Building an Australian User Community for Vivo: Profiling Research Data for the Australian National Data Service

Simon Porter[1], Lance De Vine[2], Robyn Rebollo[3]
[1]University of Melbourne, Melbourne, Australia, simon.porter@unimelb.edu.au
[2]Queensland University of Technology, Brisbane, Australia, l.devine@qut.edu.au
[3]Griffith University, Brisbane, Australia, r.rebollo@griffith.edu.au

INTRODUCTION

The Australian National Data Service (ANDS) was established in 2008 and aims to: influence national policy in the area of data management in the Australian research community; inform best practice for the curation of data, and, transform the disparate collections of research data around Australia into a cohesive collection of research resources. Research Data Australia is a web site established by ANDS for the purpose of describing data collections produced by or relevant to Australian researchers and to promote visibility of research data collections in search engines.
In order to populate Research Data Australia, ANDS has established projects with most Universities across Australia, under two main programs:

1. Seeding the Commons – activities aimed at stimulating the systematic collection of research data records so that they can populate Research Data Australia
2. Metadata Capture – activities aimed at the automation of research data capture processes to improve the collection of high quality metadata on research in the everyday process of conducting research

ANDS also has a third program of funding; Metadata Stores, aimed at ensuring that Universities have sustainable mechanisms in place, for both collecting information on research data collections, and for allowing Research Data Australia to harvest appropriate collections well beyond the initial ANDS funding cycle. As part of activities associated with the Australian National Data Service, an increasing number of Australian Universities are choosing to implement VIVO to profile information about institutional research data sets, both locally and as part of a national data commons. To date, the University of Melbourne, Griffith University, the Queensland University of Technology, and the University of Western Australia have all chosen to implement VIVO, with interest from other Universities growing.

This presentation will outline why VIVO is a natural fit within the ANDS metadata stores program, specifically:

- Why providing quality records about research data means also providing quality information about researchers, grants, publications, and university organisational units. How the needs of an ANDS metadata store can be aligned with University policy requirements for a central research data registry
- How the VIVO ontology has been extended to fully represent the Registry Interchange Format - Collections and Services (RIF-CS) Schema, a data interchange standard developed by the Australian National Data Service for supporting the submission of metadata to the Australian National Data Services' collections service registry[1]
- How re-usable data transformation and workflow components are being developed for transforming and ingesting institutional data into VIVO
- How Australian Universities driving the development and implementation of VIVO plan to assist other interested parties in their implementation efforts through the packing of a customised Metadata Exchange Solution that includes a RIF-CS ready ontology and data workflow components.
- The functionality required to integrate information collected in VIVO back into non Semantic Web information infrastructure.

VIVO

In 2009 the National Institutes of Health in the United States funded a $12.2 million project to create web based infrastructure to fascilitate the discovery of researchers and collaborators across the US. The resulting project is called VIVOWeb and is built upon Vitro, a technology developed at Cornell since 2003 and which has since become part of the larger application called VIVO. VIVO is an open source semantic web application that allows institutions to ingest and link institutional metadata and allows users to browse and search while ensuring that the institution retains control of how data is accessed. In Australia, various additions have been made to VIVO to support the requirements of the ANDS metadata stores program including a) an extended ontology which fully represents RIF-CS, b) an OAI-PMH provider for providing OAI-PMH feeds, c) customised web page templates and d) workflow modules to support data ingest and transformation.

STANDARDS ADOPTION

An important aspect of a metadata store which connects to diverse data sources is that it makes use of widely recognised standards. VIVO makes extensive use of W3C and other popular standards to fulfill this requirement. VIVO includes:

- The use of various XML manipulation standards such as XSLT and XQuery for the manipulation of metadata.
- An extensibile OWL ontology that includes data elements for populating the RIF-CS schema and for supporting the submission of metadata to the ANDS collections service registry. It makes use of other standard ontologies such as DC, FOAF, SKOS, BIBO and others.
- The use of SPARQL for querying and manipulating Semantic Web data. VIVO makes extensive use of Semantic Web standards and technologies and can leaverage off of the large amount of work being carried out in the Semantic Web community. Metadata in VIVO can be exposed as Linked Data via a SPARQL endpoint.
- An Open Archives Initiative Protocol for Metadata Harvesting [4] (OAI-PMH) provider service for exposing selected metadata for harvesting. OAI-PMH enables standardised, automated, regular harvesting of metadata and is included in VIVO via the use of OAI-CAT [2], an open source OAI-PMH provider solution.
- A well defined abstraction layer for the provision of Persistent Identifiers (PIDs). Different standard PID provider software may be "plugged in" to this layer. A Plugin for harvesting ANDS Handles for use in Research Data Australia is included.

METADATA INTEGRATION, SEARCH AND DISPLAY

The University of Melbourne, Griffith University and the Queensland University of Technology have jointly extended the core VIVO ontology to support the population of RIF-CS metadata elements. A comprehensive crosswalk table was developed from the original ANDS' Metadata Content Requirements document. The crosswalk table includes a list of RIF-CS required and recommended schema elements for describing ANDS registry objects. Additional columns that identified RIF-CS equivalent fields in a research enterprise system and the Vitro Ontology were included for the purpose of mapping each standard metadata element to another. Institutional metadata ingested into VIVO and stored according to the VIVO ontology is mapped into RIF-CS and exposed for harvesting via OAI-PMH. Only metadata elements marked as harvestable are exposed via OAI-PMH.

Metadata is displayed via dynamic web pages and can be searched using either full text search or the SPARQL query language. Web pages for each of the RIF-CS classes of objects are provided. The user is also able to "walk the semantic graph" by navigating from one linked object to another. Research data and activity networks are captured well using VIVO's semantic web technology.

DATA TRANSFORMATION AND WORKFLOW

A key part of extending VIVO for use by Australian universities has been the development of modularised software components that allow flexible construction of data workflows for ingesting data into VIVO from diverse institutional data sources. Such components have been built using the Kepler workflow engine [3] with many new components being added to Kepler specifically for fascilitating ingest and transformation of data in VIVO. The source code for these modules can, however, be used separately if so required.

PACKAGING VIVO FOR OTHER UNIVERSITIES

Work at the University of Melbourne, Griffith University and the Queensland University of Technology will result in a freely available package that will enable an institution to install and customise VIVO as a Metadata Exchange Hub for aggregating institutional metadata and for exposing selected data elements for harvesting. Documentation will be provided containing detailed instructions for installation and maintenance procedures as well as examples showing how to customise web interfaces and how to set up ingest workflows and data transformation. Examples will also be provided showing how to map institutional metadata schemas to the extended VIVO ontology.

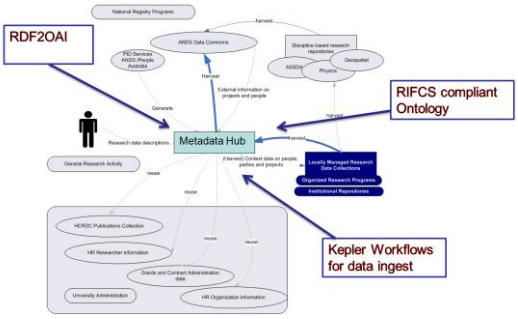

Figure 1. System Architecture

CONCLUSION

This presentation will outline why VIVO is a true 21st century technology that fits naturally within any University metadata infrastructure, and especially how it supports the requirements of Research Data Australia. We will describe how the VIVO ontology has been extended, how the metadata crosswalks constructed, what data workflow modules have been created, how interfaces can be customised and how the system can be deployed at other universities.

REFERENCES

1. *The Registry Interchange Format* - Collections and Services. Available from http://www.ands.org.au/resource/rif-cs.html, accessed 28 June 2010.
2. *OCLC's OAICAT.* Available from http://www.oclc.org/research/activities/oaicat/default.htm, accessed on 28 June 2010.
3. *Kepler* . Available from *https://kepler-project.org/*, accessed on 30 June 2010.
4. *Open Archives Initiative http://www.openarchives.org/*, accessed on 30 June 2010.