ANDS metadata stores: Describing metadata collections in VITAL

I'm at the University of Newcastle visiting repository rat extraordinaire Vicki Picasso (actually at this bushland campus she should be a repository wallaby or something) and her colleague Dave Huthnance from IT. We are working on a model for how research data collections destined for Research Data Australia might be described and managed in the local institutional repository.

(Please ANDS can we have some advice on this metadata issue? Some of you say to use RIF-CS and some say that's a bad idea.)

Vicki presented a model of how metadata about research data could be ingested into the VITAL repository they use at Newcastle at eResearch Australasia 2009; it featured the VALET system which is a very simple repository ingest tool and what Vicki calls “Institutional Data Triggers” such as events in a grants database which would fire-off a metadata ingest workflow.

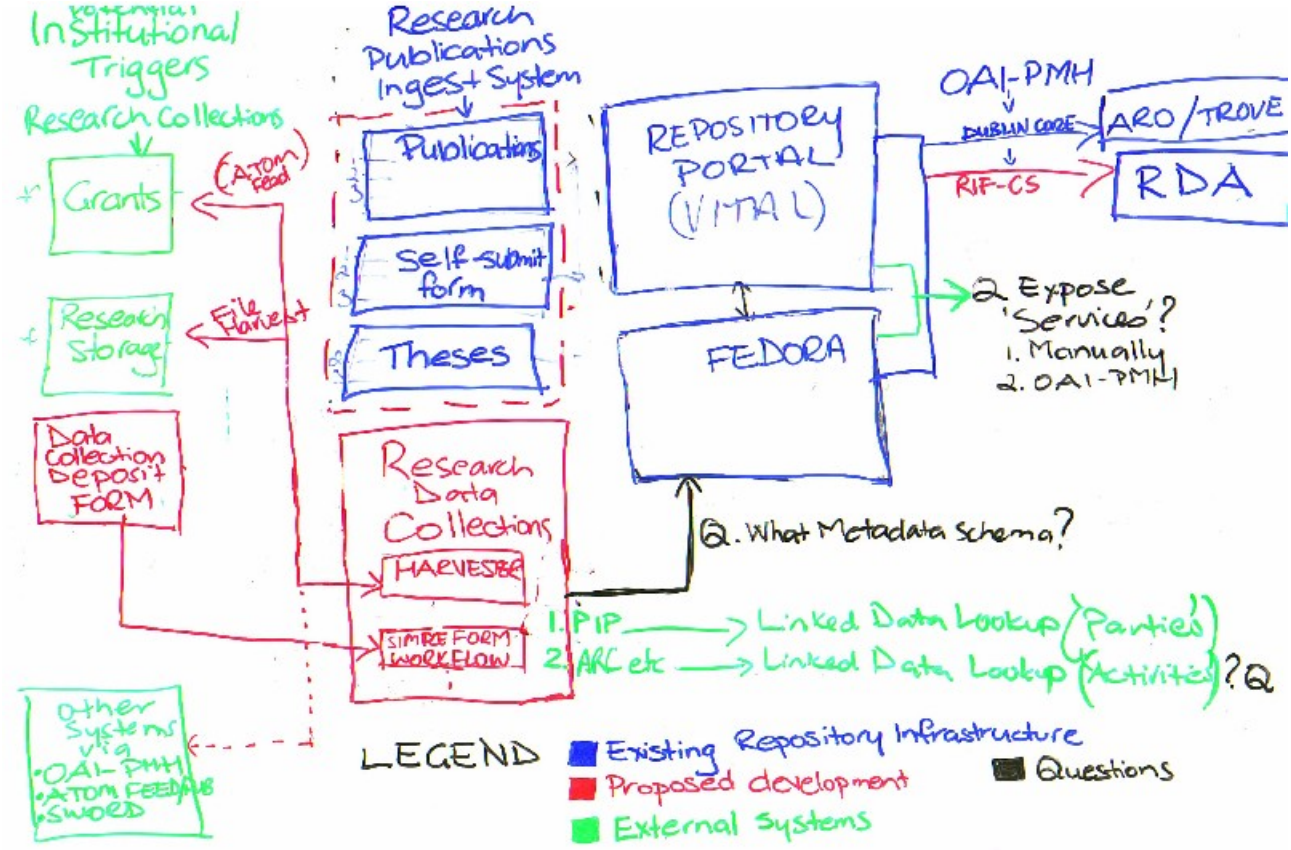
The new model

Today we refined that diagram. Like it says the **blue bits** represent the current Nova repository infrastructure (VITAL + VALET + Fedora) which feeds data to (amongst other things) the National Library's harvesting systems.

The red bits are new proposed infrastructure, to be developed, to enable collections metadata to be captured and feeds of RIF-CS metadata to Research Data Australia. The new red box labelled “Research Data Collections”, should it be built, will be a more sophisticated version of VALET, probably written in Java so it can work in the same Tomcat web container as Fedora – it would have a VALET-style simple forms interface for walk-up submissions (this *could* be used to replace the existing publications ingest and staging workflows too, as shown by the dotted red line – if this were a requirement).

Green is for external services. One of the very interesting green bits is the Research Storage system which is being provided by university IT and administered by the Research Office. I gather that this is essentially a file-store; we are proposing to add an interface that lets researchers see their files in the new (red) ingest system and add metadata to them, and flag them as candidates for RDA. I think Newcastle's policy will be that if you want data to be available via Research Data Australia then it is desirable this it goes in the Research Storage System. Sounds good to me. To bridge the gap between files on a storage system we are proposing a bit of middleware to link the file view of data to a web/repository view.

As discussed before here, the ANDS stakeholders in this project are keen for us to take a linked-data approach to metadata (slogan: Less typing, more linking!). I talked a bit about how this might work in the previous post on name identities; potential integration with services like the NLA's PIP/People Australia and possible services like an ARC website for grant information are shown in green at the bottom right of the diagram (I have some input from Basil at the NLA I need to process, but at this stage I think we're looking at having NicNames in there so institutions can manage their own metadata.



One assumption we're making here is that the core class of item we're describing here is a collection, which should fit with the kind of data that is already in the repository, which is **research outputs**, like data.

There are some questions, of course.

1. What metadata schema to use for describing data collections?
2. And where would the ISO2146 notion of Services fit in? The services listed in the RIF-CS documentation are all repository-type search/feed services so it seems appropriate to either tie them in to the OAI-PMH 'identify' verb or to let repository managers simply enter them in to an ANDS system directly.

Going further

One idea that has come up is that VITAL sites might want to use Fedora and the OAI-PMH feeds available off it but not expose them via a web portal at all. In conversation with Teula Morgan from Swinburne today, Vicki proposed a model where there is no portal interface. I call this a 'headless' approach; there would be local management interface for research data collection metadata (the red box) but it could be that the primary discovery mechanism is outsourced to RDA. This is pretty common for university web sites – USQ uses Google for our website search service for example.

I am also exploring the idea that this ingest tool, which will be able to put records into Fedora (which as far as I know nobody has ever been fired for acquiring) could form the basis for our major deliverable on our

ANDS metadata stores project; a specification for a stand-alone metadata -about-research-data-collection system.

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.