Building an Institutional Discovery Layer for Virtual Research Collections

Malcolm Wolski, Joanna Richardson and Robyn Rebollo, Division of Information Services, Griffith University, Australia

Abstract

University libraries are under pressure to ensure that their strategies and services to support researchers are aligned with the parent organization’s research goals. Important to researchers are not only research information needs—which necessarily underpin their research—but also the discoverability and accessibility of their own research outputs. While libraries have a history of designing discovery systems, new research paradigms are presenting both challenges and opportunities for libraries to reconceptualise such systems within a broader context. This paper describes a nationally funded Australian university initiative to build a research repository which feeds data into both a national research data service and university library discovery tools. Challenges and benefits are discussed.

Introduction

At a national level government funding and policy guidelines are placing pressure on universities to increase the accessibility of their research output. Research is a competitive field in which one of the keys to success is collaboration based on the ability to efficiently find and use quality data which is ready to be assimilated into a project—be it local, national or international—both in the immediate and in the long-term future. Currently there is a clear focus in the literature on the need to improve research data management, sharing, and accessibility to help meet these objectives. Concurrently attention is also being paid to services that aid in discovery.

Researchers’ Searching Behaviour

Recent library studies in the literature have focused on how best to provide information support for research, especially given the fact that the relationships between researchers and the “traditional” library support have shifted radically. Researchers use—and prefer—“easy solutions that are adequate, not optimal” (Kroll & Forsman, 2010). Researchers prefer generic services with easy interfaces 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 (2009) 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 study (Wong, Stelmaszewska, Barn, Bhimani, & Barn, 2009) reinforces these findings.
This should not be a surprise since—as Borgman (2007) 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 library.

**Libraries and Discovery Systems**

Much has been written in the literature about the role of the library in making the world’s knowledge accessible to current and future scholars (Brown & Swan, 2007; Frischer, Unsworth, Dempsey, & Staley, 2009; Wegner, 2006). A joint Research Libraries UK / Research Information Network (2011) study describes academic libraries as “a physical manifestation of the values of the academy and scholarship.” Libraries have traditionally seen their role as providing free access to the world’s scholarship (O’Brien, 2010). Comprehensive bibliographies outline the advantage of open access on citation impact (Hitchcock, 2010; Wagner, 2010).

Traditionally libraries have played an important role by exposing institutional research output through their institutional repository. The “digital repository”, with which libraries have had such a leading role, has been predominantly institution-based, although there are some notable discipline-based ones. The institutional repository has traditionally relied on the goodwill of its research community to upload open access versions of their publications. This was seen as a logical step given that the cycle of discovery has long been underpinned by sharing data, particularly scientific data, through publications. While there have been some notable successes, most institutions have been underwhelmed by the response from their research community (Lynch, 2003; Salo, 2008).

Until recently the scholarly output that libraries have focused on capturing has tended to be limited to traditionally published works. More recently new publishing paradigms are emerging, with data—supporting journal articles—as the focus. Repositories, nevertheless, do have an important role to play in supporting the research lifecycle as that support now moves to encompass more than just print research output. That role is not, though, a “singular technological solution” (Furlough, 2009). Instead repositories are becoming integrated components of larger systems and distributed infrastructure (Payette, 2010). Nicholas (2010) establishes a compelling case for not automatically assuming that the institutional repository should contain all content associated with all stages of the research lifecycle, particularly datasets. The core audience for this complex data tend to have requirements which are not currently supported by most institutional repositories; therefore they may store that data in a different repository federation, e.g. TARDIS for crystallography, which is better suited for the complexity of their needs.

How then can libraries be a major player in what Neuroth (2009) calls the “architecture of participation”? Lougee (2009) points to work done at the University of Minnesota Libraries in developing a multi-dimensional framework for academic support. This has resulted, for example, in the development of more customised library services that assist in the discovery and gathering phases of research.
However, as she also observes, “We have a record of designing discovery systems. Libraries also have experience related to tool development and integration of resources. Where libraries have been less involved is in the behavioral and community assessment that is a necessary prerequisite to developing customized environments. Further, the library’s role in catalysing collaboration is infrequent or unintentional.”

The authors describe in this paper a nationally funded Australian university initiative to build a research data—defined in the broadest sense—repository which feeds data into both a national research data service and university library discovery tools. The initiative has utilised key library skills in collaboration with other information professionals to build a service which is aligned with researchers’ needs to enhance both the discoverability and accessibility of their research.

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 (Sandland, 2009).

ANDS has built the Research Data Australia (RDA) service (Research Data Service, 2010). 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. Interestingly the Australian Research Council—a major funder of research at the national level—revised its funding rules in early 2011 for a major grants category; successful applicants may now use a small percentage of awarded funding to underwrite the costs of publishing / disseminating project outputs, i.e. publications and data.

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 (Buetow, 2009; Thelwall, Li, Barjak, & Robinson, 2008).

**Metadata Exchange Hub Architecture**

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 automated metadata feeds to Research Data Australia (RDA). These 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) (Australian National Data Service, 2010b), which is a subset of the ISO standard 2146 (International Standards Organisation, 2010). The development of a metadata aggregator (Hub) has become a core piece of infrastructure (Wolski, Young, Morris, De Vine, & Rebollo, 2010).

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 (Australian National Data Service, 2010a). 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 Hub is to expose the relationships 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.
The following diagram (Figure 2) is a simple illustration of the Metadata Exchange Hub components. VIVO, a semantic web solution 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 using the RIF-CS.

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 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 (Kepler Collaboration, 2010) for data workflow and transformation, OAI-CAT (OCLC, 2010) for OAI-PMH provision, and custom Java code for object Identifier creation.

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.
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 (http://vivo.cornell.edu/) 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 (Devare et al., 2007; Krafit et al., 2010).

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).

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 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 Hub metadata feeds for people profiles and research project records are ingested currently by national systems such as National Library of Australia’s People Australia and Research Data Australia, and eventually by the Australian Research Council. These systems then determine whether persistent identifiers exist for people, e.g. People Australia persistent identifiers, or projects, e.g. external funding identifiers. Research Data Australia automatically provides persistent identifiers for any collection records ingested from the Hub which do not include one. This is done through ANDS’ persistent identifier service called Identify My Data, based on Handle technology.

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

**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 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://adt.caul.edu.au/) 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):
The principal non-library enterprise system that contains valuable information on research activities is the Research Administrative Database, maintained by the University’s Office for Research. The database is the University’s authoritative source for information on all project-based work 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 Administrative 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 Figure 4:
It should be noted that while Kepler workflows were utilised in the initial development because of the requirement for the investigation and rapid deployment of a range of different methods for processing data, developers are currently analysing other options. In the future it is planned to replace the Kepler workflow engine with a more efficient, dedicated ETL server such as JEDOX (http://www.jedox.com/).

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. It is now possible to utilise the Metadata Exchange 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.

**Challenges**

The creation of an enterprise discovery layer in the form of the Hub has highlighted a range of challenges for University central IT, library and administration departments. Firstly authoritative source data needs to be identified within the institution, along with the associated data sources and data owners. Secondly agreement needs to be reached among the various data owners on access to authoritative source data in the various repositories and databases for publishing purposes. Thirdly the Hub and the various data feeds need to be treated as corporate systems for maintenance and support, while the Hub application itself needs a business owner within the organisation. The library is a logical owner because of its expertise in the underlying
information technologies and science required to develop this service. However there are many stakeholders within the organisation which have vested interests in the published information, e.g. research administration in its role as a major authoritative data provider, public relations and marketing using this information for advertising, especially to attract new students. Therefore, if the library is to assume the role of business owner, it will need to forge strong working relationships with these other organisational units.

It is expected that additional descriptive metadata will need to be created or captured to make published data more meaningful, e.g. a research project description suitable for publication. There are also numerous issues in relation to access management associated with ethics clearance, privacy, IP and other restrictions that need to be addressed. Levels of access are dictated by departmental, institutional, and national policies, regulations and legislation; even jurisdictional boundaries may be an issue. These need to be applied at various levels: from the published metadata through to individual research data objects. This can be a complex issue since it needs to be addressed at the research project level, unlike administrative data which can be addressed at the institutional level.

A final challenge is that many research projects cross institutional boundaries. Research funding agencies also retain and publish their own information in relation to University research projects. The challenge arises when data from these institutions is fed to federated discovery sites such as the RDA. There needs to be consistency between data published from research funding agencies and other institutions when identifying the projects and people involved.

In implementing the first stage of the Hub, research data collections have been ingested for which there is agreement on the issues raised above. However as the service expands, it is clear that this will become a catalyst for the University to improve its overall approach to research data management so that both the university and the researcher achieve short-term and long-term goals: manage their research outputs according to good practice; comply with policy guidelines and legislation no matter where they are stored; and provide assurance that access and discovery are managed.

Benefits

Initially the Metadata Exchange Hub was designed primarily to provide an automated metadata feed about Griffith University research data to Research Data Australia (RDA). However as library professionals began to collaborate with other information professionals, it became readily apparent that the project would transcend its original purpose. The role of library professionals in designing and implementing the project has been documented elsewhere (Wolski, Richardson, & Rebollo, 2011).

Implementation of the Hub has derived immediate benefits. The Hub aggregates metadata which is harvested externally by RDA and National Library of Australia’s Trove service. Locally it feeds into LibrarySearch, the library discovery service. In so doing, the Hub broadens the purpose of the library discovery tool beyond the scope of traditional library managed resources by exposing other useful resources on campus without the need to enter information into the traditional library catalogue. In addition
it contributes to Williams’ (2009) “serendipitous contextual information” retrieved in local searches. Researchers looking for information via LibrarySearch may find other relevant resources already on campus. The library discovery tool also facilitates connections between researchers by exposing research activity.

It lifts the profile of university research outputs as a useful resource not only within the academic and student community but also among library support staff. Importantly from the perspective of raising the library’s profile, it starts to address Oakleaf’s point about the need to demonstrate “library value within the context of overarching institutions” (2010).

**Conclusion**

The need to identify the multifaceted and overlapping relationships between digital objects and then make these relationships explicit is an important part of scholarship and archival work (Furlough, 2008). The implementation of Griffith’s open source Metadata Exchange Hub allows university libraries to enrich their discovery layer(s) by exposing not just institutional research outputs but also the richness of the relationships among those objects.

From the researcher’s perspective, on the one hand the Hub has the potential to present a reasonably complete profile of their research activity--and outputs--by aggregating data from a range of authoritative university resources. On the other hand, a researcher external to the university can retrieve more complete and comprehensive search results in highly specialised areas of research--results which go beyond the “just good enough” category mentioned at the beginning of this paper.

The service and architecture described here are a first step in creating a research discovery framework that not only exposes existing information in university systems to build a virtual institutional research collection but also integrates that information into a national discovery layer to build a large virtual national collection. University libraries, regardless of geographic location, are well placed to be a key participant / catalyst in helping their own institutions build these types of virtual research collections, thereby making research outputs more easily discoverable and accessible for collaboration and reuse.

**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.

**References**


O'Brien, L. (2010). *The changing scholarly information landscape: reinventing information services to increase research impact*. ELPUB2010 - Conference


