ANDS Metadata Stores: Integrating VITAL with the NLA's Party Infrastructure Project

Here is some more news on the metadata stores project for ANDS (see previous posts) and how we might build links between VITAL/Fedora and an identity service for people (parties in ANDS-speak). This is potentially a step on the way to a linked-data future not just for research data but also for institutional repositories.

Natasha Simons said, on the CAIRSS list:

> An exciting new project, which was mentioned at the CAIRSS Community Day last December, has commenced at the National Library of Australia (NLA). The Australian Research Data Commons Party Infrastructure Project (ARDCPIP) is an Australian National Data Service (ANDS) funded project being carried out by the NLA to enable improved discovery of research outputs and data through assigning public and persistent identifiers to Australian researchers and research organisations using the NLA's People Australia Infrastructure. The project will involve consultation with the research sector and more information about it will be made available to you shortly.

This is good news for the metadata stores work we're doing with ANDS. I was in Canberra in early February and we caught up with Basil Dewhurst and Natasha; seems like the timing is good to try to get a link-up between the work we're doing to sketch the design of a metadata store for data on top of the VITAL repository.

Why is this important?

One of the problems we have been talking about with ANDS is how to represent people. In VITAL names are typically not subject to authority control, and even if they are it is via some kind of people-driven workflow, not a feature of the software. (Any exceptions to this? Please let me know in the comments.)

The linked-data semantic web ideal would be for there to be some kind of resolvable identifier stored each time a party is mentioned. You want at least two bits of data (using the example that Basil gave in the comments here last time).

1. The name as it appears on a particular work that's a string, like Cappo, Michael Charles.
2. A URI – such as http://nla.gov.au/nla.party-471077
3. And maybe a canonical, normalized way of referring to the name for a particular service – the heading at the NLA's trove site is Cappo, Michael.
4. Possibly more URIs that refer to the same party.

If we had this it would meant that submitting data about data collections to the Australian Research Data Commons would be much more concise and accurate – it should not be necessary to include any party data at all in a feed from a metadata store to the commons, just URIs.

But, in the case of University X, not all the researchers have persistent identifiers. So, I have suggested that we try to create persistent identifiers for all parties as the first step in building a metadata store on top of VITAL. This is good several ways:

1. Having an early adopter on board will help the PIP project at the NLA.

2. This will bring together three ANDS funded projects and promote general metadata harmony.
3. University X will get a significant upgrade to the integrity and usefulness of their existing data.

How?

What we're looking at is a pre-load stage to populate PIP/People Australia with the identities of people already mentioned in the repository.

1. **Harvest name data from VITAL** - could be via MARC / DC / or some other format derived from those, over OAI-PMH or via some other batch load. Contextual information is likely to be limited to subject and affiliation data pretty much - you won't get birthdates. Vicki Picasso tells me that at Newcastle they record which authors belong to their institution, so for that institution this process could concentrate on the authors they know are theirs.
2. **Records are auto-matched in PIP** using whatever algorithms it has.
3. A person (probably at University X) uses the NLA's private People Australia tools to s**ort out the name references and associate the various publications with the right person ID.**
4. Once the name data has been checked (step 3) the developer writes some code to **inject the ID data back into Fedora**, in MARC (if that's possible) or into a new datastream. (If my team got the job we would do this by writing some custom index rules for our Fedora indexer which is part of modular system that is The Fascinator.)
5. The developer creates an ingest system (like VALET). As part of the form-filling process it would look up name IDs at point of ingest and trigger creation of new parties when one is not found. We will write up more detail about how this would look, using Newcastle as an example.

There is a variant on this where NicNames might be used in between 1 & 2 and PIP would be fed cleaner data.
Once the intial load was done, many names associated with data collections would already have URIs so these would not have to be re entered; for 'new' identities there would have to be some kind of system for minting a new ID then flagging it for attention from a name-curator at a later time.

Linked data more generally

I have described a process here for party identities; it would also be good to have the same kind of approach for activities (in the ISO-2146 sense) which means **research projects** in this context. Andrew Treloar of ANDS Monash branch tells me that there has been talk of having the major funding bodies like ARC provide a PIP-like service for grant funded projects, providing some kind of URI that a human or a machine could look up, getting a web page or an RDF record as appropriate. The RDF record could be used to extract trusted data like project names at both the repository and the ARDC registry but nobody would have to re-type that data into a form.

Does anyone out there know anything about this potential service for Activities? Can you comment?