Developing the Discovery Layer in the University Research e-Infrastructure

Malcolm WOLSKI, Joanna RICHARDSON, Mark FALLU, Robyn REBOLLO, Joanne MORRIS
Division of Information Services, Griffith University
Brisbane, Queensland 4111, Australia

Abstract

Governments worldwide are faced with the challenge of creating research e-infrastructures to not only manage but also make accessible and discoverable increasingly large amounts of research data. Universities in turn are under pressure to ensure that their research strategies and support services are aligned with these national imperatives. This paper describes a nationally funded Australian university initiative to build a research e-infrastructure layer which connects individual researchers and the University to the Research Data Australia service in order to expose details of their research activity as well as available research data outputs. As governments work towards fully functional e-infrastructures which will be both cross-disciplinary and cross-border, the semantic metadata exchange service described in this paper offers a model which supports the interactive discovery of, and navigation to, content that may reside locally or across the world.

Keywords: Research infrastructure, VIVO, semantic web, Vitro, discovery systems, Kepler

1. Introduction

In a submission to the European Commission, Kroes [1] writes: "Information and Communication Technologies (ICT) are the most recent transformational factors in science. They enable close and almost instantaneous collaboration between scientists all over the world and they provide access to unprecedented volumes of scientific information." ICT have helped to create a world in which knowledge—and its application—is seen as a key to global competitiveness and national prosperity is viewed as underpinned by knowledge innovation [2]. Within this context, governments worldwide are grappling with the challenges of creating robust research e-infrastructures which can not only manage this information but also ensure its discoverability and accessibility.

2. Australian National Data Service

As part of the Australian government's NCRIS (National Collaborative Research Infrastructure Strategy) initiative, the Australian National Data Service (ANDS) was formed to support the "Platforms for Collaboration" capability. The service is underpinned by two fundamental concepts: (1) with the evolution of new means of data capture and storage, data has become an increasingly important component of the research endeavour, and (2) research collaboration is fundamental to the resolution of the major challenges facing humanity in the twenty-first century [3].

ANDS is building the Research Data Australia (RDA) service [4]. It consists of web pages describing data collections produced by or relevant to Australian researchers. RDA publishes only the descriptive metadata; it is at the discretion of the custodian whether access, i.e. links, will be provided to the corresponding data. Behind RDA lies the Australian Research Data Commons (ARDC) which is the infrastructure and systems needed to support data and metadata capture, publication feeds, and applications such as data integration, visualisation and analysis.

3. ANDS Objectives

The long term (ten year) objectives for data management within the Australian National Data Service (ANDS) are to:

- Increase the amount of research data that is routinely deposited into stable, accessible and sustainable data management and preservation environments
- Enable Australian researchers to discover, exchange, reuse and combine data from other researchers and other domains within their own research in new ways
- Facilitate the sharing of Australian data to support international and nationally distributed multidisciplinary research teams
- Support the development of data management services and support within institutions that promote good data management practices for researchers

Key stakeholders in the Australian research environment—ANDS, National Library of Australia, funding bodies such as the Australian Research Council and the National Health and Medical Research Council, research institutes and universities—all have knowledge to be shared. In building its national collaborative infrastructure, ANDS has utilised a federated approach which supports multi-layers, i.e. RDA aggregates at the national level data about Australian research which has been aggregated at the local level.

Critical to the model is the ability to enhance discoverability and accessibility of all aspects of research to improve knowledge communication. The connectivity between research data and researchers is important, especially for purposes of re-use and in cross-disciplinary research. Identifying relationships between people, institutions, projects and the relevant research data created enhances opportunities for collaboration and new research [5] [6].

This paper describes how Griffith University has built a research e-infrastructure layer which connects individual researchers and the University to the Research Data Australia service. The local technical framework developed for the service is based on semantic web, triple store and open access technology.

4. Griffith University's Metadata Exchange Hub

A Metadata Exchange Hub has been developed as part of an ANDS-EIF (Education Investment Fund) funded project involving collaboration between Griffith University and the Queensland University of Technology. The Hub was built to meet ANDS' requirements for institutions to provide aggregated metadata store solutions to populate Research Data Australia (RDA). The metadata feeds encapsulate metadata providing high-level descriptions of research datasets and entities related to them, such as researchers, research groups, research projects and research services. The metadata schema used is the Registry Interchange Format - Collections and Services (RIF-CS) [7], which is a subset of the ISO standard 2146 [8]. The development of a metadata aggregator (Hub) has become a core piece of infrastructure [9].

To populate RDA, the metadata is harvested from institutions via the Open Archives Initiative's Protocol for Metadata Handling (OAI-PMH). This protocol is a HTTP REST based web service with six methods defined for interrogation and harvesting of structured metadata. The default metadata schema for OAI-PMH is Dublin Core, but other schemas may also be used. For the purposes of transporting and aggregating research metadata for RDA, the RIF-CS schema is used. RIF-CS is a high level schema that defines four classes of objects – collections, parties, activities and services. The objects of these classes may be related to each other via relationships defined in a controlled vocabulary [10]. RIF-CS can also be effectively modelled using Resource Description Framework (RDF) and related semantic web standards. See Figure 1.

An important part of Griffith University's Metadata Exchange Hub is to expose the relationships –using RIF-CS—among researchers, their projects and their research outputs, as illustrated in Figure 1. These relationships form a linked graph. For example, Mary Jane (party) has the relationship (is a participant, i.e. researcher) of a project (activity) but also has the relationship (manages) datasets that, in turn, has relationship (is part of) Collection A, etc.

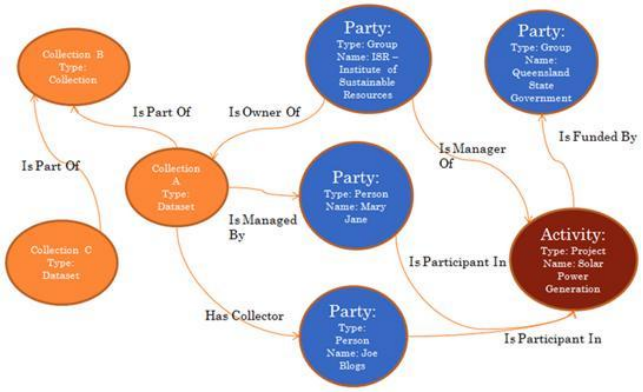

Figure 1: RIF-CS – Linked Data

As part of the ANDS-EIF project, staff analysed the pros and cons of existing software solutions as the potential foundation for the Hub. Since the major project driver was to develop an open source solution which could be used as an exemplar / good practice for Australian universities which want to be part of the national collaborative research infrastructure, the Project Team decided to use a semantic web solution called VIVO as the metadata store, which also includes mechanisms for the editing and display of Hub metadata. Other software used for the project included Kepler [11] for data workflow and transformation, OAI-CAT [12] for OAI-PMH provision, and custom Java code for object Identifier creation.

5. Architecture of the Hub

The following diagram (Figure 2) is a simple illustration of the Metadata Exchange Hub components. VIVO, which is based on technology developed at Cornell, has been implemented with minimal changes to the underlying software architecture. Research activity metadata is uploaded to Research Data Australia (RDA) using the Registry Interchange Format - Collections and Services (RIF-CS).

As part of the Metadata Exchange Hub project in Australia, a number of additions have been made to VIVO to support the requirements of the ANDS' metadata stores program, including (a) an extended ontology capable of fully expressing RIF-CS and modelling research activity in Australian research institutions; (b) an OAI-PMH provider for OAI-PMH feeds; (c) customised web page templates for presentation; and (d) workflow modules, e.g. Kepler, to support data ingestion and transformation. A more detailed explanation of key modules follows.

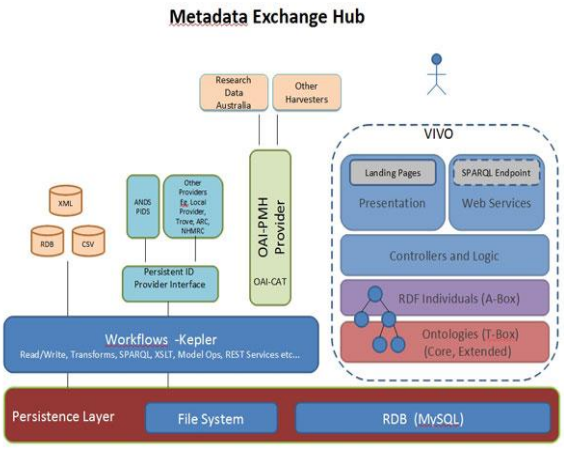

Figure 2: Architecture of the Metadata Exchange Hub

In 2009 the National Institutes of Health funded a US$12.2 million project to create a web-based infrastructure to facilitate the discovery of researchers and collaborators across the United States. This project is known as VIVOWeb and is built upon Vitro, a technology developed at Cornell in 2003 and renamed as VIVO. VIVO is an open source semantic web application that allows institutions to ingest and link institutional metadata; allows users to browse and search; and ensures that the institutions retain control over how their data is accessed. It is fundamentally a Java web application with a persistence layer that represents information using RDF and OWL (Web Ontology Language) and is built on the Jena semantic web framework, a triple store [13].

Ontologies are designed as sets of rules and languages that enable data and information sets of different machine readable systems to be implemented cohesively into one comprehensive domain. An important part of the project has involved the development of a national research-focused ontology, based on the core Vitro ontology, which has been successfully deployed in the first version of the tool. Various extensions were made to the set of standards in VIVO 1.0 to meet ANDS' requirements. Classes were added for research collection metadata and for describing subject codes. Object properties were added for RIF-CS relationships. Data properties were added for identifiers and to support the generation of RIF-CS feeds. Collectively they provide a coherent framework for mapping the bulk of institutional research activity in Australia. The table below lists external ontologies that are included in this customised version of VIVO. Ontologies identified with an asterisk (*) include RIF-CS elements.

Ontology	Namespace
ANDSHarvest*	http://www.ands.org.au/ontologies/ns/0.1/VITRO-ANDS.owl#
Bibontology	http://purl.org/ontology/bibo/
Dublin Core elements	http://purl.org/dc/elements/1.1/
Dublin Core terms*	http://purl.org/dc/terms/
Event Ontology*	http://purl.org/NET/c4dm/event.owl#
FOAF*	http://xmlns.com/foaf/0.1/
FOR 2008 Ontology	http://purl.org/asc/1297.0/2008/for/
geopolitical.owl	http://aims.fao.org/aos/geopolitical.owl#
ns	http://www.w3.org/2006/vcard/ns#
SEO 2008 Ontology	http://purl.org/asc/1297.0/2008/seo/
SEO 1998 Ontology	http://purl.org/asc/1297.0/1998/seo/
SKOS (Simple Knowledge Organization System)*	http://www.w3.org/2004/02/skos/core#
time	http://www.w3.org/2006/time#
Vitro public constructs	http://vitro.mannlib.cornell.edu/
VIVO core*	http://vivoweb.org/ontology/core#
TOA 1993 Ontology	http://purl.org/asc/1297.0/1993/toa/
RFCD 1998 Ontology	http://purl.org/asc/1297.0/1998/rfcd/
Griffith Specific Extensions	accessible via VITRO/VIVO Group

Table 1: Vitro-ANDS Ontology

The VITRO-ANDS ontology is customisable and extensible to cater for specific research requirements. Planned future developments include extensions to better characterise the research outputs of the creative and performing arts sector. Additionally the ontology has been used successfully to model non-research activities such as commercial consultancies.

The core of the VIVO system is an RDF triple store. This is used to model and store data and is an alternative to systems that use traditional relation tables. The triple store can be conceptually divided into two parts: the T-Box and the A-Box. The T-Box (Terminology Box) is the generic data model that describes the relationships between types of institutional data, e.g. projects have Chief Investigators. The VITRO-ANDS ontology forms the T-Box component of the VIVO system. The A-Box (Assertion Box) contains descriptions of specific instances of data, e.g. John Smith (party) has the role (chief investigator) in the project (activity).

There are many software tools and frameworks that may be used for data workflow and transformation. Some are built for a specific purpose or set of use cases. Others are targeted at more general applications. The Hub provides solutions that have been constructed based upon the Kepler workflow software. Kepler also makes use of other standards, languages and software for implementation of parts of its functionality, and sometimes provides a wrapper around functionality found in other software libraries. Some of the benefits of Kepler which made it suitable for the Metadata Exchange Hub project include:

a) Modular system for encapsulating functionality with data-typed interfaces for connecting modules
b) Large library of existing Kepler actors (modules)
c) Relatively easy method for extending Kepler with new actors
d) Cross platform (uses Java)
e) Execution of workflows from both GUI on the desktop and command line on a server
f) Different execution models and flow control mechanisms
g) The GUI can provide an effective means for rapid prototyping

The following Kepler actors were found to be useful for the Metadata Exchange Hub project: FileReader, FileWriter, FileCopier, String To XML, XML Assembler, XML Disassembler, XPath Processor, XSLT Processor, RestService, Open Database Connection, Database Query and PythonActor. The Project Team did have to write some custom actors, particularly for operations involving the Jena API as well as CSV to XML conversions.

A typical simple workflow for ingesting data into VIVO might look like the following: Read CSV file, transform to RDF, merge with existing model and save to Database. The workflow below (Figure 3) implements this and is an example of an ETL (Extract, Transform and Load) process.

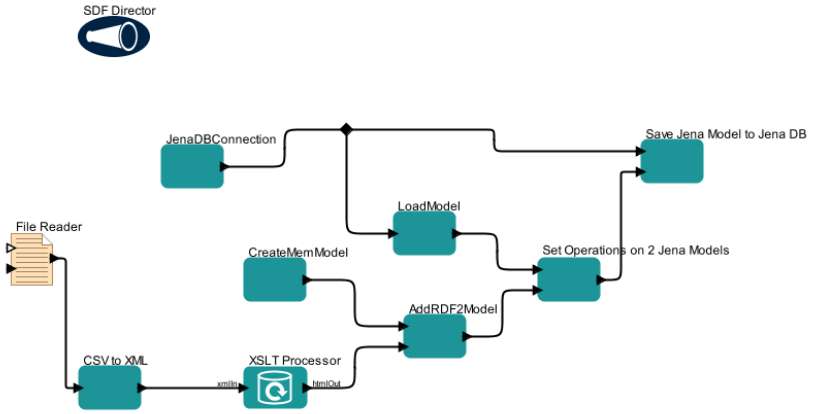

Figure 3: Kepler Workflow in VIVO for Data Ingest

The next diagram (Figure 4) illustrates the workflow that maps the subset of information required for harvesting to RDA. Metadata from the Hub (in the Vitro ontology) is mapped to a RIF-CS formatted feed available via OAI-PMH. The SPARQL query is a parameter that is supplied to the Kepler workflow. Results are automatically serialised as RDF XML and converted to RIF-CS by a custom XSLT routine that maps entities from the internal Vitro ontology to RIF-CS formatted XML, which is then capable of being processed and made available by OAICat

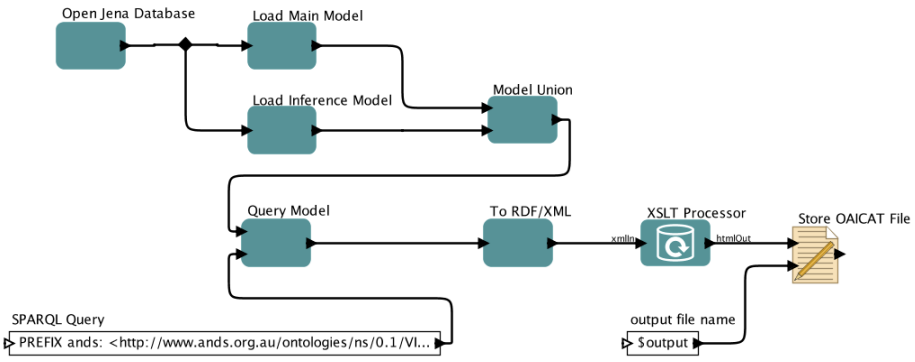

Figure 4: Kepler Workflow in VIVO for Harvesting

The architecture of the Hub has been designed to allow for automatic machine to machine communication for the ingestion of University research activity data. Nominated relevant metadata is harvested from the University's repositories, data stores and corporate systems in its native form. The process of automatic disambiguating entities within data sources can be difficult and has not yet been implemented within the Hub. The VIVO Harvester in the US has done some initial work; however, this is a much needed area of development for the future.

The system first attempts to determine whether persistent identifiers exist for any people or projects in national systems. Key national systems are operated by the National Library of Australia (Trove, People Australia), Australian Research Council (ARC), and the National Health and Medical Research Council (NHMRC). In the case of Trove and ANDS, if no persistent identifiers exist, requests are made (machine to machine) to create an ID, e.g. new researcher person ID.

Kepler workflows automate the translation of metadata from the format in institutional stores to appropriately formatted RDF triples. Kepler workflows then insert the RDF triples representing the institutional data into the triple store (forming the A-Box). This automatically creates human readable HTML landing pages on the fly based upon RDF triples. Links between entities, e.g. People who are Chief Investigators of Projects, are made explicit in the form of hyperlinks. These links are bidirectional, i.e. they link from person to project and back from project to the person. This happens automatically, even if the link was not explicit in the original data store. Kepler workflows then trigger SPARQL queries within VIVO. These queries return all of the research activity data as triples. XSLT is used to transform the serialised triples from the VITRO-ANDS ontology to RIF-CS formatted XML.

The final process is to make the RIF-CS formatted metadata available for harvest via an OAI-PMH interface using the OAI-CAT component. Research Data Australia will periodically harvest the new and updated institutional data via this interface.

6. Discussion

Although the Metadata Exchange Hub is in pilot, metadata collected to date has been harvested by both Research Data Australia and the National Library of Australia's Trove resource discovery system. In addition it is currently being interrogated internally by University researchers. University funds have been allocated as a high priority to move this system into production. Work is underway to finalise the automated updating of research activity data from enterprise systems with an anticipated rollout by mid 2011. The use of Google Analytics will not only provide feedback on system usage but also will expose trends or peaks for subject analysis.

Because the Hub is based on linked open data, the metadata feeds expose the relationships among researchers, their research groups, their projects and their research outputs, including datasets. This means that research information is available for publishing in a "profile". Therefore the Hub creates individual "Researcher Profiles", which provide a history of research undertaken by a respective researcher. Similarly a "Research Group Profile" provides a history of research undertaken by a respective research group, e.g. research centre. Both have links to the actual research data, which supports the ANDS' objectives outlined previously. These "profiles" will be uploaded to both RDA and Trove.

For the postgraduate student, for example, a Profile can be used as a tool to identify seminal research undertaken by experts within the group including their respective supervisor or indeed to select a potential supervisor. The Profiles include the use of visualisation technologies to graphically represent the relationships described above. This allows the student to rapidly follow links in a non-linear fashion without losing the original context.

The Hub also plays an important role through feeds into the University discovery services. For example, Griffith has recently deployed the Serials Solutions' Summon web-based discovery service as the library search / discovery tool. It is now possible to utilise the Metadata Exchange Hub to push key research information through to the Summon library search tool, making it another resource available for scholarly purposes.

7. Conclusion

From the perspective of knowledge communication within the new research environment of universities, it is important that research activity be exposed at the University level in a managed way that creates a rich discovery environment. Semantic Web technology is ideal for use as a federated architecture for integrating metadata from diverse systems. Over time there will be increasing pressure for this sort of automatic and reliable linking functionality in discovery systems.

Because of the ability of the Metadata Exchange Hub to ingest data from a wide range of sources and then export information via filterable views, the system offers value to a number of key elements within a university, e.g. central IT administration, the library, and the research / academic community. In addition, it has wider potential non-university applicability such as for central records, museums, government department archives, and the creation of knowledge bases, i.e. adapting the technology itself for other discovery environments. Therefore, as governments work towards fully functional e-infrastructures which will be both cross-disciplinary and cross-border, the semantic metadata exchange service described in this paper offers a model which supports the

interactive discovery of, and navigation to, content that may reside locally or across the world.

8. Acknowledgement

The authors wish to acknowledge the work of the Research Collection Metadata Exchange Hub Project Team--a joint effort among Griffith University, Queensland University of Technology and the Australian National Data Service--for the technical aspects of this paper.

9. References

[1] High Level Expert Group on Scientific Data. (2010). **Riding the wave - How Europe can gain from the rising tide of scientific data. A submission to the European Commission**. Luxembourg: European Commission.

[2] O'Brien, L. (2010). "The changing scholarly information landscape: reinventing information services to increase research impact". **ELPUB2010 - Conference on Electronic Publishing** (Helsinki, Finland – June 16 – 18, 2010). Available: http://hdl.handle.net/10072/32050 [2010, 1 Nov]

[3] Sandland, R. (2009). "Introduction to ANDS", **Share: Newsletter of the Australian National Data Service** ((issue 1), 1. Canberra, ACT: ANDS.

[4] Research Data Service. (2010). **A Window on the Australian Research Data Commons**. ANDS. Available: http://services.ands.org.au/home/orca/rda/ [2010, 1 November]

[5] Buetow, K. H. (2009). **Speeding Research and Development through a Collaborative Ecosystem**, Collaborative Innovation in Biomedicine. Washington, DC.

[6] Thelwall, M., Li, X., Barjak, F., & Robinson, S. (2008). "Assessing the international web connectivity of research groups". **Aslib Proceedings**, Vol. 60(1), 18 - 31.

[7] Australian National Data Service (2010a). **Registry Interchange Format - Collections and Services (RIF-CS)** Available : http://www.ands.org.au/resource/rif-cs.html [2010, 1 November]

[8] International Standards Organisation. (2010). **ISO 2146:2010**. ISO.

[9] Wolski, M., Young, J., Morris, J., De Vine, L., & Rebollo, R. (2010). "Metadata Aggregation – A Critical Component of Research Infrastructure for the Future", **eResearch Australasia 2010**. Gold Coast, QLD: University of Queensland,. Available: http://hdl.handle.net/10072/34856 [2010, 1 Nov]

[10] Australian National Data Service. (2010b). **Controlled Vocabulary**. ANDS. Available: http://services.ands.org.au/documentation/rifcs/1.2.0/schema/vocabularies.html [2010, 1 November]

[11] Kepler Collaboration. (2010). **The Kepler Project**. NSF. Available: https://kepler-project.org/ [2010, 1 November]

[12] OCLC. (2010). **OAICat**. OCLC. Available: http://www.oclc.org/research/activities/oaicat/default.htm [2010, 20 October]

[13] Krafft, D. B., Cappadona, N. A., Caruso, B., Corson-Rikert, J., Devare, M., Lowe, B. J., & VIVO Collaboration. (2010). "VIVO: Enabling National Networking of Scientists", **WebSci10: Extending the Frontiers of Society On-Line**. (Raleigh, NC, April 26-27, 2010).