Discovering Australia's Research Data

Stefanie Kethers
Australian National Data Service
c/o Monash University
Clayton 3800, Australia
(61) 3 9902 0546
Stefanie.Kethers@ands.org.au

Xiaobin Shen
Australian National Data Service
c/o Monash University
Clayton 3800, Australia
(61) 3 9902 0567
Xiaobin.Shen@ands.org.au

Andrew E. Treloar
Australian National Data Service
c/o Monash University
Clayton 3800, Australia
(61) 3 9902 0572
Andrew.Treloar@ands.org.au

Ross G. Wilkinson
Australian National Data Service
c/o Monash University
Clayton 3800, Australia
(61) 3 9902 0598
Ross.Wilkinson@ands.org.au

ABSTRACT
Access to data crucial to research is often slow and difficult. When research problems cross disciplinary boundaries, problems are exacerbated. This paper argues that it is important to make it easier to find and access data that might be found in an institution, in a disciplinary data store, in a government department, or held privately. We explore how to meet ad hoc needs that cannot easily be supported by a disciplinary ontology, and argue that web pages that describe data collections with rich links and rich text are valuable. We describe the approach followed by the Australian National Data Service (ANDS) in making such pages available. Finally, we discuss how we plan to evaluate this approach.

Categories and Subject Descriptors
H.3.5 [**Online Information Services**]: Data Sharing; H3.7 [**Digital Libraries**]: Collection, User Issues, Standards

General Terms
Documentation, Design, Human Factors

Keywords
e-research, metadata, Australian Research Data Commons

1. INTRODUCTION
The Australian National Data Service (ANDS) (http://www.ands.org.au) is funded by the Australian Government's National Collaborative Research Infrastructure Strategy and the Super Science Initiative. Its main goal is "more researchers re-using data more often" [8]. This includes addressing technical and social / cultural issues, as found e.g. the StORe project [15] and other initiatives [4] Specifically, ANDS partners with researchers and research organizations to help them meet their data management ambitions; and transforms the disparate collections of shareable research data around Australia into a commons of discoverable research resources.

JCDL'10, June 21–25, 2010, Gold Coast, Queensland, Australia.

To ensure that the right level of information is available for researchers to reuse data, ANDS has decided to base its services at the collection, rather than data items, level. Although the definition of collection varies across different standards (e.g. ISO 2146:2010 [12], NISO draft Z39.91-200x [14]), the ANDS preferred definition is that of a collection as "an aggregation of physical or digital objects treated as a unit for business purposes" [12]. From the perspective of a scientist looking for reusable data, it is important that collections are described in a way that meets their needs, which vary at different stages in the process of reusing data – i.e., discover, access, and exploit [8].

After discussing related work in section 2, we examine sample research questions that have led researchers to look for research data collections across disciplines and locations and the resulting information needs in section 3. Section 4 describes ANDS' approach to making data more discoverable in such situations, the Australian Research Data Commons (ARDC). Section 5 outlines further work and concludes the paper.

2. RELATED WORK
Australia is not alone in adopting a national approach to data management, but with the recent Super Science funding for ANDS, Australia is now spending more on research data management in absolute terms than any other country.

The Netherlands does have a strong offering in the form of the Data Archiving and Network Services (DANS, http://www.dans.knaw.nl/), based within the Dutch Royal Academy of Arts and Sciences. DANS provides data archiving to the social sciences and some humanities. There is no equivalent to DANS at present for the sciences in the Netherlands.

In the UK, a number of discipline-specific data centres provide support for data archiving for researchers from those disciplines. The Digital Curation Centre (http://www.dcc.ac.uk/) has moved from an initial focus on preservation to a far greater interest in data issues over the last two years, and is now a significant national provider of expertise in the sector.

In the US, the National Science Foundation has funded two projects under the Sustainable Digital Data Preservation and Access Network (DataNet) call. This aims to build sustainable infrastructure by creating a new type of organization that the NSF does not believe exists today. They are looking for librarians, archivists, and computer/ computational/information scientists who will together to build excellent infrastructure for science

and/or engineering, while engaging deeply with intended users; domain scientists will be full partners in the process.

Services similar to the ones described in this paper include a registry service for scientific data, which allocates citable DOIs to data collections and also to individual data sets, if required, provided by the German Technische Informationsbibliothek (TIB) Hannover [6]. The Information Environment Service Registry (IESR) [3] in the UK provides a registry of collections of resources including data collections.

3. DISCOVERING RESEARCH DATA

3.1 Sample Research Questions

Many large research questions need access to data that is created by researchers, by government and by community, and does not comfortably sit within any particular disciplinary boundary [17]. Indeed, some of the most important research questions require teams straddling many fields, and data that are not held by any team member. For example,

- To research birth weights in the Australian working class, McCalman et al [13] needed access to data held in the National Archives of Australia, the Archives Office of Tasmania , family histories, and medical records from the Army. How does one find and correlate this data?
- To research the resilience of Australian coral reefs, Hoegh-Guldberg and colleagues [10] needed to capture their own data, but also to refer to data from diverse sources, such as interaction between coastal ecosystems and humans [11].
- To determine effective water use in South Eastern Queensland, there is a need for hydrological data, economic data, and land use data to be examined and correlated to enable Bristow et al. [7] to investigate appropriate models.

To address questions like these, ANDS is creating the Australian Research Data Commons (ARDC) [1] to provide a location to discover data collections that are relevant to researchers. In doing so, ANDS made a significant decision to hold information about the collections, but not the collections themselves, creating a rich set of relationships to support discovery of this research data. Consequently, there is support for rich information exploration about the collections, but little support for explicit data queries, which can be better supported in the research domains, and when explicit data needs have been described. The Australian Research Data Commons is intended to enable researchers to exploit Australia's research data as a whole, rather than supporting specific individual disciplines.

3.2 Abstract Needs

There are many information needs that a national data service needs to support – e.g. direct access to data, data manipulation, or location of a known item. Some of these needs will be met by other forms of data services, e.g. discipline-specific data services, such as the Protein Data Bank (PDB, http://www.rcsb.org/pdb/). However, there are four characteristics of the examples given above where the problem space is complex:

- Data is likely to be held by many organizations and in many formats,
- The need for the information by researchers will be ad hoc (that is, not able to be anticipated by service providers),
- The information need is more complex than a simple discipline-focused query, and may be best expressed by a network of concepts, rather than a set of query words, and
- The information need is beyond a particular discipline boundary.

Consequently, no particular access method, or particular ontology is likely to be able to meet this kind of need.

3.3 Information Needs

Metadata can be classified into a number of types, according to the attributes of the object the metadata encodes: *descriptive metadata*, *technical metadata*, *structural metadata*, *preservation metadata*, *provenance metadata,* and *rights metadata* [16]. Based on the role that the metadata can perform in assisting with discovery and re-use, it can also be characterized as:

- **Discovery metadata:** to help a user find the data
- **Decision metadata:** to determine data value for the information seeker
- **Access metadata:** to help gain access to the data
- **Reuse metadata:** to support data reuse and exploitation.

These roles and the metadata needed to support them draw on the experience of the UK Data Archive and their established practices. The different metadata roles can be filled by different metadata elements, although some metadata elements can fulfill several roles. For example, the useful discovery metadata may include (but not be limited to) the following items:

- **Title**: includes e.g. data set name (for a data set), organizational name / personal name (for a person or organisation), research grant / project title (for an activity, such as a project).
- **Description**: Provides context for the data, e.g. a description from a relevant ontology, an abstract of a related paper, or excerpts of a relevant proposal
- **Field of research subjects**: includes ANZSRC code subject and/or local research key words
- **Coverage**: includes temporal and spatial coverage of the data
- **Contact information**: includes the distributor, or contact information for the custodian of the data. (This could serve as discovery metadata, decision metadata or access metadata if the data is not online).

The decision metadata can include the relationships between different collections, researchers, organisations, projects and services, as well as data quality information. For instance, a researcher may find the fact that a particular researcher or funding program was associated with the data to be an indicator of higher quality. Providing links to this related information can assist in the determination of value.

The access metadata can include a global persistent identifier (e.g. handles or DOIs), and a local persistent identifier (e.g. local ID/local database primary key) for the data. It may also include the contact details of an individual / organization, if the data providers want access requests to be mediated through an individual, rather than met automatically via a data store. Such a facility can ameliorate some researchers' concerns about misuse of the data they make available for sharing.

The re-use metadata can include a link to a paper describing e.g. study design or sample preparation, or to a document that explains the variable names or encodings used.

4. THE AUSTRALIAN RESEARCH DATA COMMONS (ARDC)

We argue that, in order to fulfill the information needs described in section 3, an interlinked set of Web pages providing the different types of metadata described above are useful for researchers to locate, assess, access, and reuse relevant data sets.

The Australian Research Data Commons aims to encourage more researchers to reuse their data more often. The term "commons" refers to a framework joining data from different sources and making them available for community use [1]. The Australian Research Data Commons is a combination of four components:

- 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 (e.g. data, researchers who produced data sets, data collection instruments, and institutions where the researchers work),
- the infrastructure needed to enable, populate and support the commons [1].

ANDS does not hold the actual data, but points to the location where the data can be accessed, e.g. in an institutional repository. Currently, ANDS provides one service for the discovery of research data, i.e. Research Data Australia (RDA, described in the next section), and three major services to researchers wishing to make their data available through the ARDC:

Identify My Data (http://ands.org.au/services/identify-my-data.html) allows researchers or system designers to create a persistent identifier, i.e. a clickable reference, for their dataset that can be maintained so that it will not be broken when the location of the dataset changes. ANDS is a partner in the DataCite initiative [5], which will provide DOIs for data sets.

Register My Data (http://ands.org.au/services/register-my-data.html) allows system designers to register collections of research materials. Descriptions of registered collections are published in a number of discovery services (e.g. Research Data Australia and the Global Registries Initiative, http://www.globalregistries.org/).

Publish My Data (http://ands.org.au/services/publish-my-data.html, a composition of Identify My Data and Register My Data) allows Australian researchers and research organizations to publicise the existence of research collections via the internet [8].

4.1 Research Data Australia

One way to access the contents of the Australian Research Data Commons is through one of its key services, Research Data Australia (RDA, http://services.ands.org.au/home/orca/rda/). RDA is aimed at describing the data collections (with rich descriptive information) produced by Australian researchers and making them accessible via browsing and via a Google search restricted to the RDA pages. Furthermore, major search engines (e.g. Google, Bing and Yahoo) can index the RDA pages. Software written by ANDS takes collection descriptions and associated information and builds a rich mesh of web pages for harvesting by search engines.

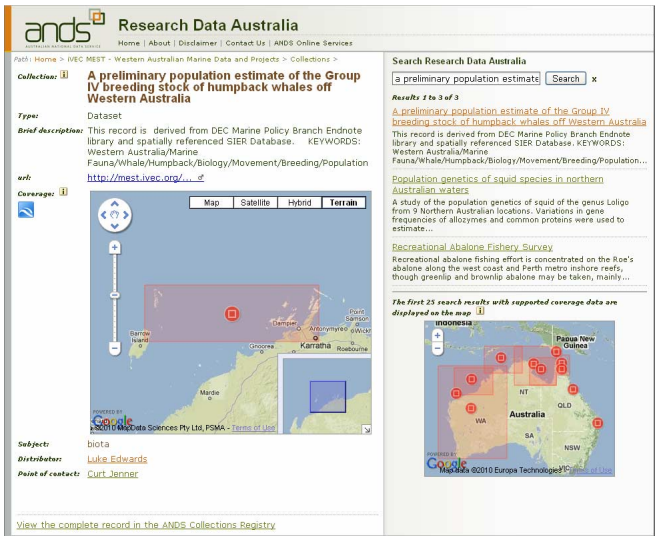

Figure 1: Sample Page from Research Data Australia

Figure 1 shows a sample page from Research Data Australia with mainly description metadata (e.g. title, description, coverage, field of research subjects). At the bottom of the page, there is also information on relationships that this data set has with other RDA elements, in this case the distributor of the data set.

4.2 Building Research Data Australia

In order to build Research Data Australia, ANDS needs to maintain a database of information. ANDS has chosen to build this on ISO 2146:2010 [12]. This is an international standard developed to operate as a framework for building registry services for libraries and related organizations. The ANDS Registry Interchange Format – Collections and Services (RIF-CS, see http://ands.org.au/resource/rif-cs.html) is defined based on ISO 2146, but it only includes elements needed for a collection service registry, so it is not a full binding to the ISO 2146 standard. Furthermore, RIF-CS is designed to be a data interchange format supporting the submission of metadata to a collection service registry. Thus, it is different from the traditional Australian institutional repository metadata storage formats, e.g. Dublin Core, MODS or MARCXML. ANDS does not recommend RIF-CS as a metadata storage format – it is designed primarily for interchange of information about ISO2146 objects.

Specifically, the current version of RIF-CS is composed of four registry object elements based on ISO 2146:2010 entities [12]:

- **Collection:** an aggregation of physical or digital objects,
- **Party:** a person or group,
- **Activity:** refers to something occurring over time that generates one or more outputs, and
- **Service:** refers to a physical or electronic interface that provides its users with benefits (e.g. work done by a party or access to a collection or activity).

These object elements can have varied semantic relationships with each other, e.g. a person can be *partOf* a group, a collection *isOwnedBy* a party, or an activity *hasOutput* collection.

The biggest advantage of RIF-CS is its flexibility in transferring information and its richness of relationships between the collection, party, activity and service concepts. The biggest disadvantage of RIF-CS is the additional work required from data providers to provide feeds of these ISO 2146 entities. ANDS is working to develop software and guidelines to assist with this.

5. CONCLUSION AND FURTHER WORK

5.1 Evaluation

Laboratory tests of our approach are not possible, as the Australian Research Data Commons is an operational service under continual development and enhancement. However, ANDS will monitor use as our collections descriptions grow in their richness and their interconnectivity. This will take two forms – quantitative and qualitative. In our quantitative measures, we will determine over time the number of collection that are described, the frequency with which collections pages are accessed, and the paths that are followed through the ARDC.

The last measure should show how valuable the connections are, and the other measures should show whether descriptions of this form are seen as valuable to researchers for describing their own data and discovering other researchers' collections. The other evaluation will be qualitative; ANDS will interview researchers yearly to determine the value of the service against a variety of their data needs, in particular the value of the service for needs that are complex, ad hoc, and beyond disciplinary boundaries.

5.2 Concluding Remarks

In this paper, we gave examples of research questions that give rise to information needs that go beyond specific data services currently provided by individual disciplines. We have described the Australian Research Data Commons (ARDC), a framework of services designed to help address such researchers' information needs that are broader, and more complex than typical data queries within a discipline a framework of services designed. The ARDC is currently in its infancy, but already contains over 800 collections and more than 200 parties. We expect a strong growth of the ARDC over the next 18 months, and are beginning to collect stories and anecdotal evidence about usage of the ARDC.

6. ACKNOWLEDGMENTS

Our thanks go to the Department of Innovation, Industry, Science and Research (DIISR) and the National Collaborative Research Infrastructure Strategy (NCRIS) for financial support. We also wish to thank the reviewers for their insightful comments.

7. REFERENCES

[1] ANDS 2009. Australian Research Data Commons. http://www.ands.org.au/news/ardcfinalprojectplan-oct2009.pdf (retrieved 14 April 2010).

[2] ANDS 2007. Towards the Australian Data Commons. https://www.pfc.org.au/pub/Main/Data/TowardstheAustralianDataCommons.pdf (retrieved 8 April 2010).

[3] Apps, A. 2005. The JISC Information Environment Service Registry. ASSIGNation 22(3) pp 9-11.

[4] Beagrie, N., Beagrie, R., and Rowlands, I. 2009. Research Data Preservation and Access: The Views of Researchers. Ariadne Issue 60, http://www.ariadne.ac.uk/issue60/beagrie-et-al/ (retrieved 12 April 2010).

[5] Brase, J.2009. DataCite - a global registration agency for research data. In Eleventh Interlending and Document Supply Conference.

[6] Brase, J., Schindler, U. 2006. The publication of scientific data by World Data Centers and the National Library of Science and Technology in Germany. Data Science Journal 5:205-208.

[7] Bristow, K.L., and Charlesworth, P.B. 2002. Effective water management vital to the long-term economic viability of the Lower Burdekin. In Proceedings of the ANCID Conference, 1 to 4 September 2002, Griffith, Australia.

[8] Burton, A., and Treloar, A. 2009. Designing for Discovery and Re-Use: the 'ANDS Data Sharing Verbs' Approach to Service Decomposition. International Journal of Digital Curation, 2009. 4(3).

[9] Burton, A. and Treloar, A. 2009. Publish My Data:A composition of services from ANDS and ARCS. In Proceedings of Fifth International Conference on e-Science, p. 164-170.

[10] Diaz-Pulido G.A., McCook, L.J., Dove S., Berkelmans R., Roff G., Kline D.I., Weeks S., Evans R.D., Williamson D.H., and Hoegh-Guldberg O. 2009. Doom and Boom on a Resilient Reef: Climate Change, Algal Overgrowth and Coral Recovery. PLoS ONE, 4(4): e5239.

[11] Hoegh-Guldberg, O. 2005. Low coral cover in a high-CO2 world Journal of Geophysical Research Oceans 110(C9).

[12] ISO TC46 SC4. 2005. ISO 2146 - Information and Documentation - Registry Services for Libraries and Related Organisations.Working Draft, 13 December 2005. http://www.nla.gov.au/wgroups/ISO2146/n197.doc (retrieved 8 April 2010).

[13] McCalman, J., Morley, R., and Mishra, G. 2008. A health transition: Birth weights, households and survival in an Australian working-class population sample born 1857-1900. Social Science & Medicine 66(5):1070-1083.

[14] NISO. 2005. NISO Z39.91-200x. http://cidoc.ics.forth.gr/docs/frbr_oo/frbr_docs/Z39-91-DSFTU.pdf (retrieved 14 April, 2010).

[15] Pryor, G. 2007. Attitudes and Aspirations in a Diverse World: The Project StORe Perspective on Scientific Repositories. International Journal of Digital Curation 2(1).

[16] Treloar, A., and Wilkinson, R. 2008. Rethinking Metadata Creation and Management in a Data-Driven Research World. In Proceedings of IEEE e-Science, p. 782-789.

[17] Working Group on Data for Science. 2006. From Data to Wisdom: Pathways to Successful Data Management for Australian Science. Report to Prime Minister's Science, Engineering and Innovation Council (PMSEIC).