Boundaryless eResearch: Use Linked Open Data

I am at the eResearch Australasia conference, official tag #eraust09. Yesterday afternoon Anna Gerber, Peter Murray-Rust and I convened a Birds of a Feather (BoF) session: Boundaryless eResearch: use the Web, use Linked Open Data.

I put up some thoughts about what I'd like to see from the session. It was well attended, and the discussion took off nicely into issues well beyond technology. Jim Richardson tweeted it, summarised here.

Anna's wrap-up is here:

> Linked Open Data is not a silver bullet that tries to resolve all issues surrounding publishing of data online: questions of data quality, privacy, institutional policy, persistent access, and so on, remain. It is a technical solution, providing a set of principles for publishing data so that it will be discoverable from and linkable (mashable) with other data available on the web.
>
> When describing the LOD vision, we often talk about a web of data, however we need to consider how to integrate this emerging web with our existing web of documents (as well as with datasets and documents that are not yet published online). Technologies like RDFa that can be retrofitted to existing web documents to link them with LOD datasets will be crucial towards this end, as will enabling researchers to publish their data with RDF generated from tools that they already use within their research practice (like word processors, workflow tools, etc).

Here are the few slides we used to seed discussion on the day, with a couple of late additions. A talk by Anne Cregan earlier in the day provided a good introduction to Linked Open Data: Linked Open Data: a new resource for eResearch.

Linked data (ptsefton)

TBL says:

> Like the web of hypertext, the web of data is constructed with documents on the web. [...]
>
> - Use URIs as names for things
> - Use HTTP URIs so that people can look up those names.
> - When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
> - Include links to other URIs. so that they can discover more things.
>
> Simple. In fact, though, a surprising amount of data isn't linked in 2006, because of problems with one or more of the steps.[...]
>
> http://www.w3.org/DesignIssues/LinkedData.html

Simple? (ptsefton)

What does this mean?

Does everything that cites the URL http://creativecommons.org/licenses/by-sa/2.5/au/ get the license?

[Professor Tom Cochrane, co-leader of the Creative Commons project for which QUT is the institutional partner for Australia just happened to be there. It turns out that the convention of linking one of the little CC badges is what signifies intent to license a work, although my wording is probably good, but the point stands that there is not unambiguous way to tell a machine the difference between a link which is intended to license the work and one which is citing or pointing at the license for some other purpose.]

Our goal: insulate users from this stuff (ptsefton)

```
<!-- /Creative Commons License -->

<!--

<rdf:RDF xmlns="http://web.resource.org/cc/"

    xmlns:dc="http://purl.org/dc/elements/1.1/"

    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">

<Work rdf:about="">

   <dc:date>2005</dc:date>

   <dc:creator><Agent>

      <dc:title>Peter Murray&amp;#45;Rust</dc:title>

   </Agent></dc:creator>

   <dc:rights><Agent>

      <dc:title>Peter Murray&amp;#45;Rust</dc:title>

   </Agent></dc:rights>

   <dc:type rdf:resource="http://purl.org/dc/dcmitype/Text" />

   <license rdf:resource="http://creativecommons.org/licenses/by-nc-
nd/2.0/" />

</Work>

<License rdf:about="http://creativecommons.org/licenses/by-nc-nd/2.0/">

   <permits rdf:resource="http://web.resource.org/cc/Reproduction" />

   <permits rdf:resource="http://web.resource.org/cc/Distribution" />

   <requires rdf:resource="http://web.resource.org/cc/Notice" />
```

```
   <requires rdf:resource="http://web.resource.org/cc/Attribution" />
   <prohibits rdf:resource="http://web.resource.org/cc/CommercialUse" />
</License>
</rdf:RDF>
-->
```

Is that valid? http://www.w3.org/RDF/Validator/ARPServlet [PMR]

1. Issues with URIs for everything (AG)

- persistence of URIs - what is the lifespan of the data
- how to mint new ones (persistent identifier services eg ANDS)
- human readable vs non-human readable (eg http://example.or/data/myexperiment vs http://handle.net/109348934587893745349587329 4)
- human readable may become outdated/ no longer accurately describe the content
- who makes choices on these policies? (researcher/institution/community)
- how to discover the URI used by other members of the research community for a given concept?
- use of dictionaries/thesaurus or agreed databases eg LOC, dbpedia
- academics may not be comfortable using terms from non-academic sources eg wikipedia/dbpedia

3. Issues with RDF as URI (AG)

- how much data/metadata to describe using RDF?
- raw RDF vs styled/presented vs embedded (RDFa)
- what about data/documents that have already been published - adding RDFa
  - level of granularity of markup
  - maintaining integrity of documents if embedding RDFa
  - publishing at existing URIs adding #thing1 #thing2 for existing URIs
- what vocab/ontology to use (reuse, mix them up)
  - even if there is an established ontology for the domain, applying ontology terms to a given data set may not be straightforward

4. Issues with links to other resources (AG)

- providing these links at the URI as RDF vs querying stand-off triple stores: the same data may be used by several research communities - should data specific to each community be directly available at that URI or in discipline-specific datastores?
- Provenance of data : can it be trusted? in many cases the data being published is not the raw data - need to identify who has interpreted/cleaned up the data

Licensing – the 'open' bit (PMR)

The biggest danger is not making the assertion that the data is open.

The Panton Principles: http://wwmm.ch.cam.ac.uk/blogs/murrayrust/?p=1939

[Via Jim Richardson more resources: http://bit.ly/2eZlOG]

Example DbPedia (PMR)

```
 SELECT ?resource

WHERE {

?resource <http://dbpedia.org/ontology/Person/birthPlace>
<http://dbpedia.org/resource/Sydney> ;

<http://dbpedia.org/ontology/Person/deathPlace>
<http://dbpedia.org/resource/Sydney>

}

ORDER BY ?resource
```

http://dbpedia.org/snorql/?describe=http%3A//dbpedia.org/resource/Sydney

```
 http://dbpedia.org/snorql/?query=SELECT+%3Fresource%0D%0AWHERE+{+%0D%0A
%3Fresource+%3Chttp%3A%2F%2Fdbpedia.org%2Fontology%2FPerson%2FbirthPlace
%3E+%3Chttp%3A%2F%2Fdbpedia.org%2Fresource%2FSydney%3E+%3B%0D%0A%3Chttp
%3A%2F%2Fdbpedia.org%2Fontology%2FPerson%2FdeathPlace%3E+%3Chttp%3A%2F
%2Fdbpedia.org%2Fresource%2FSydney%3E%0D%0A}%0D%0AORDER+BY+%3Fresource
```

Discussion points (all)

- Who has used RDF? How did it go?
- AG:
  - Embedding data in documents and web pages eg via RDFa is crucial in linking existing web of documents with emerging web of data
  - Don't sweat the details just get the data out there in an open format

- ptsefton:
  - **Getting tools into the** hands of researchers so they can 'do' linked data.
  - **Getting the web into the scholarly communications process as a first-class citizen:** Scholarly HTML.
  - **Bringing the web to the desktop.** The Fascinator Lensfield Anna & team's Firefox Add In for creating compound semantic-web objects, for literary scholars.

Examples

- **The Aus-e-Lit project** is a NeAT-funded project that aims to address the eResearch needs of researchers involved in the study of Australian literature and Australian print culture. The project enhances and extends the existing AustLit web portal with data integration and search services, empirical reporting services, compound object authoring, editing and publishing services and collaborative annotation services.
- **The OREChem project** is a Microsoft-funded collaboration between Cambridge, Cornell, Indiana, Penn State and Southampton Universities that aims to make existing chemistry data sources available as LOD, to develop new LOD creation resources using grid computing, to develop and converge on standard ontologies for chemistry knowledge representation, and to further the state of the art in extracting semantic chemistry data from published PDF.
- **Talis Connected Commons** is an initiative whereby Talis offer free Linked Data infrastructure for open datasets. Nick Day and Jim Downing at the University of Cambridge are working to publish semantic data from the CrystalEye[8] system through Talis Connected Commons. No, we need a r*elation* something like *has-license* – the challenge is to let ordinary researchers.

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.