Journal 2.0: Embedding semantics in documents

I started a series of posts on Scholarly HTML but I think it makes more sense to call the idea I have for web based journal publishing systems Journal 2.0. I don't have enough time to write this all up in as much detail as I would like and I know from the comments that I am confusing people, but I wanted to get down one more idea before I go on leave for ten days.

One of the goals of Scholarly HTML, sorry Journal 2.0, will be to have embedded semantics in documents, like being able to say that the string *Peter Sefton* represents the name of the author of this post, or that when I say Toowoomba I am talking about a town on the Darling Downs in Queensland, not some farm in South Africa.

There's a lot of it about right now, this push to embed semantics in documents. Chris Rusbridge, for example would like to be able to do this too, and not just in HTML, in PDF as well. Chris commented on my exploration of citation using simple links:

> There is a developing, if stagnant microformat standard for citations, but they can surely best be expressed in RDF.

Yes well, they can be expressed in RDF and RDF can be embedded in HTML using RDFa, but I can't see how we might get authors to deal with RDF or RDFa directly because constructing RDF is hard even if you leave out the syntax issues. That's why people are looking at tools like Microsoft's Ontology add-in for Word 2007, about which a few words have been exchanged on this blog. Chris reminded me of Bryan Lawrence's exploration of adding RDFa to his wiki which I think illustrates that you can do RDF in a wiki, but that only very dedicated authors will master the syntax.

I want to explore the idea of using simple links for embedded semantics as I did for citations.

Here's how it might work in a simple case:

1. I want to say that I am the author of a paper, so I type a version of my name: P. M. Sefton. This could be in a word processor, a wiki, WordPress anything that supports simple hyperlinking.
2. I go to a website, lets says it's called http://ontologize.me and use the relation builder tool there to pick from a list of ontologies, like geographical, species, chemical, etc I pick “Dublin Core Metadata”, and choose the “Creator” relation. The site then asks me to find a web page for my creator at which point I can put in a link to my own site, to an author page in a repository, or search one of the proliferating researcher identity services.
3. The ontologize.me website will then give me a link I can use in my page to link the text P. M. Sefton. It would look something like this (I have left the URL unencoded so you can read it)

   > http://ontologize.me/metadata?o=http://ptsefton.com/about#me&type=http://dublincore.org/2008/01/14/dcelements.rdf#creator

4. If I click on that link, the the ontologize.me site could show me a 'sentence' saying:

   > Peter Sefton **is the creator of** <referring page>

   Behind the scenes there would be some RDF which a machine could ask for instead of the HTML view.
5. If I upload a draft document to a journal website it could take the link apart and turn it into RDFa so it can be part of the semantic web. This is a bit complex as there is a literal string “Peter Sefton” involved which is related to the URI for me which has a type of Dublin Core creator and the page in question has “Peter Sefton” as it's creator. I think that's right.
6. And to disambiguate the string Toowoomba, I could choose a geographical ontology and navigate

to the right page. To make a link like this:

http://ontologize.me/content?o=http://www.geonames.org/2146268/toowoomba.html

Now this scheme would work technically (I think) if people had to hand-code these links, but obviously **nobody would use it**.

But with a good interface at the mythical ontologize.me it could be relatively simple to use. Even better would be to embed support into a writing tool, so you could make a Word or OpenOffice.org Writer add-in which would make it easy to manage. I think the Microsoft Word ontology add-in finds strings in your document that match those in an ontology for example and auto-tags them.

But **still nobody would use it** unless (a) the results are really useful, which is unlikely unless there is large community doing this or (b) there is an incentive, like a journal requires it, or a department insists on it for PhD theses.

Now, I'd like some feedback on this, and particularly I would love some help writing out the RDFa for examples like the author/creator link above.

Anticipating some feedback.

Won't the documents be incredible ugly full of all those links?

Remember the links are for the authoring stage. How the semantic relationship are presented once the work is up as a Scholarly HTML work would be completely up for grabs, I can image an author image and summary popping up if you hover over a name, or a map if mouse-over Toowomba for example. If this is built into a tool then the links might not even show up as such, but are there for interop.

Wouldn't it be better to use real embedded RDF?

Well of course, and that's what Microsoft are doing with their ontology add-in but at the expense of interoperability. Others have dropped by the blog to suggest using the new features of OpenDocument, and again I say that's fine but we don't have tool support yet.

Please tell me if I making some fundamental error apart from the error of judgment in inventing a whole new infrastructure web application that needs to be written.

Finally, a note on how I got to here. I have been talking around this idea for a while, trying to figure out how you could add RDF semantics to an existing URI, like a URL for an author page, by appending parameters or suchlike. But the problem is that this could break things. Then it struck me that RDF is all about having URIs – so what if we had URIs for these relations. The initial idea was inside out relative to the one above and it involved adding parameters to a URI like so (again this is not URL encoded so you can read it):

http://ptsefton.com/about#me?
http://ontologize.me/metadata=true&type=http://dublincore.org/2008/01/14/dcelements.rdf#creator

The idea was that this would probably continue to work, ie it would take you to my author page, but when used to anchor some text it would encode a bunch of RDF triples. But to support his you'd need a web service anyway so I settled on doing everything from ontologize.me.

In summary, I think this could potentially help prime the semantic web infrastructure. A site like ontologize.me would be a kind of wizard for relating things to each other, and could serve as an online gateway to established ontologies, with an extreme approach to interoperability which need not be incompatible with existing laudable efforts like the Word add-in.