INTEGRATION OF SEMANTIC WEB SYSTEM WITH RESEARCH DISCOVERY TOOLS TO ENHANCE POSTGRADUATE STUDY

Malcolm Wolski¹, Joanna Richardson²

¹Griffith University (AUSTRALIA)
²Griffith University (AUSTRALIA)
m.wolski@griffith.edu.au, j.richardson@griffith.edu.au

Abstract

The Australian National Data Service (ANDS) has taken a national leadership role in implementing a national collaborative research information infrastructure strategy. The strategy is designed to build the eResearch capability of Australia’s researchers through national initiatives and strengthened engagement in international initiatives. With the objective of increasing the discoverability of Australian research data collections, ANDS is building the Research Data Australia (RDA) service. With the benefit of an ANDS grant to help populate the RDA, Griffith University has developed a framework for collecting and exposing research data.

The technical framework for the service is based on semantic web and cloud services. The system collects appropriate metadata from research collections within the University through customised feeds from the various University content management and corporate systems. This system then acts as a central University repository to feed information in a standard format to Research Data Australia. It can also expose this data to University library discovery tools and other research information federations where appropriate. The end result is that this service not only integrates and aggregates data within the institution but also provides a key link into national and international open systems.

In addition to facilitating connections between researchers by exposing research activity, the system creates individual “Researcher Profiles”, which is a valuable tool for prospective postgraduate students when selecting their preferred university and area of specialization. The integration of the metadata aggregator system with library discovery tools enhances the overall postgraduate information discovery experience.

This paper describes the project, including the underlying infrastructure and processes. It discusses the use of semantic web technology to aggregate key institutional research data, along with the importance of enhancing discoverability through integration with a variety of discovery tools. The paper concludes with a discussion of how the service supports the research discovery needs of postgraduate students.

Keywords: metadata, research infrastructure, discoverability, library discovery tool

1 INTRODUCTION

Universities worldwide are being challenged on two major fronts. On the one hand there is international consensus that the quality and performance of a nation’s higher education system will be key determinants of its economic and social progress. Universities therefore are under pressure to improve the quality of graduates. On the other hand government funding and policy guidelines are placing pressure on universities to increase their research impact. Governments worldwide are investing in national research information infrastructures to drive national innovation. Because universities clearly have a central role in the generation of knowledge and innovation, they are major stakeholders in national innovation strategies.

In Australia two recent government reports are beginning to have a major impact on the way in which universities are responding to these challenges. The Bradley Report [1] highlights the need “not only to encourage a major increase in the numbers of the population with degree-level qualifications but also to improve the quality of graduates”. The Cutler Report [2] is a key driver for determining the role that universities will play in the revamped national innovation system, with a strong research program seen as fundamental to producing “nationally important research”.

2 CREATING A NATIONAL INFORMATION ARCHITECTURE TO SUPPORT RESEARCH DATA

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 a combination of the set of shareable Australian research collections, the descriptions of those collections including the information required to support their re-use, the relationships between the various elements involved (the data, the researchers who produced it, the instruments that collected it and the institutions where they work), and the infrastructure needed to enable, populate and support the Commons.

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. Core to the model is enhancing discoverability and accessibility of all aspects of research.

This paper describes how Griffith University has built a research e-infrastructure layer which not only connects individual researchers and the University to the Research Data Australia service, but also feeds data into university discovery tools. It examines the flow-on effects for postgraduate students.

3 METADATA EXCHANGE / RESEARCH ACTIVITY HUB ARCHITECTURE

A Metadata Exchange Hub--known at Griffith University as the Research Activity 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 automated metadata feeds to Research Data Australia (RDA). These 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) [5], which is a subset of the ISO standard 2146 [6]. The development of a metadata aggregator (Hub) has become a core piece of infrastructure [7].

To populate RDA, the metadata is harvested from institutions via the Open Archives Initiative’s Protocol for Metadata Handling (OAI-PMH). This protocol is an 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 [8]. RIF-CS can also be effectively modelled using Resource Description Framework (RDF) and related semantic web standards. An important part of Griffith University’s Metadata Exchange / Research Activity Hub is to expose the relationships –using RIF-CS—among researchers, their projects and their research outputs, as illustrated in Figure 1. The connectivity between research data and researchers is important, especially for purposes of reuse and in cross-disciplinary research. Identifying relationships between people, projects and institutions, for example, enhances opportunities for collaboration and new research [9] [10].
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.

As part of the Metadata Exchange / Research Activity 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 which fully represents RIF-CS; (b) an OAI-PMH provider for OAI-PMH feeds; (c) customised web page templates; and (d) workflow modules to support data ingest and transformation.

**Fig. 1: RIF-CS – Linked Data**
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. 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 persistence layer that represents information using RDF and OWL; it is built on the Jena semantic web framework [13] [14].

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 (Attribute Box) contains descriptions of specific instances of data, e.g. John Smith is a Chief Investigator.

The architecture of the Hub has been designed to allow for automatic machine to machine communication for the ingest 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 disambiguating entities within data sources can be difficult and has not yet been captured 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 then determines 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 none exist, requests are made (machine to machine) to create an ID, i.e. 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 A-Box. This automatically creates HTML landing pages for human readable versions data. 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.

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 periodically harvests the new and updated institutional data via this interface.

4 RESOURCE DISCOVERY LAYER

Like most universities Griffith has an institutional repository which currently houses text-based publications as well as other enterprise systems which hold information about projects, publications, research data and staff (researchers). Until recently it did not have an integrated enterprise solution to deal with the data deluge, let alone a system which could collect the relationships among all these entities.

The implementation of the Metadata Exchange / Research Activity Hub has seen the meshing of research data sets driven primarily from research grant work (emerging concept), enterprise corporate data (new concept) with digital publications (traditional library concept). The latter include additional research-centric outputs such as the Australasian Digital Theses collection (http://www.griffith.edu.au/library/support-for-research/digital-thesis-submission) which publishes digital versions of higher degree theses via a searchable database. Theses are an important resource for Griffith Graduate Research School faculty, who serve as lead supervisors for graduate research student thesis publication. Another important resource is the institutional repository, Griffith Research Online (http://www.griffith.edu.au/researchonline/).

The following diagram is a simple illustration of how the Hub links to a number of key enterprise systems as well as the national Research Data Australia database (ARDC):

Fig. 3: Resource Discovery Layer

The principal non-library enterprise system that contains valuable information on research activities is the Research Administration Database, maintained by the University's Office for Research. The
database is the University's authoritative source for information on all project-based worked held at the University, including donations, research grants, and consultancy and commercial research policy. Only projects identified as in progress or completed are candidates for ingest to the Hub.

A crosswalk mapping between the Research Administration Database and the Hub was developed prior to the set-up of automated ingests to ensure the efficient transfer of data between the two systems. An extract, load and transform (ETL) process is initiated to control data workflow between the two systems. The Kepler workflow engine (https://kepler-project.org/) is used to automate the entire extract and ingest procedure. A typical Kepler workflow for ingesting data into the Hub is seen in the image below:

![Kepler Workflow Diagram]

Fig. 4: Kepler Workflow; File Reader = ETL process

While the objective of the above project was to populate Griffith research activity information for discovery on the Research Data Australia (RDA) service, the information and technology architecture developed has opened new opportunities for populating other discovery services.

Griffith has recently deployed the Serials Solutions' Summon web-based discovery service as the library search / discovery tool; Summon has been evaluated in depth in Library Technology Reports [15]. It is now possible to utilise the Metadata Exchange / Research Activity Hub to push key research information, e.g. data on Griffith research data collections, through to the Summon library search tool—known as LibrarySearch, making it another resource available for scholarly purposes. The first stage of ingests to the Griffith LibrarySearch includes research data collections published to the Hub. This is an important step toward substantially enhancing the support not only for researchers but also for postgraduate students.

5 INFORMATION SEEKING BEHAVIOUR OF POSTGRADUATE STUDENTS

Recent studies in the literature have reported on the searching behaviour of researchers; a number of these studies have included postgraduate students in their coverage. Researchers use—and prefer—“easy solutions that are adequate, not optimal” [16]. Researchers prefer generic services with easy interfaces as opposed to specialised databases with challenging search and retrieval capability.

In addition they utilise Google for many different types of information searches even if the search results may not be precise or complete but just “good enough”. Williams [17] refers to Google as the “ultimate enabler” and corroborates the fact that researchers seem unconcerned that it yields a partial and potentially unmediated set of results. This limitation seems to be offset—from a researcher's perspective—by Google's ease of use, its word-search capability, and the fact that searches often deliver “serendipitous contextual information in addition to what is expected”. The JISC UBiRD [18] study reinforces these findings. This should not be a surprise since—as Borgman [19] argues—in principle anything on the Internet can be considered part of the content layer of the scholarly information infrastructure. It should be irrelevant whether that content is on a web site or in a repository or in a digital repository.
Several recent studies which focus specifically on the information behaviour of postgraduate students corroborate these findings, with some slight differences. George [20] reports that their behaviour is greatly influenced by people, especially academic staff, who introduce them to key resources. In terms of library resources, there is a clear preference for convenience and / or the need to have information quickly. A barrier to their search for information is knowledge about or access to resources.

Vezzosi [21] undertook a regulatory study of doctoral students in the field of biology at the University of Parma. Her findings also indicate that people play a crucial role in their information behaviour, principally in relation to recommending relevant documents. Doctoral students in this discipline rely heavily on the internet for their research work. Google is a “crucial information tool” and appears to be the starting point for information seeking in most situations. Students appear aware of the importance of bibliographic databases and the possibility of doing advanced searches; however they appear uninterested in devoting time to reading the relevant instructions. Like George, Vezzosi highlights the importance of citation chaining as a searching technique, i.e. the use of a relevant article of book in order to track references, endnotes, footnotes and bibliographies.

Jamali’s study [22] reports on doctoral students in the field of physics and astronomy. There are differences in the information-seeking patterns of students who conduct theoretical research and those of whom are involved in experimental research. The findings however show a high reliance on electronic journals in both areas. In the study by Jankowska [22], the focus was on user satisfaction and expectations concerning library services at the University of Idaho. Findings suggest that some of the negative perceptions are a result of the fact that “most users do not have a full understanding of the complexities of library systems’ and a result of “graduate students’ lack of awareness of certain services and resources and / or their lack of proficiency in navigating the library’s information systems effectively”.

The overarching message from these studies is very clear. Postgraduate students—like researchers—want a Google-like approach to their information seeking needs: comprehensive and intuitive in ease of use.

6 IMPLICATIONS OF GRIFFITH PROJECT FOR POSTGRADUATE STUDENTS

Implementation of the Hub has delivered two immediate major benefits. By feeding metadata to discovery tools, e.g. Google, Summon, it enhances discoverability of discipline specific resources through easy-to-use interfaces. It enhances the ability of the user to refine their search as well as to explore other areas of enquiry.

Secondly it enhances connectivity. 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 goes beyond the common practice of citation chaining by providing direct access to a broader range of resources, e.g. datasets, which may not necessarily be referenced as part of a traditional scholarly publication.

Because valuable research information such as research projects (e.g. research grants), data collections, services and related publications (and their relationships) is already extracted from other University systems and collated in the Hub for uploading to Research Data Australia, there exists the potential to deliver a comprehensive set of research “activity” information. The Hub is capable of pulling data from the various data sources mentioned, regardless of the data format, and then combining and presenting it in a centralised web format.

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 outputs. For the postgraduate student, a Profile can be used as a tool to identify seminal research undertaken by experts within the group including their respective 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 Profiles are also a useful tool for prospective postgraduate students, when selecting their preferred university and their particular area of specialization.
7 CONCLUSION

Griffith University has developed a semantic web service--Metadata Exchange / Research Activity Hub--that not only integrates and aggregates research data within the institution but also provides a key link into national and international open systems. The system collects appropriate metadata from research collections within the University through customised feeds from the various University content management and corporate systems. This system then acts as a central University repository to feed information in a standard format to Research Data Australia, thereby supporting the creation of a national collaborative research infrastructure. From an institutional perspective, it exposes data to LibrarySearch, the University library discovery tool. In so doing, the Hub broadens the purpose of this type of discovery tool beyond the scope of traditional library managed resources by exposing other useful resources on campus.

Studies show that both researchers and postgraduate students require easy-to-use interfaces to specialised information sources. The Hub, using the inherent features of semantic web technologies, provides this capability. It is particularly useful for postgraduate students, who are just starting their research activities, especially the formulation of their research hypothesis. The ability to provide high quality, easy-to-use information support for the research needs of postgraduate students is important not only to enhance resource discovery but also to build their expertise as future researchers. This supports the national agenda of improving the eResearch capability of university postgraduates.

REFERENCES


