More on Microsoft Word and non-interoperable standards compliance

Glyn Moody was pleased with my response to his rant on Microsoft Collaboration.

Others were less so. And Jim Downing would like me to expand on what I think might be good ways to do interoperable plugins. I'll start by dealing with the comments, which leads to looking at how plugins like Chem4Word might work. I don't have time right now to do a lot of background research for this post so I will pose a few questions and I'd appreciate more feedback.

First up, Ian Easson has some advice for me.

> The essence of your objection to custom XML embedded in OOXML files is this:
>
> "This custom XML is an insidious trick in my opinion as it makes documents non-interoperable. As soon as you use custom XML via Word 2007 you are guaranteeing that information will be lost when you share documents with OpenOffice.org users and potentially users of earlier versions of Word."
>
> You should learn more about something before you make a big fuss about it.

Thanks Ian, I'll try to fit that into my busy schedule.

He then says:

> First, it is entirely a question of the *consumer* of this information as to whether the custom XML gets lost. In the case of older versions of Word using Microsoft's free add-on to read and write OOXML, THE INFORMATION IS NOT LOST. The "write" part of the add-in preserves any such information.

Right. The information is written out into OOXML then when you open it in an earlier version of Word you can't see it. I tried this with the NLM plugin and as far as I recall there was no structure apparent if you open the file in an earlier version of Word. The result is that if you try to edit the document you break the structure at which point THE INFORMATION IS LOST.

And then there's OpenOffice.org:

> As for Open Office not preserving it (if it does so, I don't know), the fault is with OpenOfiice, not with Word or with the OOXML ability to embed custom XML. Talk to them about it.

I'm pretty sure that the official Sun OOo build can read OOXML but not write it (thanks Sun for your commitment to interoperability) and whether it discards the XML is irrelevant in that case as there is no equivalent mechanism in the Open Document Format. There is another version of OOo with some code contributed by (I think) Novell – which may preserve the XML but the situation is the same as with earlier versions of Word – the XML is unlikely to survive editing of the document.

> As for the general argument that the custom XML in IS29500 documents acts to reduce interoperability, you are simply wrong, for two reasons:
>
> 1) Applications that consume and generate IS29500 (OOXML) documents, if properly written, are supposed to ignore (not strip out) any such custom XML if they don't understand it. So, the custom XML is totally invisible to such applications.
> 2) For applications that are specially written to consume or generate a specific variety of custom XML, the ability to have such XML embedded in ordinary office documents acts greatly to IMPROVE their specific kind of interoperability.

> So, the end result is: improved interoperability, with no downside.

I disagree. This doesn't promote interoperability in practice, as I noted above, it promotes sales of Word 2007. I think this 'feature' is a trick.

What **does** promote interoperability is the approach that Microsoft itself used to use with OLE embedding. We saw this in action with Peter Murray-Rust last year. He brought a Word document with embedded ChemDraw chemistry pictures in it. Guess what? When you open them in OpenOffice.org Writer you can view the rendered image, even if you don't have ChemDraw. Even better, when you crack open the ODT file the original ChemDraw binary is there – we were able to feed that to some open source software that PMR and team were involved in writing and extract data. The result of using OLE was: workable interoperability. That's because OLE objects are embedded using standard interfaces not vague hand-wavy whatever-you-like Custom XML.

If he'd turned up with a document authored with the Microsoft Word ontology plugin we would have seen none of the embedded semantics in OOo Writer. So, the end result would have been: reduced interoperability and no upside.

> You should rethink your attitude towards this matter after you have checked out the facts first.

Thanks for visiting. Feel free to correct any factual mistakes.

Doug Mahugh of the Microsoft Office interoperability project also dropped by.

> I agree with your view of the value of interoperability with open software. Unfortunately, that isn't possible yet for this issue, because there is no approach to semantic tagging that has been implemented in open-source applications. When the RDF approach in ODF 1.2 becomes widely implemented (assuming it does, of course), there may be some interesting options there, but for now implementers need to either use a product like Word, or overload styles (or something similar) to get the job done.

Actually, Doug, my question was for the specific case being covered in the ontology Add-in. If I want to link a to bit of my text to a node in an ontology why could I not just use a link,as in Linked Data? Say I'm talking about my long suffering mongrel dog Spensa. I might want to link 'Spensa' using, say his OpenId and maybe the words 'mongrel dog' to some kind of taxonomy and/or ontology (I guess a taxonomy is an ontology or is that wrong?). This link could be styled to look unobtrusive in text, and could possibly be automatically footnoted for print viewing and linked in some cool way on the web. I had a go at this but I was unable to find a useful endpoint for my link for whatever species he is, canus-spensis I suppose.

Or the case I mentioned of wanting to assert that authorship. What if my local ePrints repostiory had a page for me-as-author. It might look like http://eprints.usw.edu.au/authors/PeterMacolmSefton and resolve to a page that describes me and the works that are attributed to me with a note there that says "if you want to indicate that this person is an author of a paper, link their name as it appears on the work to this page". That is way simpler to implement than the stuff we talked about in our paper on embedding semantics in word processing documents.

Now, this linking is something anyone could do with a little training but you could build a Word plugin to make it easier. I'd have no objection to that – particularly because it is possible to build code that can work in both Word (for Windows) and OOo as we have shown in our ICE project.

If I'm wrong about simple hyperlinks being a workable solution there are a couple of mechanisms in Word for embedding semantics. Styles are probably no use in the case, but fields could work. A custom field could contain the semantic info, it will get stored in the underlying XML just fine and it will work with any version of Word. Why didn't the ontology tool get built using fields to store the info? It wouldn't have anything to do with wanting to sell more copies of Word 2007 now would it? (I know, OOo dumps Word fields but it would still be an improvement)

Bookmarks are another (clumsy) way to do interop. I know this stuff because I did a lot of checking of facts when we worked to build the first interoperable citation toolbar for Zotero that would let you round trip documents from Word to Writer and back.

I am well aware that this is a jungle. Microsoft is a big cat whose DNA tells it to sell more copies of Word 2007. That's fine, none of my business. My business, amongst other things is to help the people at USQ make web pages and books from their writing. And they don't all have Word 2007 and they don't all want Word 2007. I suspect that with the ontology project there would be a more interoperable way to do the in-document tagging. Maybe someone from the project could share some of the background. Did you try to use fields? Did you consider interop with OpenOffice.org? Do you test tagged documents that have been passed to older versions of Word, edited, saved and reopened?

Now, onto the final bit where I think Doug is talking about the use case where you want to author a document in Word that is later going to be mapped onto a complex schema. After working in this field since 1995 I am pretty well convinced that attempting to write a fully featured editor layered on a word processor for the likes of the NLM DTD or DocBook is not going to end well. I have seen enough of these things marched out at SGML and XML conferences with great hopes and waited in vain for the follow up presentation next year. Anyone remember BladeRunner? Microsoft SGML Author?

> In most scenarios people have a specific schema in mind for tagging the content of the document, so with the styles approach you end up imitating XML structure through nested styles. That can get very complicated, and can cause interop issues for consumers that don’t handle nested styles well. I like the microformats/customXml approach because it’s simple for developers to work with, and consumers that don’t care about custom semantics can just ignore it without losing any formatting (as would happen if styles were ignored). It’s also easy to round-trip the customXml element, since it’s in the WordprocessingML namespace.

Easy to round trip in Word 2007, but what happens if someone without the plugin edits a document with the custom XML in it, particularly one like an NLM document where the stuff is all over the place? What if they reorder sections, or delete a section heading? Has this been tested or is the intention with the NLM plugin to recommend that people only use Word 2007?

Doug is right that overly complex style systems don't work for authors. Been there. Done that. That's why we have devised a simple, generic set of styles that users can use to create structured HTML, on the ICE project. Ian Barnes used the same styles to create DocBook, and I'm pretty sure we could do the same for NLM but all we would be doing is producing a small subset of those formats.

There are so many words flying around here it may not be clear what I'm complaining about. So I will try again. What I am saying is:

We in the Academic community have an interest in promoting interoperable authoring practices so our users across platforms can collaborate on documents together. Therefore if we get involved in working with proprietary software we should make a serious effort to make sure that what work *we* do will work as widely as possible. *Our* goal is not to promote Word 2007. If we could get the work done using an older version or a free alternative then we might save some money that we can use for whatever it is we do here in the academy.

I think that the XML format under Microsoft Office is being used as a distraction from real practical issues of interoperability. Whether this is good for Microsoft in the long run is for them to work out but I am pretty convinced that some of the solutions coming out of Microsoft Research are no good for me or my colleagues and I will continue to say so and continue to work with my team to do things in a way that benefits all of us.