Open Repositories 2009 trip report

Here's my summary of my experience of OR09 it Atlanta Georgia USA. Tim McCallum and I came over from USQ and arrived on Friday night after 30 plus hours of traveling, for a Monday start. Tim discovered that if you lose the posters due to severe fatigue then getting them printed out on the Georgia Tech library's plotter is easy and cheap.

There were only a few Australians this time. I was surprised that there was nobody to represent/promote Fez or Mudadora, two antipodean repository solutions based on Fedora Commons but I did meet a Muradora user, Juan Rodriguez from the Memorial Sloan Kettering Cancer Center who knows the Muradora team – sounds like it's alive and well.

I'll go through some general impressions of the conference then summarize my direct contribution; moderating a session, giving a couple of papers, presenting a poster, serving as a judge on the Developer Challenge and attending workshops and meetings with Microsoft Research. This is, of course, a personal view. As with any conference I missed stuff while I was working on presentations, plotting, having a jetlag-management nap, looking out the window, or judging the competition etc.

Overall impressions

Lots of people I have talked to have remarked on the movement towards modularity, where repositories are not monolithic systems but sets of services. I can't remember who it was who reminded me of Clifford Lynch's 2003 definition of a repository as a 'set of services':

> In my view, a university-based institutional repository is a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members. It is most essentially an organizational commitment to the stewardship of these digital materials, including long-term preservation where appropriate, as well as organization and access or distribution. While operational responsibility for these services may reasonably be situated in different organizational units at different universities, an effective institutional repository of necessity represents a collaboration among librarians, information technologists, archives and records mangers, faculty, and university administrators and policymakers. At any given point in time, an institutional repository will be supported by a set of information technologies, but a key part of the services that comprise an institutional repository is the management of technological changes, and the migration of digital content from one set of technologies to the next as part of the organizational commitment to providing repository services. An institutional repository is not simply a fixed set of software and hardware.
>
> http://www.arl.org/resources/pubs/br/br226/br226ir.shtml

I agree. It's not a computer program, it's a lifestyle; what Lynch is calling *organizational commitment*.

I think the title of a presentation from John Kunze, Stephen Abrams and Patricia Cruse of the California Digital Library (CDL), Permanent Objects, Disposable Systems summed this up really nicely. I liked the stuff from the CDL and Library of Congress looking at simple ways to describe and move data; the 'non-repository' movement. We'll be looking into BagIt, Pairtree, Dflat, etc, particularly for our work on The Fascinator Desktop where we need tested, safe, documented ways to organize data in way that is as lightweight as possible.

I had a little moment in the spotlight when keynote speaker John Willbanks referenced my 'Scholarly HTML' idea. This was reported in Twitter thus:

> akosavic **Wilbanks** at **#or09**: rename "semantic web" as "scholarly HTML"

So there you have it, meet the saviour of the semantic web. Move over Sir Tim.

Actually I wouldn't go that far – what I am trying to get at with this Scholarly HTML is that the research article – our unit of academic currency should be a web page, not a bit of pretend paper, a PDF. Journals need to be reinvented. Articles should be web pages (yes we need ways to time-stamp and version them). Peer review and editing are both important, but I can think of better ways to get those done than we typically use now. Then there's the idea of embedding machine readable semantics in the form of statements of fact, links to data etc, not to mention machine readable metadata. More on this soon here on the blog – I think I'll write a series of papers on this, with appropriate collaborators, in the open then we'll see if we can get them to count as scholarly literature via peer review. A couple of people told me they're watching the Scholarly HTML posts so I think I'm onto something with this one.

Repository sustainability

I was asked by email before the conference to moderate a session, Strategies for Innovation and Sustainability: Insights from Leaders of Open Source Repository Organizations. Chairing sessions is my least favourite part of conferences, but I said yes. What they didn't explain was that this was not just paper session, it was a panel session, where the moderator had to do more than just introduce the speakers. On stage with me were the leaders of the three major repository platforms. Michele Kimpton of the DSpace Foundation, Sandy Payette from Fedora Commons, Les Carr of ePrints fame from the University of Southampton and Lee Dirks of Microsoft Research.

The big news this conference was the recent merger between the DSpace and Fedora Commons organizations to form DuraSpace. That meant that I got to introduce Lee from Microsoft as the new player in the open source world battling a creeping DuraSpace monopoly (Microsoft's new repository Zentity is likely to be released as OSS). Before I left for the US my partner advised me not to call Microsoft the 'underdog' – so I didn't.

The three organizations on the panel each had ten minutes or so to talk about how they are set up to sustain their open source software. I don't think there was much definite there. The DuraSpace crew are still working out a sustainability model, while ePrints remains driven very much by Southampton, but with some cash coming in from selling services. The Zentity repository is too young too need a sustainability model – it needs an adoptability model.

My question to the panel was basically to ask them to play devil's advocate and ask them 'what's the worst thing that could happen'.

In the case of the new MS Repository, Zentity Lee was upfront – if there's no uptake then the product will not be supported. That's basically the same as with any repository, but this one will be a bit different if it takes off, as it's what Sandy called 'open at the edges' in that it runs only on the Microsoft software stack – what this might mean long term I don't know, but if your organization decides to move platforms then the repository won't be going with you.

The answer from the other two organizations as to what could go horribly wrong, was 'not too much' – at least that was my reading of the answer. My summary is that with ePrints, Dspace and Fedora commons there are enough users that if the central organizations crumbled or gave up then someone would invent a new one.

But I reckon the best thing that repository managers can do is get familiar with the ways you can import and export data while intoning to themselves Permanent Objects, Disposable Systems.

Papers and posters

At OR07 the reviewers didn't think that my proposed presentation about dragging repositories onto Web 2.0 was worth scheduling but we kept working on dragging repositories from Web 0.5 collections of PDF into the twentieth century (stay tuned for the twenty-first). This time, I was able to put all the stuff I did for the conference straight into ePrints. Each item was authored in ICE, in OpenOffice.org, (although I could have used Microsoft Word) with an embedded slide-show. Even my poster had an embedded sideshow, a straight HTML view and PDF.

First up Jim Downing from Cambridge and I showed off the work we did with our teams on the ICE-TheOREM project. Not only were we able to show a thesis going onto the web in HTML as well as the dreaded PDF, it had granular chapter-level embargo, and we were fully buzzword compliant, with ORE and SWORD built in. And for the first time we made our work available as a ready-to run virtual machine, a few copies of which I handed out. We'll definitely do more of that, and keep updating our machine with all the software we work with – at the moment it runs ICE, ePrints and The Fascinator, but I'd love to see DSpace and OJS and Moodle on there as well all integrated.

I gave a presentation on The Fascinator in the Fedora user group stream (great rooms with power at every seat) which was gratifyingly well attended.

And there was the poster, which I supplemented with a metaphor – a collection of 40mm & 50mm PVC waste pipe and various connectors. David Flanders used it to build a data grid which included a pipe going straight to repository hell a place he has apparently spent a fair bit of time drinking microbrew with too much malt. The idea was to drive the point that we want to make research data plumbing as easy as PCV pipe network engineering. Here I am spruiking the poster with a fistful of PVC.

At time of writing there are a couple of minor usability issues with the HTML-in-ePrints approach which I'm sure will be fixed soon.

Microsoft workshop

I attended Microsoft's workshop on their growing set of academic tools. I'll reserve judgment on the new Zentity repostiory, but I am very, very pleased to see Microsoft Research working on academic workflows. As part of that, the idea of building HTML conversion into repository deposit tools is getting a serious airing in the repository community and I'm very pleased about that. More after my visit to MS Research tomorrow. (I have already spent a little time in Seattle with Pablo Fernicola, wandering in the sculpture garden and talking in general about academic computing).

Judging the competition

The final thing to mention is the Developer Challenge, run by the unstoppable David Flanders (I was going to say indefatigable but I can't spell that) with Rachael Rodenmayer assisting. This was judged by a team of five based on five minute screencasts. There were some good ideas in there, in a field that spanned entire repository developments that were already done to small prototypes. Read about the winners at the JISC site.

First place went to Tim Donohue's Mention It, which is a kind of trackback mechanism for repositories, the kind of thing that should drive repository deposits by being useful, as opposed to the kind of thing that tries to load up repositories by forging bigger shovels to, you know, shovel more stuff in.

Second place went to an astoundingly useful, clever idea from Rebecca Koesar, which was to turn the Fedora repository into a filesystem with her FedoraFS. The demo will appeal to geeks, shows how you can use ordinary system calls to do stuff to the repository. Lots of potential here and I hope she is able to set it free for download. The ePrints team could use it to stick Fedora under ePrints, even (although with their new modular storage layer it's going to be easy to do that anyway – I bet there's a a Fedora layer before OR10).

One of my favourites was the ePrints app store, having just waited several months to get new code into our ePrints to support ICE integration I welcome the idea of a central store that makes it easy for repository managers to drop in new modules but other wiser judges pointed out some of the risks. I think calling it an app store is a bit of stretch, but something like the WordPress plugin mechanism would be very welcome in the repository world, with the proviso that managers should be encouraged to test everything in a sandbox, then a formal test environment, then think very hard before turning on any new plugin.

While we saw some interesting stuff, I think that we could probably break this challenge into two bits – one prize for development pre-conference that could be tied-to and judged alongside poster presentations and another for in-conference collaboration, with development focused on a set of challenges decided before hand. There should be two tiers, one small set of prizes for well-specified small challenges to address and bigger prizes for rising the challenge. That's what I think anyway.

Conclusion

Good conference. Good venue apart from a lack of power points (no lack of PowerPoints unfortunately) in the main venue. I was pleased to be showing off HTML in ePrints at last, and have ICE in production as an eResearch workflow tool. I think the repository world is making a welcome move to (re)embracing the idea of small pieces loosely joined.