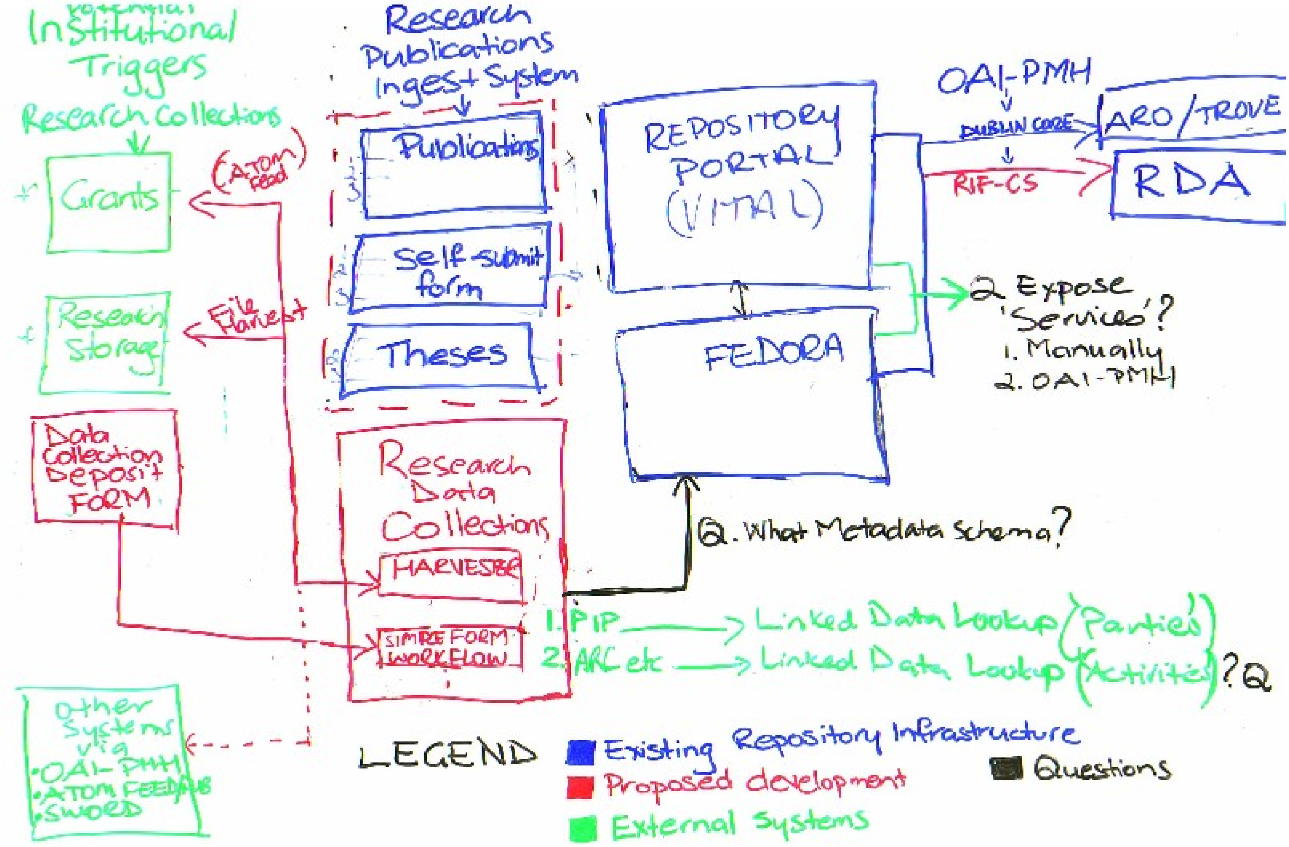
More details on a metdata store for data in/alongside VITAL

Here's another post about the ANDS metadata store work I've been doing. I was at the university of Newcastle last week working with Vicki Picasso and Dave Huthnance, with calls to Teula Morgan at Swinburne. Together we fleshed-out a model for how Newcastle might run a data-registry alongside their VITAL repository.

I want to use this blog post to expand on the last blog post and refine the details of what will become an ANDS project plan for someone to develop this stuff. I'm going to talk about the diagram from the last post, in particular the application which has become known as "The Red Box".



1 Requirements

The main requirements for this system, as collected from the stakeholders are that it:

1. Can provide a **university-wide registry of research data,** (the data collections will reside wherever they currently reside under access control, and I believe there is a policy in development that open data will all be served from the research storage service) with two main inputs:
    - Form-based deposit, with dead-simple workflow using a system which is as easy for the library and research office to customise as the current VALET system, with an option to port the existing ingest workflows as well (that's why there's a dotted red line around the current forms system).

- Discovery of data files and data collections on the university's new research storage facility.

2. Work alongside of the VITAL institutional repository by storing metadata (and potentially small collections themselves) in Fedora, which is the underlying storage layer for VITAL
   - **VITAL can be used as the portal for access to research collection metadata or not**, as appropriate. Teula Morgan told us that at Swinburne they would likely connect the research data collections into their discovery layer rather than using VITAL. The following models need to be supported:
     - VITAL as the authoritative source for digital objects, with RedBox deleting objects which it has handed off to VITAL.
     - RedBox as the authoritative source for digital objects, with VITAL or another portal acting as the discovery interface for research data collections. With the addition of a portal, this configuration could form the basis of a standalone metadata store.
   - The system needs to **stay out of VITAL's way as much as possible** to avoid the risk of unapproved material 'leaking' through VITAL and to avoid affecting performance.
   - **The system should provide for batch-changes to repository content**, in order to assist in cleaning up the document data that is already in there prior to adding research data. (This component came up both in discussions about how we can work in a linked-data way – adding URIs for people mentioned in the repository and as a requirement from Swinburne where they lament the lack of batch-editing tools in VITAL).
   - The system needs to have **no dependencies on the proprietary parts of VITAL**, to allow the possibility of switching repositories.
3. As far as possible **deal only with the collections side of things**, without attempting to become management system for activities (research projects) or parties (people). Following from this, to be a “Linked Data” application where Activities and Parties are referred to by URIs, and other terms also use well-define URIs rather than strings. (There is a potential side-project to this one to upgrade the NicNames system – more on that soon).

Caroline Drury the CAIRSS/ANDS liaison person pointed out that we should discuss how these requirements go beyond encouraging people to register data via the ANDS 'register my data' service for Research Data Australia. The big things are:

1. This allows an institution to have additional management metadata that is not in the RDA system, such as details of how long data should be held. We're working with Simon Porter from the University of Melbourne on this. The idea as that when this is implemented with University X they will work out the metadata they need and the developer would assist in creating the input forms and mappings from one metadata format to another.
2. This provides for curation by a data librarians who can leave some submissions in the queue while they sort out details of the metadata or wait for data to be made available on the storage facility.
3. This also encompasses data which is not destined for or not ready for listing on RDA.
4. But most importantly it will allow the institution to meet its obligations under The Code.

2 Implementation

The deliverable for the work we're doing at USQ for ANDS is an ANDS project plan. That makes sense, as it makes starting up the next phase of work straightforward. But to put forward a complete plan, we need to make some assumptions about the design, mainly in what technologies we would use. So here's a proposed broad-brush architecture for a metadata stores solution work with VITAL. Remember this is only a proposal an the reason it's up on this blog is so you can comment on it.

The large-scale assumptions are:

1. RedBox will use **Fedora 3 as an internal storage component –** with data synchronised to VITAL as needed. This meets the requirement that VITAL is functioning as a portal and will reduce stress on the VITAL repository as much as possible. Fedora is an obvious choice for ARROW/VITAL sites. We are choosing to work with the latest version for the RedBox component, but it will need to synchronise with Fedora 2 which underpins the VITAL product.
2. The application will be **developed in Java,** building on The Fascinator platform which was originally sponsored by the ARROW project in 2008 and which has been under development at USQ since then. Benefits include:
   - Being Java it can sit in the same Tomcat web-server as Fedora and the Apache Solr indexer used by most Fedora repositories these days.
   - The ingest component we're developing for The Fascinator, while incomplete, meets the requirement that the system be configurable in a similar way to VALET – where extending the forms and integrating them with external systems like CrossRef is trivially easy. Existing VALET forms can also be ported to the new system (it's a manual process, but not difficult).
   - It's highly modular and so can be used, for example, without a portal, a role which VITAL, or a discovery service can take on. One of the most important modules will be plugin harvest technology to pick up content that's on the storage system and provide a view to researchers and data librarians to begin describing it. The Fascinator has as extensible system of plugins and we already have file-indexers and a framework for extracting metadata from files, which can be extended to work with new kinds of research data as they appear.
   - Our developers know the system, meaning we can be up and running with this application very quickly.

   We are aware of other Fedora software components, but none that meet all of the above criteria. The closest would probably be Muradora – if there are ANDS contributors using that who want help setting it for research data then I think that would be in scope for us to look at in our ANDS work,
3. While there was some discussion about using VALET or VITAL as the foundation for the RedBox early in the project neither of these systems has an architecture which can work on a university-wide scale if all data sets are to be described, or the ability to harvest metadata about files on the research storage service.

2.1 Components

I'll take a brief look at some of the components in turn.

2.1.1 Batch editing

The Fascinator already has the basis for a batch-change tool. The Fascinator, whether it is sitting on top of Fedora or our simple file back-end has indexer component which can either watch a queue for changes in a repository or do a complete re-index.

This indexer has an extensible set of rules which are small scripts that can be fired off to deal with various kinds of content. In our desktop work we use this for things like generating web-ready versions of video content and HTML versions of documents, but it could also be used to transform datastreams in the repository to make bulk changes.

Here's an example of how it might be used. I talked previously about setting up a system with the NLA or locally using NicNames to assign unique IDs to researchers before we start the major part of this project. Once we had a set of name identities, they could be put back into the repository by writing an 'indexing' rule that for each document does a look up to the name authority system and puts it back into the MARCXML datastream in an appropriate field. Then you tell the system to re-index.

For a batch edit, you'd back everything up, then set this up and run it on a copy of the repository and swap in the new data and test thoroughly with VITAL, before either running the process on the live repository or swapping in a new Fedora database underneath.

2.1.2 OAI-PMH feeds

One of the key deliverables for this project is to get data flowing to Research Data Australia. Part of the project scope will be to make sure that we have fully-functional OAI-PMH feeds, complete with support for deletions, working from both VITAL's copy of Fedora and from the RedBox itself to meet the requirement that the system is able to run in “headless” mode. Xiaobin Shen from ANDS in Melbourne has been working with OAI-PMH providers, so we don't expect this part to be hard., just a matter of careful selection, configuration and testing.

2.1.3 VITAL configuration

One of the major tasks is to create customisations for VITAL to display metadata records about data and make sure that the result fits in to the rest of the VITAL repository portal.

2.1.4 The Forms interface

While we're not proposing to use the Squire forms interface sponsored by ARROW directly, the code we will be using is based on and informed by the same VALET model. This system consists of:

- Form-templates using the same Velocity template engine for Java as VITAL uses, to make this easy to deploy for ARROW sites.
- Form widgets that allow for linked-data lookups. The idea will be to use sources such as People Australia or a NicNames instance to provide a URI for a name; type in “John Smith” and it will give you a list of John Smiths and the areas in which they work to pick from, with similar lookups for research projects. Where there is no suitable John Smith then the forms application will create a new one with a temporary URI via a call to the name-authority.

  Another thing the ingest system will need is a link to the Identify My Data service or a local equivalent for those sites who are sold on the benefits of using Handles as persistent identifiers for their collections[1].

[1] I'm not convinced that using the ANDS handle infrastructure is a good idea. Firstly, ANDS is only funded for a short time and while ANDS staff express their aspirations for keeping things going there are no guarantees, secondly, data has to have URL pointing to where it is stored, and that has to be maintained as well, with redirects and so when things change. In a lot of use-cases I think that using Handles just increases complexity, cost and risk.

- Simple workflow descriptions like the following sample:

```
"stages": [
   {
                "name": "init",
                "security": ["owner", "admin"],
                "visibility": ["guest", "owner", "admin"],
                "template": "${user.home}/.fascinator/workflows/templates/basic-init.vm"
           },
           {
                "name": "live",
                "security": ["owner", "admin"],
                "visibility": ["guest", "owner", "admin"],
                "template": "${user.home}/.fascinator/workflows/templates/basic-live.vm"
           }
     ]
```

2.1.5 Research storage harvest & grants database triggers

The University of Newcastle has a research storage facility (a SAN) and policies are under development which will likely see lots of RDA-ready data made available on the facility. One of the features of the RedBox application will be to he able to harvest metadata about such files; at the very least, file-paths, sizes and dates, and any other metadata that can be automatically extracted. For data that's on the storage facility the idea is that researchers can add it to the metadata store by finding the data files themselves and clicking “Describe this data” to fill out a form.

The Fascinator has a number of features in this area such as the ability to generate thumbnail versions of files and web-ready renditions of various formats, which is of peripheral interest to the metadata stores activity.

The harvester system is highly configurable, so it could be set up to index, or watch other kinds of storage service; one that Newcastle require is the ability to harvest an email account with messages from the grants database about grant completions; which are events that should trigger a metadata librarian to chase-up data from researchers.

3 Risks

There will be a detailed risk assessment for ANDS, but here are some notes about the main risks:

Risk	Mitigation Plan
The software we develop here might end up being only used at one or a handful of institutions, which would then bear the maintenance load.	• Work to promote this solution to other ARROW sites. • Release all code as open source with tested documentation. • Use components of this solution as part of the standalone solution we have also been asked to look at, broadening the installed base for the the RedBox application. • Consider funding a program to port VALET workflows to the new system for ARROW sites to build a sustainable community. • Work with the University of Melbourne and collaborators to see if some of the component developed for the RedBox application could be used at their sites, and to ensure compatibility with VITRO as a data store. • Document the metadata storage system and batch transformation system so that new Fedora-compatible ingest or portal tools can be swapped-in later.
University X unable to supply full metadata schema on time; there is no clear consensus on best practice for describing research data collection to meet the demands of The Code, while being able to serve RIF-CS to Research Data Australia.	• Stakeholders to vigorously encourage ANDS to produce metadata guides, possibly after workshops. • If all else fails implement RIF-CS in the repository., with the possibility of doing a batch-update later.
Linked Data infrastructure including URI endpoints for parties and activities doesn't come on-line in time for implementation	• Work with the PIP project to make sure that this does not happen (at least for names) possibly entering name data into People Australia semi-manually. • Work with ANDS to set up interim systems where possible. • If all else fails store strings in metadata fields just as we have been doing in IRs for years and update to URIs later when the infrastructure is available.

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.