Metadata Stores – CAIRSS presentation - by Simon Pockley

I’ve been thinking that the most useful thing I could do with this session about Metadata Stores is to take you through what ANDS means by Metadata Stores, what we want these funded projects to achieve, why we think these aims are important and how they fit into the context of research data management in Australia. I will also touch on how the selected universities are going about their projects, when we expect these projects to be completed and where some challenges are emerging.

ANDS has offered research intensive institutions up to $225,000 to assist with the deployment of:

- a metadata store for research data collections,
  - with connectors to institutional sources of truth,
  - with connectors to external sources of truth,
  - with coverage over the entire institution,
  - and - with feeds to the Australian Research Data Commons.

What are Metadata Stores?

The term is shorthand for a store of metadata about research data collections and associated entities.

Just as the deployment of repositories has been characterised by discussions about a service based approach, for ANDS, the term Metadata Stores really points to a data management infrastructure rather than Institutional Repository space.

ANDS has always been very clear about its high level objective to have Australian researchers easily publish, discover, access and use research data through what it calls its Four Transformations of data that are:

- Unmanaged - managed
-Disconnected - connected
-Invisible – findable
-Single use - reusable

But if we unpack what we could mean by data management infrastructure, you might get a better sense of where we see the Metadata Stores Program fitting and why it is important.

What do we mean by data management infrastructure?

We can make a distinction between 'capacity' - the overall ability to manage research data, and 'capability' - a measurable attribute of capacity. We think that data management capacity can be defined though a set of nine capabilities. These nine capabilities can be broken down into a range of activities or sub-components. None are exclusive. For example, licensing could be a component of storage, policy and discovery. While the following list separates the capabilities, it should not be thought of as a linear sequence. These capabilities are interrelated and three dimensional. When they are closely aligned in their objectives and are working together, they form a coherent data management infrastructure:

- **Researchers:** Identify researchers and their research
- **Data:** Identify the creation & location of data
- **Governance:** Library, IT and Research Office working together
- **Policy**: Activities supported by effective Policy and Procedures
- **Tools**: Deployment of integrated tools
- **Discovery**: Metadata for discovery and access
- **Storage**: Data stored/shared with access controls
- **Skills**: Training programs - staff and researchers
- **Integration**: Capabilities integrated across institution(s)

One approach to building capacity is to build services in the hope that people will use them. Another approach is to fund projects with the proposition that capacity can be built by working on, and completing these projects.

If we look at the footprint of the *Seeding the Commons* Program in relation to the nine capabilities, we can see a strong emphasis on discovery capabilities but little on integration and almost none on storage.

![Figure 1. ANDS’ *Seeding the Commons* capability footprint (universities ordered by research output)](image1)

The *Data Capture* Program was focussed on building and integrating tools.

![Figure 2. ANDS’ *Data Capture* capability footprint (universities ordered by research output)](image2)

The *Metadata Stores* Program has a much more defined footprint with fixed deliverables and a focus on integration through the requirement that collection records come from at least three Faculties or their equivalent. There is also an emphasis on connecting researchers to their data through internal and external sources of truth.
Combining these Program footprints gives us a sense of the relative strengths of capability that has been enabled by ANDS in the Australian University sector. Although the signal strength is stronger at the Go8 end where most funding has taken place, there are clear indications that there may be stronger integration at the less research intensive end.

N.B. this ANDS 'enabled' capacity is not necessarily the only source of capabilities. Many universities have developed their own capabilities and these are not reflected in this matrix.

What I’m hoping you will better understand is the context in which the Metadata Stores Program is being conducted. Under the EIF funding rules ANDS can fund:

- The building of software infrastructure that supports research activities. i.e.
- Specification development for software
- Software development at enterprise level
- Installation, configuration and testing of software

The funding rules put some constraints around what these projects. ANDS can only encourage the other non-software dimensions of these projects that are so integral to the success of these projects. Here I am specifically referring to Governance, Policy and Integration.

Nevertheless, there have been a range of software choices and to a varying extent two communities of interest have emerged around these choices. ReDBox is a system developed with ANDS funding by Queensland Cyber Infrastructure Foundation Ltd (QCIF). ReDBox is described by its project manager, Duncan Dickinson as a research data catalogue. It was two parts, ReDBox for data and Mint for
people and other authority files. If you are interested, then I recommend you watch the short video interview with Duncan: [http://www.youtube.com/watch?v=Nfz1MBOpKJ]. The CSIRO have developed their own software as has the University of NSW who will be deploying Mint for people. University of Queensland has taken a component approach because they have some major software deployments on the horizon and need to be flexible. ANU is using the ORCA software that supports Research Data Australia. I am about to post some of the rationales for these software choices on the Metadata Stores blog.

Emerging challenges

Although all projects share a set of mandatory deliverables, one of the challenges associated with community building around the Metadata Stores projects is that they are mostly at different stages.

- Dirty data – where integration with sources of truth is exposing shortcomings in those sources
- Change management – where project managers are not in a position to influence cross institutional decision making
- Skills – where there is an emerging shortage of skilled people with data management expertise
- Continuity - where the high level decision makers at universities can suddenly resign and leave a vacuum in cross-institutional oxygen supplies
• Sector silos – where the university sector is poorly integrated with community and government sectors where data reuse could be of most value.