eResearch at ADFI

At the Australian Digital Futures Institute we work on eLearning and eResearch. In this post I will summarize where we are at with the latter, how I suspect it fits with the work that's going on at the Australian National Data service (ANDS) and where I think we could do more, in the form of a few project suggestions that might spark some debate if not some funding.

There are two software applications that form the backbone of our eResearch work. ICE is an established content management system which we are taking into the research realm and The Fascinator is a new bit of infrastructure designed to bridge the gap between the desktop (or the laptop, or the small lab) and the data commons – it is designed to bring repository services to the desktop, but it can also work at the server level. Over the next year I would expect these two systems to merge somewhat so that ICE content services are available as part of The Fascinator.

eResearch research themes

There are a two broad themes to our research:

1. Scholarly HTML, endorsed by Jon Wilbanks as a good tag for what we need to do to nudge the web towards semantics in his keynote at Open Repositories 2009: getting academic writing onto the web with as much inbuilt machine-readable meaning as possible. Under this heading we have:
    i. ICE is all about being able to make web pages from academic documents. That might sound like a problem that was solved a long time ago, but no, most research is still published in metaphorical paper format (that's PDF). As I argue in my forthcoming paper for Serials Review (Peter Sefton, Towards Scholarly HTML, Serials Review (2009), doi:10.1016/j.serrev.2009.05.001 (doesn't resolve to anything yet)) , only the big publishers have the tools to make HTML from research articles as a matter of course.

        It is important to get HTML into our scholarly communications platforms because HTML is what the web is made of. If we want to have rich interactive documents with embedded semantics and linked data then that's going to happen with HTML, not PDF, or Flash.

        With ICE we now have the basics sorted out what's needed is some work on ways to adapt journal and thesis templates, and support for LaTeX.

        Maybe we could do something with Microsoft Research along the lines of their article authoring add-in (more on my visit to Redmond soon).

        I know that this word processing / authoring stuff doesn't always seem important to eResearch but I think it's key, as articles and these are the jumping off points for people to discover a lot of data. We need to work on researcher practice, probably starting with the newest ones.

    ii. Related to ICE we have the work we have been slowly chipping away at getting theses onto the web, not just as pretend paper (PDF) in a repository, but as real scholarly HTML that can play nicely with the semantic web. The ICE-TheOREM collaboration with Peter-Murray Rust's group at Cambridge has wrapped up, but produced a really useful proof of concept for Scholarly HTML thesis publishing, presented by me and Jim Downing at OR09 a couple of weeks ago.

    iii. Recent work on ways to encode document semantics in an interoperable way. I have made some progress on jamming semantic relationships into URLs so they can be used in all manner of authoring tools, and Duncan Dickinson had a go too. For example, Linda Octalina wrote some code that can take semantic stuff embedded in a .docx file using the MS Word Add-in

and turn it into a plain-old link. The idea is not to litter the web with ugly links but to allow tools like Word or a repository to recognize when a link is encoding some semantics and let you do useful things with it while still allowing interop with tools which do not understand the magic links.

I have now convinced myself that we could really give the semantic web a good kick-along if there was a trusted service which could construct links that encode metadata, citations, references to concepts, and data embedding. More below on a potential project.

1. **Closing the gap** between the desktop and the data grid with our work on The Fascinator Desktop. This new, in progress application indexes local files, and watches the file system for changes – then exposes all your stuff you via a web interface. The idea is to allow researchers to tag and classify their research materials then create virtual collections of material which will be routed to the right downstream repositories automatically. This ties in with seeding the Commons program:

   Program Aims

   - To improve the state of data capture and management across the research sector
   - To improve the fabric for data management and the amount of content in the data commons

   http://www.ands.org.au/repositories.html

   We've had feedback from all over the place, including inside of ANDS that the Fascinator Desktop's 'sucker upper' functionality is a missing link in the eResearch software stack. There are obvious links to the ANDS work:

   i. The Register My Data service. Using the RIF-CS collection description language can we get authors to label their collections and then see their data routed from the desktop to the data commons and become discoverable?

   ii. The Identify My Data service. Can we get some real life researchers (including ourselves) assigning IDs to bits of data when we write about them? I have wanted to be able to do this with stuff like sample word processing files for ages; these are the data we work with a lot. Can we associate a paper with a bunch of sample data, identify it all and then make sure it is all deposited in a repository at the right time? Not to mention stuff like the 7GB virtual computer we created for ICE-TheOREM. Where can I put that?

Some potential projects

Building on (1) our ongoing work on Scholarly HTML and associated services, and (2) closing the gap between the desktop and the data Commons, there are a few projects which I think might help the ANDS agenda.

Project: A *real* version of ontologize.me

The point of my toy site ontologize.me is to show how a service might be created where people can do semantic web stuff. I think this would solve several real problems in a simple way:

- Let people embed metadata in a document that is produced using pretty much any authoring system. With the People Australia IDs now coming onstream this is becoming a practical reality.

  Here you go: I assert that I, Dr Petey, am the author of this here blog post.

  If I link to another Sefton, my sister Catherine then I'm not asserting anything I'm just linking.

Click and you'll see. And note how we have all these different forms of our names – she publishes as Cath[1], even though she's really a Catherine, I publish as Peter, but use Petie a lot and the NLA has me as P. M. Sefton. The URI-as-identifier makes this no longer a problem.

(Every time I mention this idea Bruce D'Arcus shows up on my blog and points me to RDFa – yes RDFa is a way to encode this stuff in a web page and yes we will do that so that it can be indexed, but it's too hard and too fragile for authors and flat out impossible in word processors to use in authoring workflows.).

- Provide standard ways for people to link data into a publication (email, thesis, paper, blog post, Google Wave :-) in such a way that downstream sites can choose to do something useful with the link. Like, for example if you link to a sound file, do so in a way that makes it clear to downstream applications that that's what you have done, so they can embed a player, or if you link to chemistry then they could embed a chemical viewer (that rotating 3d molecule trick is still very popular around here).
- Allow you to mark up the content of a document to show what it's about, like the Microsoft Word Ontology Add-in, but in a form that will work for everyone.
- Provide some standard machine-readable ways to do citations using links.

I'm imagining a kind site with a nice web wizard that lets you construct these links so you can drop them into a blog post or a paper or a wiki. The ANDS Identify my Data service for example could give you a wizard to not only identify the data but what you mean by it. And ANDS and/or People Australia could provide URIs not only for a person, but for that person in different roles; author, editor, subject etc. Duncan Dickinson points out that as usable interface between authors and ontologies is going to be really important. Most of them wouldn't know an ontology if it crawled into their lap.

I think that the ANDS crew might be working on some kind of online storehouse for ontologies and taxonomies and I reckon my extra wizard service would be a really simple but useful addon. We would love to be able to build support for this into our word processing toolbars, our conversion services and into institutional repositories, not to mention The Fascinator, which could expose these semantically rich URIs for every bit of data on researcher's local disks.

Project: Trying the ANDS RIF-CS collection metadata schema with real desktops

We have been working on The Fascinator desktop, with a broad range of users in mind, but we have one key user (not that she has the software yet but we're getting there). That's Leonie Jones – who has loads of video, transcripts, geo data around the battle of Fire Support Base Coral. Leonie has made a film with this material but there is lots more that goes with her PhD thesis. I think we should look at how Leonie and colleagues in Chris Lee's Public Memory Research Centre might be able to describe their collections in such a way that we can advertise them to the data Commons using the *Registry Interchange Format - Collections and Services* (RIF-CS). To pull this off we need ways for users to label data and state which bits can be released publicly and we will probably need a research-centre-level repository server which can do the talking to the ANDS systems.

We're hoping to build software that can be customized to work with any data set but which will generalize across disciplines.

Project: ASHT - “Australasian Scholarly HTML Theses”

Finally, something I mentioned in my post on what USQ could do to assert itself as an open access

1 I dare you to call her Cathy. Go on.

institution; theses. The Australasian Digital Thesis program (ADT) is now mature, and theses from across Australia routinely go onto the web, if only in PDF format. One of the best ways we could seed a data commons, and influence researcher practice would be to start with the earliest of early career researchers. I'd love to see a project which took a few small cohorts of PhD, Masters and Honours students and gave them the kinds of tools we are developing to manage their stuff and to write about it, in a way that will bring on the semantic web. This is definitely under consideration at USQ, but it may be of interest to the broader community here in Australia. (USQ has some responsibilities to ADT via our hosting of CAIRSS but we are still working out what our role will be).