Trip report: visit to Microsoft

I have just returned from the USA, where I attended Open Repositories 2009 in Atrlanta. The second part of my trip was a visit to Microsoft Research in Redmond, a suburb of Seattle. It was my idea to visit Redmond, and Microsoft Research kindly set up a day's visit for me. USQ funded my trip and accommodation. Pablo Fernicola took me sightseeing and to lunch on the weekend[*] – thanks Pablo – and Microsoft fed me lunch and put me in a town car to the airport.

Discussions

I spent Tuesday 28th May with Lee Dirks , Pablo and Alex Wade. I wanted to talk them mainly because I would like to see Microsoft Word able to play better with the web – more on that below. I got to meet a few other people too:

1. Sumit Chawla talked about the Microsoft interoperability group and a number of their projects. I am particularly interested in the attempt to build an OOXML (.docx) to HTML converter which can work without Word. One of the bits of homework I set myself was to look into how well this actually works. If it can handle embedded graphics and so on then we could look at using it with ICE, or with a Word Add-in for scholars (I have to say I'm pessimistic but I have not got it working yet). I think the converter is a product of The Planets project.
2. Chris Wilson who spent a long time on the Internet Explorer team and chairs the W3C HTML working group made himself available to see if I had an particular issues with IE – we don't really apart from the the well known ones that are being attended to. Talking to Chris put our little scholarly corner of the web in perspective, for example ORE is not on his radar at all. I think ORE is useful and interesting but it is very very far from being mainstream – something we need to remind ourselves of lest we become over excited and over estimate our own importance. Chris did encourage me to bring our requirements and concerns to the W3C though, for consideration in HTML 5 and beyond.
3. Brian Jones of OOXML fame joined us lunch in an pretty good Indian place on the Microsoft

[*] I was unlucky enough to be stuck in Seattle for the Memorial Day long weekend, and suffered unrelenting sunshine and blue skies, instead of the hoped-for drizzle. This was exactly as Tom Robbins put it in Jitterbug Perfume:

> "*With the absence of the cloud cover that normally caused the sky over Seattle to resemble cottage cheese that had been dragged nine miles behind a cement truck, the city, for the first time in memory, would have an unobstructed view of one of nature's most mystical spectacles.*" .

My hotel was right near the Seattle Centre where the NW Folklife festival was going on. Sort of like taking the Woodford folk festival and dropping it in Brisbane's Southbank, I'm not sure if it was one of the world's most mystical spectacles but it was very inconvenient if you're trying to walk through the park. Buskers kept blocking my path, such as the Black Death Allstars with their infectious but downright unpatriotic "This van is your van" (I didn't tell the travel office I had been exposed to the black death, but they didn't want me to come back to work anyway). As you can understand movement was so difficult that I got stuck there at the festival at times. I met some musicians, for example guitar god Yusuf Kilgore who told me about a Bob Dylan tribute night at the Conor Byrne Pub where people kept talking to me instead of politely ignoring me like they do in Australia. One lady even misquoted the above line about clouds, apparently in defence of Seattle's weather. To be fair there was a duo there that did a great stripped back ukulele-extreme-Bossa-Nova version of Tambourine Man, and Yusuf's playing was pretty good when he turned up, about 5 hours after he told me he was going to be on. I didn't run into Tom Robbins as I was hoping but Yusuf played at his birthday party and says he's really old so that's something.

To keep away from the swine flu virus and the Black Death Allstars and maintain the peak fitness required for my job at USQ I rented a bike from Classic cycles and took myself on a tour of Bainbridge Island past all the rich people's summer houses, at least 20 miles worth I reckon.

campus. I have exchanged a few blog comments and emails with Brian over the years; it was great to meet him in person. Talking to Brian about the way custom XML is embedded in Word documents and the interface design challenges that introduces reinforced some of my opinions of how the various Add-ins MS research are producing are likely to fare, more on that below. I also tried to fill in a bit of my knowledge about the various XML and nearly-XML formats that Word has supported over the years[**].

Thanks all. It was a really interesting and useful visit for me.

In this post I will report and reflect on discussions with and about Microsoft Research's work on eScholarship. There are a few things of interest:

1. The Zentity repository, which is exciting in that is built on top of an RDF-style data model that lets you express pretty much anything, and potentially a huge maintenance problem because it is built on top of an RDF-style data model that lets you express pretty much anything.
2. The MS Word Ontology Add-in, which I think is an fine idea, about which I have commented here before. After my visit I am still of the opinion that it is going to be hard to make it usable, that its lack of interoperability even with older versions of Word is a major problem – nothing will make people drop a tool like this quicker than a few disasters sharing with colleagues or taking a document home and finding that an older version of Word mangles it. I am also concerned that the fragility that results from imposing a strict hierarchy model on top of Word's native implied hierarchy will be an ongoing headache for developers and it will be hard to provide a bomb-proof implementation. I hope the MS Research people will take a look at the work we did to transform the markup the Add-in uses to plain-old links but if they don't like that it's open source so we can go ahead and do it anyway.
3. Chem4Word, being done by Microsoft with our associates from Cambridge. I think its heart is in the right place, but I am concerned about the same things as I am with the Ontology Add-in; interoperability and fragility of embedded custom XML being the two big ones.
4. The SWORD Add-in which lets you post documents straight from Microsoft Word to a repository. I'm very interested in this one. I'd love to see it integrated with some kind of HTML conversion so that people can put web pages into web based repositories instead of filling them up with virtual paper. There was a meeting at OR09 to look at how Word might work with repositories, with an incredibly strong response from the ePrints team, who are very enthusiastic about supporting this kind of ingest. Me I think it's a reasonable thing to work on, but it is only one workflow, and I think that we are probably going to see a lot of action in intermediate content management systems that manage authoring and data rather than the typical repository focus on dissemination and preservation – most of the interaction from Word is likely to be with those kinds of content systems, in my opinion.
5. The Microsoft Word Article Authoring Add-in. I'd have to say that nothing that I saw in Atlanta or Seattle has made me change my mind, I still think that this approach of trying to edit documents conforming to a large complex XML schema inside Word is going to end up as an unhappy compromise – if it takes off at all it will be at the publishers who use XML, rather than with ordinary authors. Pablo Fernicola thinks otherwise, obviously. Time will tell.

Scholarly HTML

Against this background I will confine myself to the dimensions I really care about, which is how to make word processors produce good quality HMTL, and document interoperability. I've been over and over why this is important here, but here's a summary.

1. **On the authoring side, offline word processors like Microsoft Word and OpenOffice.org Writer**

[**] I still think the Save As HTML format in Word 2000 could have been a round trippable XML format which would have been much more approachable for developers as it was HTML based. I wrote an article for XML.com back in 2004 about how to transform this format in and out of XML – and I think that might still be a useful technique.

are probably still the best all round compromised for academic authoring in those disciplines which don't use some other format like LaTeX. For now. I expect this to change soon, we are starting to see document drafting in Google Docs (which lacks citation services and styles and easy embedding of diagrams so far) , and if Google Wave realises its promise then I think it could be an end-to-end scholarly communications platform.

2. **On the delivery side, academia is one of the few places where PDF is considered acceptable** as a means of communication whereas on a normal website it is regarded as an impediment to usability. We need to be getting scholarly works into HTML so we can do more with them; meshing them with data and visualisations and delivering them to mobile devises.

While we wait for Google Wave to take over the world, what I'd like to see is a Word toolbar much like the ICE toolbar to support scholarly authoring but with better integration into Word than we have had the resources to make so far here in Toowoomba. It should let people create well structured documents which can be pushed to academic systems; journals, repositories and learning systems and not just in PDF, or Word format, in some kind of formally specified Scholarly HTML. I think that idea had some support at our meeting, but Lee Dirks in particular pointed out that it would need to be done with reference to a stakeholder group who can help define and own this Scholarly HTML thing. I'd be interested in ideas on who these stakeholders might be;

1. Publishers obviously, where MS Research have great contacts.
2. Repository owners particularly the discipline repositories like arXive and Pubmed Central.
3. The eResearch community; I hope that I can get the Australian National Data Service (ANDS) interested in this stuff.
4. The Electronic Thesis and Dissertation (ETD) movement. (My group is involved in this via our CAIRSS repository support service, the Australasian Digital Thesis program in Australia will come to CAIRSS at some point.)
5. The eLearning community, maybe.

But actually, where this matters most is on the long tail:

1. Thousands of small repositories and journals are stuck with paper-on-screen because that's all their tools support.
2. The small but growing group of users who want to do more with the versions of their documents they deposit in repositories.

I'd appreciate any thoughts about who might be interested in defining a scholarly profile of HTML – a few people told me they're following these posts so please speak up in the comments.

We are not working together

I'd like to make it clear that while we had a good talk there is no project immanent between MS Research and ADFI; I think the discussion was reasonably encouraging that there might possibly be some room for collaboration.

First some ground rules about the kind of collaboration I think we should entertain with any commerical entity.

I think it would be fair that this works only in the latest version of Word, **provided the documents it produced could be used in other editors**, such as OpenOffice.org Writer or earlier versions of Word. I will quote an earlier post here:

> In conclusion I offer this: I would consider getting our team working with Microsoft (actually I'm actively courting them as they are doing some good work in the eResearch space) but it would be on the

basis that:

- The product (eg a document) of the code must be interoperable with open software. In our case this means Word must produce stuff that can be used in and round tripped with OpenOffice.org and with earlier versions, and Mac versions of Microsoft's products. (This is not as simple as it could be when we have to deal with stuff like Sun refusing to implement import and preservation for data stored in Word fields as used by applications like EndNote.)

  The NLM add-in is an odd one here, as on one level it does qualify in that it spits out XML, but the intent is to create Word-only authoring so that rules it out – not that we have been asked to work on that project other than to comment, I am merely using it as an example.

- The code must be open source and as portable as possible. Of course if it is interface code it will only work with Microsoft's toll-access software but at least others can read the code and re-implement elsewhere. If it's not interface code then it must be written in a portable language and/or framework.

http://ptsefton.com/2009/03/16/opening-up-microsoft.htm

A potential Add-in

So in a world where we did have an idea of what Scholarly HTML looks like what would a Scholarly HTML Word Add-in do?

The basic requirement is that it would allow scholars to:

1. **Write papers, books and theses as well as more ephemeral or less formal documents** using a single interface, built on top of Word, working with its strengths rather than against its limitations.
   - It should be able to adapt itself to pre existing journal templates,
   - but also be able to help journals and institutions build useful, usable templates that will produce Scholarly HTML automatically and if necessary map to formats like NLM XML.
1. **Post that work to the web**, including to content management sites, blogs, repositories, journal submission sites – everywhere – from within Word where that makes sense (and it doesn't always).
2. **Make documents which are as data-integrated and machine readable as possible**, with the things I outlined for Scholarly HTML; citations, metadata, embedded semantics and linked visualizable data. The idea would be to encode all this stuff in a way that could be safely moved between systems and then to build web based plugins that sites could use to make the documents come alive.

Some of this stuff we have already looked at in ICE, obviously, but it would be good to look again at how we have done things – as Alex Wade said we should think about requirements and leave implementation aside. So, in a departure from form, I'll just leave the requirements above as-is and spare you my thoughts about implementation other than to comment that it would be interesting for MS Research to make their plugins work with the new Open Document Format support in Word – via simple interop strategies like encoding information in URIs or in styles.