Before Beyond the PDF: Authoring tools for document semantics

By Peter Sefton

Summary

In this post I am going to demo some infrastructure that I think will be useful in scholarly communications; a way of encoding an RDF triple, in this case some metadata, in a plain-old URL. Why? Because I want to show how we can encode semantics in ways that will survive being put into word processors, blogged, emailed, saved as PDF and maybe even being put in Google Wave. The plain-old-http-link is universally supported across all sorts of environments so it's a good way to get interoperable document semantics.

Note that this encoding scheme could be supported with easy to use interfaces plugged in to Microsoft Word, OJS, repositories and WordPress and so on, but it can be deployed as simply as having a website that says:

> To assert that Peter Sefton created this resource, wrap this link around his name in the text:
>
> http://ontologize.me/meta/?r=http://purl.org/dc/terms/creator&o=http://trove.nla.gov.au/people/541658

The same technique could be used to refer to scientific terms, or proteins, or data sets, but with an explicit statement of what the link means.

I will go through some of my thoughts about the Beyond the PDF workshop, describe this approach in more detail and finish up with a demo.

Relationship to Beyond the PDF

Over on the Beyond the PDF list there has been copious discussion on the merits of PDF. There has been quite a bit of support for the PDF format, from various camps including those who like the explicit formatting, and those who value the atomic integrity of PDF as a kind of unit of academic currency. I agree that the a significant appeal of PDF is that it is a well-bounded single file, so it is 'easy' to deal with. (Of course it's all too easy to lose track of, unless it has decent embedded metadata and you're using a tool that supports it, which is not all that common, but tools are coming for that.)

In contrast, **web-content**, even though it is potentially much easier and cheaper to make dynamic and engaging than PDF **is not easy to capture and save in a single unit**. That's changing though, there are two obvious promising candidates for making atomic web documents, or aggregations of web resources that people can actually use because tools are here, right now.

1. **EPub which is a zip-based format** with HTML inside, has the ability to load-in any kind of extended content you like, so is a potential candidate for re-flowable content plus supporting data, plus, even a PDF. EPub is supported on all major platforms including mobile devices.
2. **HTML 5 manifests** also let you make packages – so an article page in a journal can include a list of all the stuff your browser needs to download if you decide to 'Save as App... ' - again, including all sorts of data and visualisations, even annotation clients and, of course, PDF. I linked to a simple demo that my team at ADFI put together in my last post.

There are (at least) three obvious container formats for research articles. arguing about which is best is pretty pointless, because it is easy enough to support all three – they're just simple transformations of each other, give or take some inherent limits in PDF as a packing format.

The real issue, I think, is how we capture, husband, and nurture all this content. Arguments about whether or not PDFs are worthy are somewhat missing the point (as I understand it) of the workshop – to look at research and publication processes, workflows, business models and tools for doing a new kind of research.

We need authoring and workflow tools to deal with all of this stuff – **before the PDF.**:

1. Capturing and identifying data – so it can be referenced in papers, theses, blog posts, emails, etc.
2. Capturing and identifying the documents we're working on.
3. **Labelling and describing the above using unambiguous metadata frameworks.**
4. Packaging all of the above using some abstract model that captures relationships between different resources.
5. Maybe, in some disciplines where there is a demonstrable ROI, being able to embed machine-readable semantics in publications.

I think that one of the best places to start looking at what goes **before the PDF** is with **metadata** – all research articles need it and we can use it to explore how machine-readable semantics embedded in documents might work in other contexts.

A task: embed machine readable metadata inline in a word processor

So, in this post I want to look at one task: how do I assert that I am the author of a paper I'm writing a machine-readable, robust way that lots of downstream services can support once the paper leaves my care.

Some assumptions:

1. I'm an Australian researcher.
2. I use a word processor to write my papers.
3. I tend to blog works-in progress, and I'm working in a domain where my impact depends on this. That is, blog posts are an important form of scholarship.
4. I do write scholarly articles, often they start life as conference papers and are then turned into journal articles later.

I'm going to use this post as an example. I'm writing it using OpenOffice.org Writer, I have access to the conversion services provided by ICE, which allow me to post the document to an Atompub content management service, in this case WordPress.

The aim is to **add some metadata identifying me as the author.** Some colleagues and I explored a number of ways to do this using a word processor (Sefton et al. 2009), and they're supported by the tools I'm using but actually they're clumsy and fragile – and I think I have found a better way that does not depend on styles.

I want to:

- **Use my people Australia ID** minted by the National Library of Australia. http://nla.gov.au/nla.party-541658 . Yes there are other forms of ID but it's widely recognised that the major providers are going to work in a distributed way so if I use this one – if I get an ORCID, the the NLA.
- **Say** that the person identified by " http://nla.gov.au/nla.party-541658" is **the author**.
- Have my blog (and other downstream services) **understand and advertise** that I am the author.

So, here goes. Let's add some semantics to this blog post.

I will ad a by-line to the top of my blog post using my word processor.

> By Peter Sefton

Now, I can link that to People Australia.

> By Peter Sefton

Done!

Actually, no. All I have done is add a link – good enough for an English speaker to work out that I'm the author, and to provide a nice identifying endpoint.

So what I'd really like to be able to do here is make the relationship explicit using semantic web technology.

I head off to the RDFa primer. Looks like I have to do something like this:

```
<span xmlns:foaf="http://xmlns.com/foaf/0.1/"
xmlns:dc="http://purl.org/dc/elements/1.1/" property="dc:creator"
rel="foaf:maker" resource="http://trove.nla.gov.au/people/541658">

Peter Sefton

</span >
```

This is way better than my first attempt, and it seems to work at the Sindice RDFa inspector; people on the list have been helping me, particularly Paul Groth.

But how do I do that in a word processor?

I don't. It's just not supported. Even experienced HTML wranglers would have trouble with stuff if they had to do it by hand.

Hey, what if I could go to People Australia and on that page it said something like:

> To assert that P M Sefton is the author of a resource, use this link: http://trove.nla.gov.au/people/541658?prop=dc:creator

Then in my word processor I could use that link. I could add something to my WordPress site so that when I send it links like http://trove.nla.gov.au/people/541658?prop=dc:creator it would generate the RDFa for me. And OJS could look for links like that, and the local ePrints site, and so on.

But People Australia doesn't support that (yet) – it's not really the done thing to add extra attributes to a URL that resolves to someone else's site.

What if there was a site somewhere else, though, that let me generate URLS like that?

As far as I know there is no such 'real' site, but I do have a demonstration of what it might look like. Try this:

> http://ontologize.me/meta/?r=http://purl.org/dc/terms/creator&o=http://trove.nla.gov.au/people/541658?prop=dc:creator

That's my (clumsy) attempt to encode a triple in a link. My demo service is pretty basic – and doesn't do content negotiation, although Duncan Dickinson did do a better version of the service.

Subject	<The referring page>
Predicate / property	dc:creator

Object	http://trove.nla.gov.au/people/541658?prop=dc:creator
Format (not implemented)	http://purl.org/triplink/v1

So, now I can put in my byline. (And note that this scheme doesn't care if I am calling myself Peter M Sefton or Petie Sefton or ptsefton.)

> By Petie Sefton

And now all I have to do is hack WordPress to recognise this kind of little microformat (nanoformat?) - which might be called RDF-l or Triplink or something.

I went ahead and did that, using a very simple WordPress plugin (no doco, definitely no warantee needs WordPress v3.0.1 which loads some jQuery-based Javascript into the browser/client and processes the link-nanoforrmat into RDFa. This approach is only suitable for a quick demo, as a lot of the clients that might consume RDFa are not going to run Javascript – this will include RSS feeds. I really need to write some PHP to run in WP itself– looks like QueryPath is what I want.

You can try it out for yourself. When this post loads in your browser, then the HTML wrapped around the link to my name should look something like the sample RDFa above.

There a lots of things to do if anybody but me and Duncan Dickinson think this is worth pursuing further:

1. Define an extra parameter in the nanoformat so that these kinds of **links are identifiable**, whether they resolve to Ontologize.me or some other site. I will probably register a PURL for this, the links would be identifiable by `&triplink=http://purl.org/triplink/v1` or the like.
2. Work out the **syntax** of the nanoformat in more detail with input from RDF experts.
3. Most importantly, make tools.
    - **Word processor plugins** that make this transparent. I have argued that the the MS Word ontology plugin could use this to make the tool (a) much more interoperable, and (b) add relations to the mix – at the moment it links text to ontological terms with no predicate, giving no more semantics than a plain old link.
    - Websites for ontologies and name authorities that provide **meaningful links people can use** to express useful relationships.
    - **Plugins for content management systems**, PDF readers etc. But note that even without plugins these 'triplinks' are usable, you can click them and have them resolve to something useful, and harvesters could look for them in HTML, Word, PDF, email etc and process them.
    - **Services like SWORD** and repositories to look for link-encoded semantics when you upload stuff to them.

References

Sefton, P. et al., 2009. Embedding Metadata and Other Semantics in Word Processing Documents. *International Journal of Digital Curation*, 4(2). Available at: http://www.ijdc.net/index.php/ijdc/article/view/121 [Accessed October 22, 2009].

This post was written in OpenOffice.org, using templates and tools provided by the Integrated Content Environment project and published to WordPress using The Fascinator.