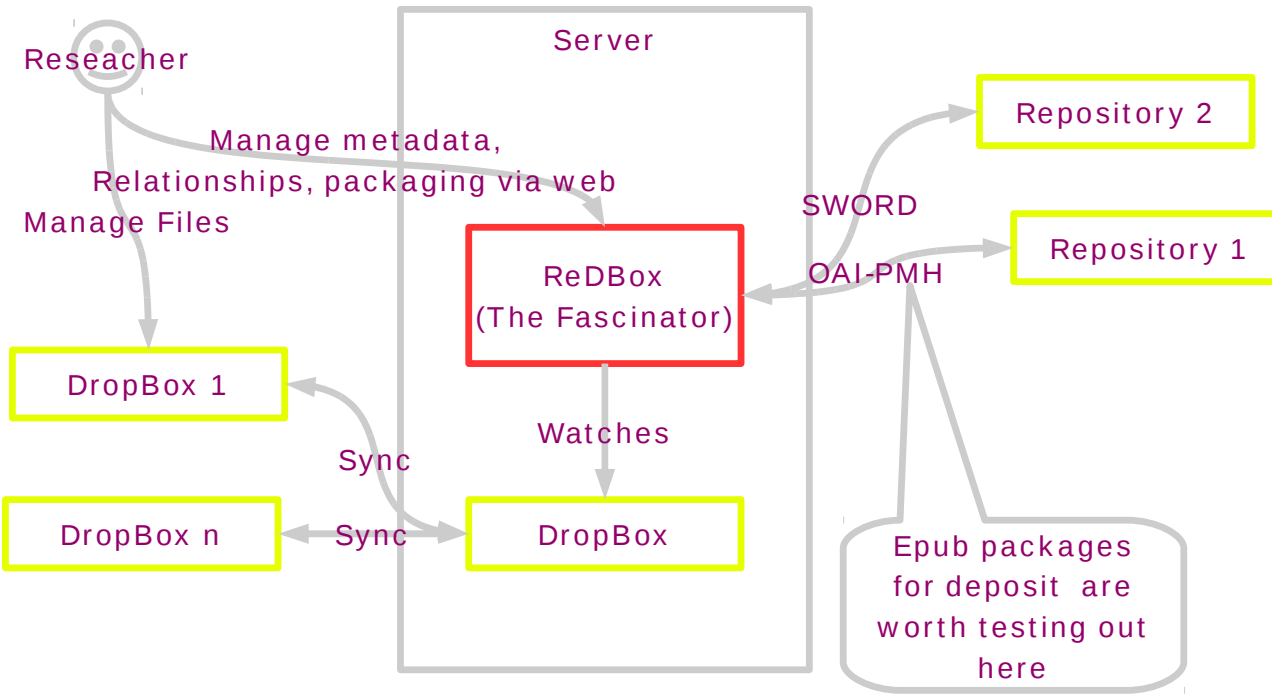
  - help with a major bottleneck in creating scholarly content
  - be as simple as possible (low hanging fruit, try to solve 80% of the use cases, go for the long tail in small labs, etc.)
  - have a working prototype no later than in 6 months
  - be useful for all scholarly disciplines

Back at ADFI

These are the things I will be working on with my team:

- Setting up a demonstrator of ReDBox (The Fascinator application for describing research data) watching a shared DropBox folder as part of the demo platform for the writing workshop. This will show how researchers can make their stuff part of the web (but private), sort it out, package it, and provide a platform where we can start talking about what to do next:
  - What are the parts of a research object?
  - How do we capture semantics in our word processor?
  - How do we provide services for automated markup and annotation.
  - Where do you send a research object to start the publication process?
  - What is the 'publication process'.
- Ditto with the Admiral project for Cambridge which provides managed research data storage packaged as a virtual machine. The Admiral work is potentially a useful component for ReDBox installation such as that at Newcastle, so we should explore it.

Demonstrator architecture



Benefits for USQ

The work we're doing on on the using ReDBox to manage files on a shared drive will:

1. Directly benefit USQ's researchers as we undertake our ANDS Seeding the Commons project, which will demonstrate data management best practice via a pilot, and generate new USQ policy.
2. Potentially benefit our online teachers and students by giving them very low-friction ways to create web resources and open educational resources including packages of multiple parts.
3. Continue to build the ADFI Software R&D team's reputation and standing as a leader in scholarly technology work.

Some issues I want to follow up

- Someone from one of the big publishers said to me, talking about business models “basically we curate and filter stuff”. I thought that sounds like what libraries do (in addition to functioning as coffee lounges – sorry, I mean *learning commons* – with bean bags and internets – seriously though, I said this before on the CAIRSS blog, librarians should be getting involved).
- There's clearly a debate to have on the benefits of HTML vs XML vs PDF as the format of record, for different purposes and at different stages in the research lifecycle, and some design work on new formats and interop options.
- Mid last year I proposed that we look at Epub as a packaging format for scholarly communications but

didn't really do anything about it. I'm pleased to say that momentum is picking up on this. After Beyond the PDF Martin Fenner talked about ePub, then implemented a save-as ePub for WordPress. I really want to look at interop between the work he's doing and our demonstrator – resources permitting. Reminder: there is a screenshot of some of the stuff from the workshop demo files, packaged as ePub, in my talk.

And finally, in breaking news – it looks like I might be in the UK for a hackfest around the weekend of the 11th of march, with a theme along the lines of Hacking / exploring Scholarly HTML, sponsored by Peter Murray-Rust's group at Cambridge. More on that soon.

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.