What should we call this name-authority, vocabulary-server, linked-data URI factory service we are building for ANDS?

Today I am at the National Library, at the "Names Round Table". I'm part of the ARDCPIPAG, which stands for "Australian Research Data Commons Party Identifier Advisory Group". We're advising a team at the library who are building party-identifier services for researchers and research institutions in Australia.

At my work, at ADFI we're working on developing specifications for Metadata stores for ANDS. The first cab off the rank is an application architecture developed with The University of Newcastle and Swinburne University. I posted about this earlier this year. ANDS agreed that it would be a good idea to take a linked data approach to the design of the application. Linked Data is a non-threatening way of talking about the The Semantic Web, which is definitely coming real soon now. No, really, 2010 will be the year, or at the very outside 2012.

Or 2013.

The linked data rules (from Tim Berners-Lee)

Like the web of hypertext, the web of data is constructed with documents on the web. However, unlike the web of hypertext, where links are relationships anchors in hypertext documents written in HTML, for data they links between arbitrary things described by RDF,. The URIs identify any kind of object or concept. But for HTML or RDF, the same expectations apply to make the web grow:

1. **Use URIs** as names for things

2. **Use HTTP URIs** so that people can look up those names.

3. When someone looks up a URI, **provide useful information, using the standards (RDF, SPARQL)**

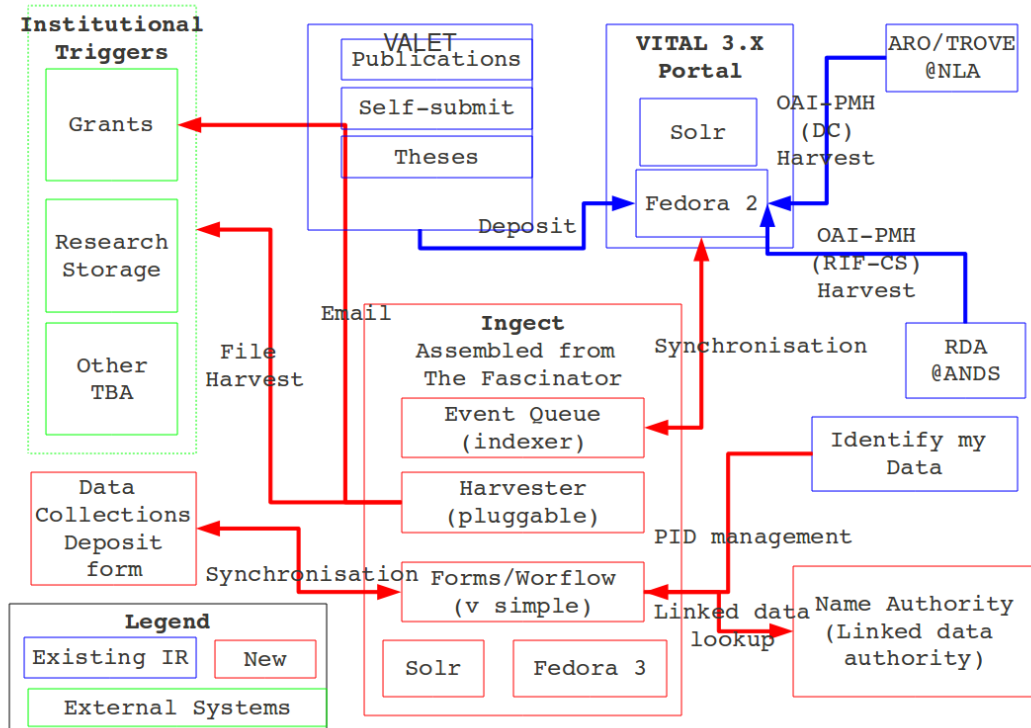4. **Include links to other URIs**. so that they can discover more things.

Simple. In fact, though, a surprising amount of data isn't linked in 2006, because of problems with one or more of the steps. This article discusses solutions to these problems, details of implementation, and factors affecting choices about how you publish your data.

http://www.w3.org/DesignIssues/LinkedData.html

Our metadata stores work is covered in blog posts here.

My presentation today focussed on the name-authority part of the architecture and looked at the process of establishing name-identities at an institutional level **before** joining in a broader federation. I asked the audience, what is the name we should use for this class of service? (We're still looking for a name for the metadata-stores app as well, something better than "The Fascinator, Research Metadata Store Edition, Pro" or the current working title of "Ingect".

Architecture - "Ingect" (working title)

Institutional Triggers
- Grants
- Research Storage
- Other TBA

VALET
- Publications
- Self-submit
- Theses

VITAL 3.X Portal
- Solr
- Fedora 2

ARO/TROVE @NLA

OAI-PMH (DC) Harvest

OAI-PMH (RIF-CS) Harvest

RDA @ANDS

Deposit

Email

File Harvest

Ingect
Assembled from The Fascinator
- Event Queue (indexer)
- Harvester (pluggable)
- Forms/Worflow (v simple)
- Solr
- Fedora 3

Synchronisation

Synchronisation

Identify my Data

PID management

Data Collections Deposit form

Linked data lookup

Name Authority (Linked data authority)

Legend
- Existing IR
- New
- External Systems

What does a Linked Data Approach mean for a metadata stores project?

1. **No more typing name-strings** into web forms.
2. **Agreed names/URIs** for things like resource-types.
3. **Sorting out URIs** for things so that (unlike with Institutional Repositories) we can:
   i. **Agree on terms** before we start.
   ii. **Match-up terms later** if we don't reach agreement.

In our forthcoming collaboration with the UoX we wanted to make sure that when we named the things in the institution with the role of 'research' or 'owner' of data we used URIs. In ANDS speak, these things are 'parties'. Some of the are people, some are institutions, organisations, or organisational units.

Goal: Establish URIs for person-parties at UoX before we start

So, a simple matter of:

- **Extracting name data** from the existing institutional repository
  (the UoX model is to build their metadata store so collections are described in the IR)

- **Disambiguating name-strings**, for the usual reasons; there is typically more than one string used to refer to the same person and often the same string used to refer to more than one person. Sorry, party.

- **Establishing new URIs** for each party that the UoX cares about (they don't care about parties from UoY).

- **Injecting the URIs back in to the IR**.

Establishing IDs? How?

We considered these options.

1. **Use People Australia** / ARDC-PIP services. (It was a bit early)
2. **Use Nicnames** (plus).
3. **Use a combination** of the above.

We chose option 4

Build a name-authority server inspired by and informed by the NicNames work.

Why?

1. Can ship as an integrated part of the broader application we are building, in the same language (Java).
    i.   **Easy-install** under the same web container as Fedora and Solr.
    ii. **Same configuration files** as the "Ingect" application.
2. Having a local service means we can use:
    i.   **Private data such as staff-IDs** internal to the application.
    ii. **Local URI-schemes** for local things, such as internal projects.
3. Our service will **deal with JSON formatted , not RDF** to make it easy for web-interface designers.

The process – disambiguating names

1. Import all the name-string/publication pairs from the IR into a new repository (in EAC format) .

   That's one for each name on a paper. Records will be presented as citations keyed by a name.

   ```
   name:<name-string> id:<pID> title:<dc:title> (subject:<dc:subject>)*
   ```

2. Import a set of canonical names we care about into the new repository.

3. Turn all the canonical names from a local directory into master-records/name-packages.

4. Allow a data librarian to drag all the name-strings into the name-packages.

   ```
   [Jane Hunter CONTAINER RECORD

   Hunter, J: Gerber, A. and Hunter, J. (2008). A compound object
   authoring and publishing tool for literary scholars based on the
   IFLA-FRBR. In: C. Rusbridge, A. Trefethen and D. Berry, Proceedings
   of: 4th International Digital Curation Conference "Radical Sharing:
   Transforming Science?". 4th International Digital Curation
   Conference (IDCC 2009), Edinburgh, Scotland, (1-10). 1-3 December
   2008.

   Hunter, Jane: Hunter, Jane (2001). Adding Multimedia to the
   Semantic Web: Building an MPEG-7 Ontology. In:International
   Semantic Web Working Symposium (SWWS), Stanford University,
   California, (). July, 2001.

   Hunter, J.: Hunter J. (1999). An "Improved" Proposal for an MPEG-7
   DDL. ISO/IEC JTC1/SC29/WG11, 47th MPEG Meeting M4518, .

   ...

   ]
   ```
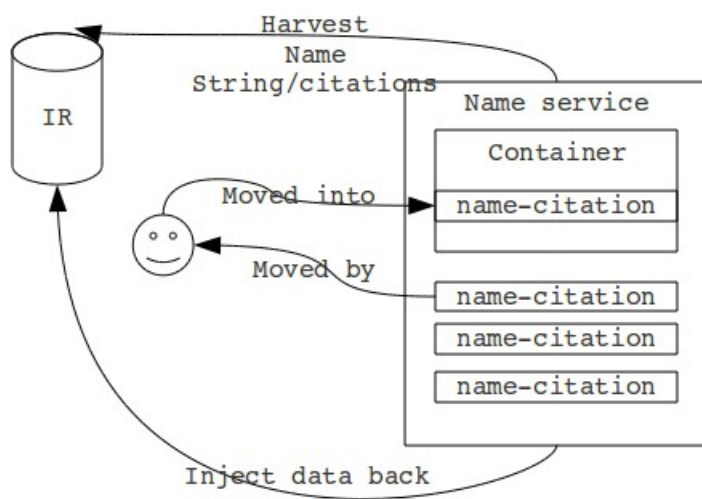
Process

But wait! There's more!

This new module, which is called, um … will:

- 'Mint' a new URI whenever someone types a string (in desperation).
- Provide locally available, fast services for accessing any ontology. Eg:
  - FOR codes.
  - Geonames.

One final thing

(And this bit is not ANDS funded)

We want to make it possible to encode and decode web-semantics in a URI. Eg:

http://ontologize.me/meta/?r=http://purl.org/dc/terms/creator&o=http://nla.gov.au/nla.party-541658

I can use this to assert that "ptsefton" is the author of this resource.

Would ANDS and/or the NLA consider supporting services like this?

Oh, and nobody at the ARDCPIP round table had an answer for me about what we call this class of name-authority linked-data-endpoint-factory application.