

Jun 23rd, 1:00 PM - 2:00 PM

Creating a university research data registry: enabling compliance, and raising the profile of research data at the University of Melbourne

Simon Porter

University of Melbourne, simon.porter@unimelb.edu.au

Anna Shadbolt

University of Melbourne, annams@unimelb.edu.au

Follow this and additional works at: <http://docs.lib.purdue.edu/iatul2010>

Simon Porter and Anna Shadbolt, "Creating a university research data registry: enabling compliance, and raising the profile of research data at the University of Melbourne" (June 23, 2010). *International Association of Scientific and Technological University Libraries, 31st Annual Conference*. Paper 3.

<http://docs.lib.purdue.edu/iatul2010/conf/day3/3>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

CREATING A UNIVERSITY RESEARCH DATA REGISTRY: ENABLING COMPLIANCE, AND RAISING THE PROFILE OF RESEARCH DATA AT THE UNIVERSITY OF MELBOURNE

Simon Porter

The University of Melbourne, Australia
simon.porter@unimelb.edu.au

Anna Shadbolt

The University of Melbourne, Australia
annams@unimelb.edu.au

Abstract

The University of Melbourne is one of the richest sources of research data in Australia making it a highly desirable contributor to Australia's emerging Research Data Commons – an initiative of the nationally funded Australian Research Data Service (ANDS). This paper will outline how The University of Melbourne partnered with (ANDS) to test a framework for exposing a number of research data collections from a variety of research communities at the university. It will identify how the project evolved with multiple agendas including;

- 1) The need to enable university research data and records policy compliance.
- 2) Participation in the national research data commons, and
- 3) Creating virtual research profiles for cross organizational research themes, as a way of strengthening cross disciplinary research.

Underpinning these agendas is an approach to populating the research data registry based on the reuse of already collected data on research. In this project we considered it critical that processes used for collecting information about research datasets leverage existing information that the University already collects about research such as grants and publications. Using this information, we tested how readily we were also able to detect the existence of research data sets, along with the probable associated researchers, project description, departments, and research classifications before individual researchers were directly engaged. Such an approach required command of research administrative datasets collected by the University's Research Office, but also the clever use of Library technologies to quickly source and scan publications for descriptions of research data. The result of these 'linked data' connections between research data sets and the rest of the research information framework was stored in an RDF triple store using an instance of the VITRO platform[3] created by the University of Cornell. The paper will also cover the choice of VITRO as an appropriate platform to base a research data registry.

Key words

Research data management; research data; research data registry; research information;

1. Introduction

In the age of the mashup, systems that collect and administer information now rarely serve a single purpose. The reasons for which data was first collected may be complimentary, or unrelated to the value that society places on its use in the future. This message is of course one of the key arguments for establishing practices that ensure access to research data. As the case is put in *Ensuring the Integrity, Accessibility and Stewardship of Research Data in The Digital Age*:

“Advances in knowledge depend on the open flow of information. Only if data and research results are shared can other researchers check the accuracy of the data, verify conclusions, and build on previous work. Further more, openness enables the results of research to be incorporated into socially beneficial goods and services and into public policies, improving the quality of life and the welfare of society.” [Committee on Science Engineering and Public Policy, 2009, 5.]

This paper reports on the progress towards the establishment of a university wide research data registry at the University of Melbourne that applies this philosophy not only research data sets that it registers, but also to the registry itself. In order to be successful, it is our belief that a research data registry must serve at least three separate purposes

- 1) The need to enable university research data and records policy compliance
- 2) Participation in the national research data commons, and
- 3) It must play a key role in helping the University communicate its research identity both internally and externally.

These goals are met within a broader pattern of research information reuse

2. Multiple Drivers for a Research Data Registry

2.1. Policy and compliance

In 1996 the University of Melbourne Research Records Project was initiated in response to requests from academic departments for assistance in the development of procedures for the storage of research data as required by the Melbourne University Code of conduct [Grady, McRostie & Papadopoulos, 1997]. This project initiated the development of model guidelines for the management of research data which were endorsed in 1998. Despite the processes followed, in time it became clear that the ‘guidelines’ label suggested that maybe they were not viewed as a requirement for the management of research data and records. Consequently, a review of the 1998 guidelines was conducted in 2004 and through strong advocacy from the Research Office; the Academic Board of the University endorsed the full guidelines as University policy in 2005.

In February 2009, Professor Peter Rathjen, the Deputy Vice Chancellor (Research) (DVC-R), requested an investigation of compliance with the research data and records management requirements of the 2007 Australian Code for the Responsible Conduct of Research [NHMRC, ARC & Universities Australia, 2007]. These national requirements have been captured in the University of Melbourne Code of Conduct for Research in University Regulation 17.1.R8 – Code of Conduct for Research and the University Policy on the Management of Research Data and Records [University of Melbourne, 2005], approved by Academic Board in 2005. The University of Melbourne policy makes particular reference to the responsibilities of researchers and Heads of Department. With the establishment of the Australian National Data Service (ANDS) in the second half of 2008, there has been renewed interest in how universities are handling research data: our investigation has been timely. Faculties were surveyed and a soft audit was conducted in a few departments to assess the level of compliance to the policy. We noted that compliance was generally off the radar for most

researchers interviewed. Departmental compliance with some aspects of the policy was patchy and in many cases data storage was ad hoc often requiring short term strategies at the project level to build resources to secure data beyond the life of a project with local risk management mitigation unclear. In particular, it was noted that local registers were absent so there was actually limited departmental awareness of what research data actually existed within their programs. Similar concerns were reported in Shadbolt et al (2006) and again in 2007. The outcomes of these investigations were reported to University Council and provided the DVC-R with a strong argument to acquire resources to build central infrastructure to support research data management and storage. The case was put that the missing key to compliance was well-resourced central services. To this end Information Technology Services (ITS), the University Library and the Research Office are working together to consolidate and expand central support for research data storage and management. These services include:

- ✓ **Information:**
 - Establishing a web based access point to resources, tools, standards, support (work in progress).
 - Awareness and Training – information forums will be designed to meet community needs and interests.
 - Building a community of practice to facilitate increased awareness of innovations and strategies occurring in research groups across the university.
- ✓ **Advice:**
 - Establishing a triage service that will record requirements and direct researchers and research administrators to key resources, locally, regionally, and nationally.
- ✓ **Building Capability:**
 - Expanding the librarian role to work with researchers to facilitate metadata capture and research data management.
- ✓ **Central data storage:**
 - Building a multi-petabyte capacity storage facility over the next 3 years.
- ✓ **A Central research data registry**
 - The provision of a Central Research Data and Records Registry begins to address one compliance requirement while at the same time providing an excellent opportunity for the systematic collection of quality information about research activities complementing data already collected via other research reporting mechanisms. To this end, the Research Portfolio, through the Director of eResearch, the Library, Information Technology Services (ITS) and VeRSI (Victorian eResearch Strategic Initiative), are developing a centralised research data management service which will include a University-wide research data register and digital data storage. This registry will also provide an ongoing communication layer between the University of Melbourne and the Australian Research Data Commons.

2.2. The University's Public Research Profile

In 2006, the University of Melbourne recognised that the information on research that it had laboriously collected for internal administration and government reporting purposes could also be used to effectively communicate public statements about the Universities research profile. In combination with supporting policy, public elements of a researcher's identity were identified from centrally collected information (see Figure 1.) Based on this information, a web page for each researcher in the University was created, and updated nightly. Each page lists a researcher's publications, grants, qualifications, research expertise; international linkages, associated research classifications, and contact details. This set of web pages called Find an Expert (www.findanexpert.unimelb.edu.au) was launched in 2007, and last year recorded an average of just under 1700 visitors per day.

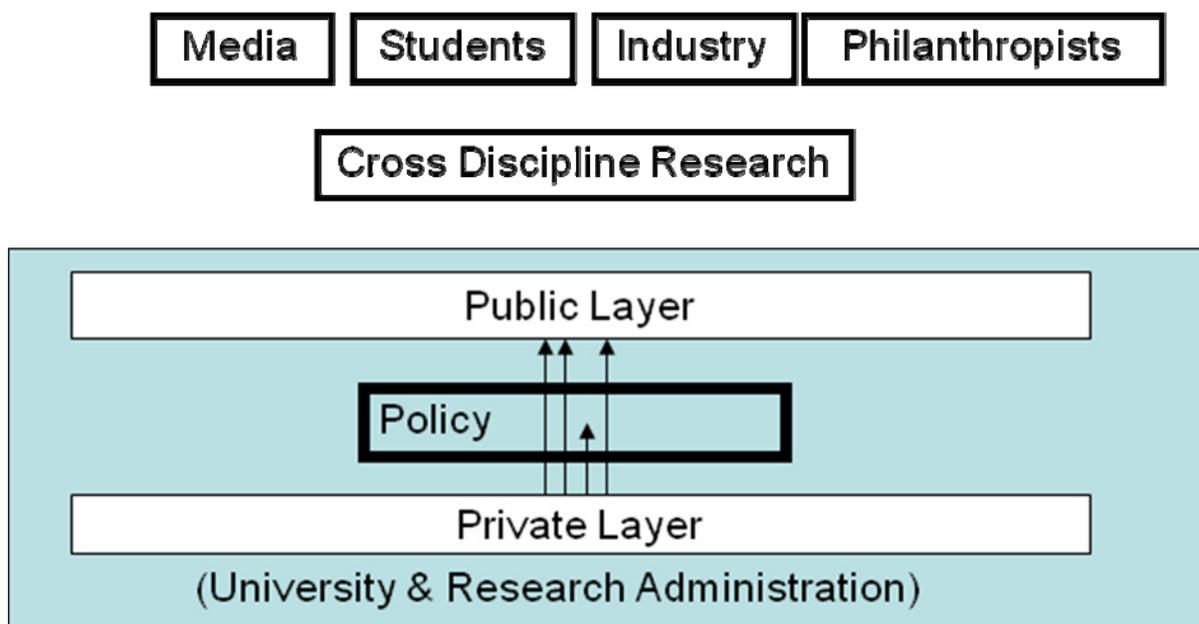


Figure 1: Policy driven communication of research profile data

Adding information about the research data collections that are associated with University of Melbourne researchers to the existing profiles, provides an extra dimension; one particularly useful in the case where knowledge of the existence of a research data set could lead to a new collaborative research opportunity for university researchers.

2.3. Research Data Australia

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

One high profile ANDS activity is to establish the population of Research Data Australia, a set of web pages describing data collections produced by or relevant to Australian researchers. It is designed to promote visibility of research data collections in search discovery engines such as Google and Yahoo, to encourage their re-use.

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

As the University of Melbourne is one of the richest sources of research data in Australia, it is a highly desirable contributor to Australia's emerging Research Data Commons, and with ANDS has undertaken significant projects in both Seeding the Commons and Metadata Capture.

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. It became apparent that a central university research data registry (established to facilitate compliance with University Policy) was also a good fit with these objectives providing a single University harvest point to Research Data Australia.. A central research data registry is also likely to be sustainably maintained by the University well after ANDS funding has dissipated.

Under the metadata stores program ANDS has provided the University of Melbourne with funding to extend the functionality of its research data registry to allow it to be harvested by ANDS, and to improve the connectivity with University research administration systems.

3. A broader vision for a research data registry

The following directions are intended to be pursued within the next year to extend the capabilities of the registry.

3.1. Integration into a National Information Infrastructure

Research is increasingly collaborative making it imperative that in addition to local Institutional based sources of information on research a data registry would need to have methods in place to uniquely identify researchers from other institutions. This additional information might be sourced from national registries including the Research Data Australia [ANDS, 2009], the People Australia Project [National Library of Australia, 2008] as well as appropriate international registries.

3.2. Integration with well maintained research collections

An institutional research data registry will also require the ability to harvest locally maintained registry collections; supporting research programs with established processes for registering their own research data assets.

An Institutional research data registry may also exchange data with discipline based repositories. An Australian example of such a repository is the Australian Social Science Data Archive (<http://assda.anu.edu.au/>). In some cases this will involve acting as a staging area for publishing research data to the repository. In other cases, a domain based repository may aid the discovery of research data sets created by researchers that belong to the institution.

3.3. Integration with research storage services

In order to effectively manage the transition between active research project storage, it is crucial that the metadata collected around the provision of research data storage is collected with the requirements of the research data registry in mind.

4. The Melbourne Research Data Registry pilot

4.1. Defining the parameters

In order to satisfy University of Melbourne Policy on the Management of Research Data and Records, the ability to describe at a minimum, the following information elements was identified as being essential to provide in a research data registry:

1. Project Title – official name of the Project – with additional or alternative titles where appropriate;

2. Description of the project providing brief details of the project
3. Custodian – the lead Department responsible for the research activity and output;
4. If more than one Department involved, include other departments;
5. Name of all institutions collaborating in the project;
6. Country(ies) of collaborating partner, if located outside Australia
7. Principal Investigator;
8. Names of other researchers, including student researchers and externals;
9. Supervisor (where applicable);
10. Funding body/bodies (if applicable);
11. Grants/Contract identifiers;
12. Research Classification/Codes
 - a. RFCD code
 - b. FOR code
13. Date the research commenced
14. Date completed or thesis submitted;
15. Date of publication of related publication or public release (if applicable);
16. If available link to publication;
17. Retention Period – how long does this data need to be kept including disposal requirement and delegation;
18. Description of data and format(s) - Include sufficient detail here to ensure that the full set of materials can be identified and retrieved. Format of data may include computer printouts, laboratory notebooks, files, maps, electronic data files, photographs, video and audio recordings, charts, models, disks, magnetic tapes.

Description of data/records and format	Quantity	Location	Date Stored	Restricted / Confidential*
				<input type="checkbox"/>

19. Access Rights and Restrictions
 - a. Identify any ethics, confidentiality, and/or IP restrictions that may limit access to the data
 - b. If they exist, identify any contracts that may be associated with these restrictions on the data /project. Include Themis code if appropriate
20. If relevant, Ethics Approval number(s);
21. Identification of contracts associated with the data/project if appropriate
22. The location of Data Management Plan if available
23. Identification of the digital data stored in the central storage facility (if any)
24. The location of non-digital data associated with the project (if any)
25. The location of digital data associated with the project not currently stored in the central storage facility (if any).
26. Details of relocation of data outside department – if applicable. Original data and records may be relocated to another department within the University, but may not be removed from the University.
27. Key words to identify this dataset

Beyond this common set, it is recognised that research data collections from different disciplines will have evolving metadata profiles to best enable discovery and reuse.

4.2. Populating the research data registry

In recognition of the large amount of information required to contextualize a research data set, in populating the registry, our goal was to minimise the amount of information that researchers have to provide to build the information infrastructure of the register. Within the Australian context, a research data registry operates in an environment where a large amount of information is already known (and laboriously collected) via the need to satisfy government reporting requirements. The challenge for a research data registry is to integrate seamlessly with this rich source of institutional

information on research and researchers including grants, publications, researcher descriptions, and government reporting classifications. As department structures, employed researchers, and even classification schemes change over time, this information will also need to be managed in an information context that is both historical and current.

The initial set up of the registry we essentially used the public set of information already defined by the 'Find an Expert' initiative. From this set of information, a decision was made to try and derive the existence of research data sets before even approaching researchers. Beginning with the Faculty of Architecture, and the Melbourne Sustainable Society Institute, a student was hired to scan the text of a publication looking for evidence of an associated research data set such as a reference to survey data. Publications were sourced using SFX open url resolver links created from already collected publication metadata within our research information system.

Once a potential data set was identified, a method of data entry was piloted whereby a minimum set of information was entered. Using the link between the research data set and the publication, a further set of data could be derived based on a set of assumptions, including:

- ✓ The authors of the publications are likely (but not always,) to be the creator of the data set
- ✓ The owning department associated with the publication for government reporting purposes is likely (but not always,) to be the owning department of the research data set
- ✓ The research classifications and keywords associated with the publication are likely to be the same or similar to the research data set.

This approach was used to identify just over 500 potential datasets and each were recorded in Central Research Data and Records Registry.

To allow research data sets to be found against a larger amount of background information, green boxes ■ beside a the name of a researcher, department, or publication to indicate the number of associated data sets which is also illustrated in Figure 2.

The screenshot shows the Central Research Data Registry interface. At the top, there is a navigation bar with the University of Melbourne logo, a search bar, and links to 'University Homepage' and 'Search the University'. Below the header, the main content area is divided into several sections. On the left, there is a sidebar menu with options like 'Home', 'Data sets', 'Potential Data Sets', 'Department', 'Institute', 'Research Group', 'People', 'Data sets for the Data Commons', 'Register My Data', and 'Index'. The main content area features a search bar and a list of research data sets. The first section is titled 'Architecture, Building and Planning Department' and lists several research data sets, each with a green box indicating the number of associated data sets. The second section is titled 'Research Data Custodian' and lists more research data sets. The third section is titled 'employs' and lists several researchers, each with a green box indicating the number of associated data sets.

Figure 2: Screen shot from the Melbourne Research Data Registry

4.3. Engagement with researchers

In order to confirm that potential data sets were in fact real, and request additional information such as the location of the research data, meetings with individual researchers were scheduled, with the support of the associate Dean of Research.

Initial feedback from researchers has been mixed with some happy to enter the extra data themselves and others preferring to utilise the system for new projects rather back tracking with legacy data. The general feeling was that it needed to be easy to complete and not a duplication of other reporting requirements. A broader roll out has been suspended while the user interface is structured to better accommodate self-deposit.

5. Overall technology approach

To implement its Research Data Registry in its pilot stage, Cornell University's Vitro software platform was selected. Vitro is a general-purpose web-based ontology and instance editor with customizable public browsing, first developed for a research and scholarship portal at Cornell University (www.vivo.cornell.edu/). The initial focus of vitro development is to communicate and manage descriptions of research activity, at the broad level of people, projects, and publications. Via a related project called DataStaR, the platform is also being used to manage the more specific processes around supporting collaboration and data sharing among researchers during the research process, and to promote publishing or archiving data and high-quality metadata to discipline-specific data centres, and/or to Cornell's own digital repository (<http://datastar.mannlib.cornell.edu/about>). The combined focus of these projects was a natural fit with both the registry and public profile drivers identified above.

In 2008, The University of Melbourne demonstrated the applicability of the Vitro Platform to manage expressions of research activity, and to communicate this information up to the Online Research Collections Registry Australia ORCA (a precursor to Research Data Australia).

With respect to ANDS, creating a research data registry on top of the Vitro platform offers several advantages for Australian Research Institutions

- Vitro is a Semantic Web application designed to work in an open information context. It has been designed to both harvest and author information in equal measure
- Vitro is driven by user defined Ontologies, allowing implementation level flexibility in the way that information about research, and research data is expressed.
- Vitro's initial design as a web based discovery portal for Cornell Research, making the public promotion and discovery of research data sets a process decision rather than a separate application.
- All data in Vitro is stored in an rdf triple store, along with the definition of the Ontology that it relates to. The clear separation of data, data model (ontology), and presentation and editing application, makes the long term preservation of the information stored in Vitro, or subsequent migration to another registry tool some time in the future easy to manage.
- Strategic Direction. Through its Vitro and DataStaR initiatives, Cornell University via the Albert Mann Library has demonstrated a commitment to the development of solutions that support the improved handling of research metadata in ways that are strongly in line with the ANDS program. By working with the Vitro software, the University stands to benefit from Cornell's own development agenda. This is particularly true given the recent awarding of the NIH project: "VIVO: Enabling National Networking of Scientists" (U24 RR029822).(www.vivoweb.org)

To increase the chance of other Australian universities and creating a local support network, the ontology associated with the registry was chosen from existing external ontologies, rather than creating a unique University of Melbourne model. Initially, the choice was made to implement the

swrc ontology however the registry will soon be migrated to be inline with the Vivoweb ontology.

5.1. Building an Australian user community for Vitro

Since implementing the first version of our Registry in 2009, two additional Australian universities: Griffith University and the Queensland University of Technology have opted to build a metadata hub on Vitro for their Joint ANDS funded metadata harvest project. With three Australian Universities now working together to extend Vitro functionality to meet ANDS objectives and more showing some interest, it is hoped that a larger Australian Vitro user community will develop.

Acknowledgements

The authors acknowledge the project funding from the Australian National Data Services in 2009/2010 without which this work would not have been possible.

References

Australian National Data Service (2009) Research Data Australia, from:
<http://services.ands.org.au/home/orca/rda/>

Committee on Science, Engineering, and Public Policy (2009) *Ensuring the integrity, accessibility, and stewardship of research data in the digital age / Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age, Committee on Science, Engineering, and Public Policy.* Washington, D.C.: National Academies Press, c2009.

Grady, K., McRostie, D., & Papadopoulos, S. (1997), "Hunters and Gatherers: From Research Practice to Records Practice." *Archives and Manuscripts*. The Journal of the Australian Society of Archivists, Volume 25, Number 2, November 1997, pp242-265.

National Library of Australia (2008) People Australia project, from:
<http://www.nla.gov.au/initiatives/peopleaustralia/>

NHMRC, ARC and Universities Australia (2007) Australian Code for the Responsible Conduct of Research. Code is accessible at: <http://www.nhmrc.gov.au/publications/synopses/r39syn.htm>

Shadbolt, A., van der Knijff, D., Young, E. & Winton, L. (2006) "Sustainable paths for data-intensive research communities at the University of Melbourne: a report for the Australian Partnership for Sustainable Repositories", Technical Report, Information Services, The University of Melbourne. Accessible at: <http://repository.unimelb.edu.au/10187/1870>

University of Melbourne (2005), 'Policy on the Management of Research Data and Records'. University Policy accessible at: <http://www.unimelb.edu.au/records/research.html>