The International Journal of Digital Curation
Issue 2, Volume 3 | 2008

Access to Data for eResearch: Designing the Australian National Data Service Discovery Services

Andrew Treloar, Ross Wilkinson

Australian National Data Service

November 2008

Summary

Much work on data repositories has derived from effort on document repositories. It is our contention that people do not access research data for the same reasons that they access research publications. We argue that it is valuable to understand information needs, both immediate and contextual, in establishing both what information should be collected, what metadata are captured, and what discovery services should be established. We report on the information needs that we have collected in our efforts in establishing the Australian National Data Service. These needs cover much more than data – there are needs for information about the data, their creators, a need for overviews, and further requirements to do with proof, collaboration, and innovation. We provide an analysis of those needs, and a set of conclusions that has led to some implementation decisions for ANDS.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

The rise of data-driven research (Hey & Trefethen, 2003) has meant increasingly that publication of the research as a largely textual document is not enough. The reader (and possibly even the reviewer) needs access to the data (and maybe the computational models) used to derive the conclusions being argued for. The scale of research data are increasing, both with respect to volume and heterogeneity (Association of Research Libraries, 2006; National Science Board, 2005). It is our contention that people do not access research data for the same reasons that they access research publications. Consequently the approach that has been taken to data repositories will not necessarily mirror that of the various publication repositories that have been set up over the last decade (Bailey, 2008; Lynch, 2003). We will describe our approach to understanding access needs in the context of the Australian National Data Service, so that we could understand a set of access methods, which finally leads to an information model and associated derived repository requirements.

The Australian National Data Service

The Australian National Data Service (ANDS)[1] is being established by the Australian Commonwealth Government under the National Collaborative Research Infrastructure Strategy. ANDS commenced in September 2008 and will run until June 2011. Among other aims, ANDS will work to ensure that Australian research data are well managed, made available for access, and discoverable so that:

- Researchers can find and access any relevant data in the Australian 'data commons'.
- Australian researchers are able to discover, exchange, reuse and combine data from other researchers and other domains within their own research in new ways.
- Australia is able to share data easily and seamlessly to support international and nationally distributed multidisciplinary research teams.

ANDS thus has a role in ensuring that all relevant data are made available, are discoverable and curated for the long term. The data themselves are expected to be placed in stores and repositories – both institutional, such as exists at Monash University[2], and also by discipline, such as the Atlas of Living Australia (ALA)[3]. Furthermore, there is a need to find paths to the data from collection discovery services, again exemplified by ALA, but also via a proposed national collections description.

An Information Needs Approach

Drawing on the traditions of library science and information science, Peter Ingwersen and Kalervo Järvelin have recently (2007) provided a new perspective on the importance of knowing why information is being sought, not just what the information is. This context should be explicitly modelled and used in evaluation of effectiveness. This is just as true for data discovery as document retrieval. For example, if a researcher is formulating an experiment, and needs to find out all relevant data sets to the experiment, then it is important to obtain wide coverage using

[1] The Australian National Data Service (ANDS) http://ands.org.au/

[2] Monash University http://arrow.monash.edu.au/

[3] Atlas of Living Australia (ALA) http://www.ala.org.au/

any system to discover the data. On the other hand, if researchers are in the middle of an experiment, and need to access research data from a colleague, then they need to get exactly that data set. Naïvely, exactly the same query might be issued of some discovery service, despite the widely differing information needs. Less naïvely exactly the same query might be issued to different discovery services, and each need would be appropriately satisfied.

This need for context in order to deliver data that are needed, has very often lead for calls for comprehensive metadata that support not only the discovery but the management and use of the relevant research data. However there are potentially very significant costs in creation of the contextual infrastructure and then the capture of the relevant metadata to fully describe the context. With a fixed budget, how does one decide what metadata to capture? Information systems researchers have extensively investigated a variety of frameworks that enable a cost/benefits approach to deciding on features of an information system. A good paper on this approach by Delone and McLean (1992) describes the many costs and benefits that need to be considered. However the overwhelming costs are usually human costs – those of data creators, those of data managers and curators, those of data discoverers, and (often importantly for the research context) those of data transformers. Comprehensive metadata capture loads effort on the creators, managers, and curators, whilst minimal metadata load the costs on data discoverers. How does one strike the right balance?

We believe that a good way of addressing these issues is to concentrate on information needs in the first instance, and then information, before turning to the technologies and data that might be needed to support these needs. In that context, we elicited views from Australian researchers and data managers on the information needs, the information, the users and finally the sorts of technologies are needed.

We first explore the needs. Clearly the dominant need is simply access to known data, held in a known location. The reasons for this access may vary however – there is a need for creators to access the data to check their work, to re-analyse using different tools and parameters. Other researchers access data to apply their techniques to existing data, and to agglomerate the data. Importantly there is a need for independent researchers to access data to test the claims of a research outcome.

However there were many other needs that were identified:

- Finding resources within a discipline using subject-specific descriptors
- Finding resources across disciplines using general descriptors
- Rapid and easy access to new data as they become available
- A need for consistency - a pathway to consistent data in consistent format - ideally the data would be in well described raw formats, (for remote sensing researchers, say) and consistent secondary formats (for end users such as ecologists and climatologists, say)
- A need for consistency - a pathway to consistent data in consistent format - ideally the data would be in well described raw formats, (for remote sensing researchers, say) and consistent secondary formats (for end users such as ecologists and climatologists, say)
- A data review that would be used as part of the start of any research program, just as a literature search is carried out.
- A perspective beyond the domain of inquiry - users might be issuing

queries within a domain, but are provided with references to collections that are beyond the domain but potentially relevant using cross-walks provided by a collections registry – the needs are met within the domain but exploiting a discovery service that walks across very high-level ontologies to find other potential collections

- Basic access to data using tagged text for all researchers
- Access to data that support collaboration via complementary data
- Researchers needing information that enables them to connect with other researchers and their data to expand their capability
- An overview of collections as a whole
- An overview of the collections together with the research creators, institutional custodians, and data services available
- Finding out about the topic as a whole – where is research being done, who is doing it and what are they doing it on? - e.g. research bodies looking for novelty or otherwise in grant applications
- Novel information perspectives to support innovation that comes from accessing information created for one purpose being used by another person for a different purpose (therefore avoiding barriers, such as domain language and access methods, is important)
- Access to data sets outside the discipline with no “expert” intermediation
- Access to data and collections that support both a national/international perspective
- A view of collections as a whole that is intended to raise general research awareness across research areas
- A view of research outside any particular discipline to support cross disciplinary awareness through data awareness

It is clear that there is a wide set of needs – they might be categorized as data needs in a variety of forms; there is also a need for information about the data, the context within which those data were created, as well as various forms of overview. It is very interesting to note that sitting under these explicit needs, there are also requirements – proof, efficiency, collaboration and innovation that are all very strong drivers of value for an e-Science. Understanding this value is perhaps key to making a decision about just what data and information are collected. Finally, note that it is very unlikely that any single access method would meet all of these needs.

What Information Should be Captured?

From the above, it is clear that the job is not done when research data are moved from the desktop to the repository! (This might still be a significant step forward.) From the descriptions of need, we can see that the following items should be easily findable: data, documents, people, research projects, institutions, relationships, and collections of any of the above.

As well as these items, there is a need to keep descriptions of these items – either structured through metadata schema, or perhaps simply as a document that describes the item. It is also the case that each of the above items can be first-class – worth finding independently so as to satisfy an information need. It is worthy of note that this approach is well supported by ISO 2146 (2008), an international standard currently under development by ISO TC46 SC4 WG7 to operate as a framework for building registry services for libraries and related organisations. In addition to these ’first-class objects’, it is clear that it would be valuable to create information artefacts on the fly that provide overviews, summaries, and structures to help provide a wider understanding of the context of a data set, or other item.

What metadata will we therefore need to capture? For the purposes of this discussion we will ignore metadata necessary for such purposes as access control, management, preservation, etc. and instead concentrate on the metadata that help identify items in a discovery process. Some possibilities include:

- Unstructured textual descriptions - during the interviews, it was suggested that sometimes the best description of the data set is provided in the research proposal, and this should be made available
- Broad metadata that ensured controlled access beyond disciplinary boundaries
- Generic, but specific metadata, such as geo-coding that enables accurate access, or time-based coding, such as might be used to describe a relationship between an institution and a project, or the validity of the data
- Specific metadata, often through highly specific ontologies, that exploit strong shared descriptions of a field or approach, such as SNOMED

Since there is much metadata that could be captured, how do we choose? Understanding value from an information needs perspective can help to provide a basis for making a decision. The metadata necessary to support direct access may need to be augmented by more discursive metadata that help the development of finding aids, overviews, trails, etc. However it might be prohibitive to generate such metadata separately, whereas this information might be derived from research plans, research data plans, or research outputs, such as publications. Finally, some captured information might not be directly related to a particular data set – information concerning the research group or project captured using web site descriptions may be appropriate unless it can be gleaned form other sources. Curation decisions are thus made in the context of the competing needs of e-Science.

ANDS Response to Information Needs/Capture

In Treloar and Wilkinson (in press) we argue that just what metadata are captured depends crucially on the cost of capture, and the information needs that are being supported. Nevertheless in the context of ANDS, our understanding of needs suggests that getting good coverage is important in order to support the requirements that all relevant data be described at some level. More detailed metadata are acquired as the business cases dictate. We thus do not dictate what metadata should be captured, but will simply harvest what object metadata are available for data and documents. We will also work with repository managers and other trusted sources to ensure that we can also harvest information about the other first-class objects of interest that we have identified: primarily collections, but also people, research projects, institutions, and their various relationships. Our aim is to satisfy the coverage information needs, as they are not well serviced by domain collection or web search services; so it is important not to be prescriptive and to glean information from many sources.

What Access Methods Are Needed?

Discovery services are important – in the long term there should be support for easy access to data and other information in reliable and consistent forms. We need a variety of repositories holding different data types with access methods that suit:

- Document repositories with free text search augmented by metadata refinement
- Data collections organized by their metadata and discovered through relational queries
- Spatial data collections principally accessed through spatial queries
- Collections with other specialist interfaces, such as visual fly-throughs, or image collages
- Combinations of the above

Many of these collections will reflect their institutional or disciplinary requirements. However it is also valuable to discover collections – the form of this discovery may take many shapes. Some are:

- Publish web pages from a collections registry to enable web search using Google etc.
- A discovery service that sits directly on top of a collections registry supporting both free text search and metadata search.
- A search interface that enables domain or problem-specific portals such as ALA to augment their searches with searches to the collections registry that provide pointers to collections outside the domain that may be of interest
- Browsing support with pathways through the collections using such tools as finding aids and other methods.
- Domain and institutional search services should be referred to directly from a portal rather than attempting to provide a national service

These discovery services should be offered as part of a portal that provides other services on top of the collection database. The form of discovery is likely to reflect the nature of the need. Finding a specific data set in an unknown location may well be best found using a search service such as Google that has previously spidered pages that have been published from a collections registry with a relevant link directly to the data.

ANDS Response to Information Needs

We have indicated earlier that ANDS is working toward ensuring that institutional and domain-based repositories will be in place. The responsibility for managing these data rests with the holder. Data that form a logical collection will be organized accordingly, or described in metadata in such a way as to allow the members of a collection to easily be identified. Associated with each repository will be the native discovery services that are provided.

These repository collection descriptions will then be harvested into a national collections registry, based on ISO 2146. The collection descriptions (and the associated ISO 2146 first-class entitities) will be output as a set of webpages that can be spidered by web search engines. The metadata in the collections registry will also be harvestable by and searchable through an ANDS national discovery service, as well as discipline discovery services.

The range of information needs described above that require coverage, such as data surveys and overviews, are typically not well supported by standard web search engines, which are optimized for finding single items. Queries concerning all data about an area can be covered with spatial queries; queries concerning all data on a topic, typically need metadata support to ensure confidence that all data has been discovered.

We see a need to provide what we call "see also" services – where a researcher raises a query of a collection that is institutional or disciplinary, and is provided the answer, but gets additional references to collections, researchers, institutions, etc. that may be relevant to that need. To date we have seen this sort of service provided with faceted searches – but it is actually more like the sort of services provided by commercial services such as Amazon. We thus target early provision of discovery services to find a collection, and then a data set; find all relevant collections to a need; and find collections and items of interest beyond the boundaries of a given collection, as they appear to best meet the information needs discussed previously at a reasonable cost. Other services that may well add significant value such as generation of overviews, support for explicit pathways, etc. are the subject of business analysis at this stage.

Conclusions

In this article we have argued that data collections are different to publication repositories and that it is valuable to understand information needs in establishing both what information should be collected and what discovery services should be established. We have reported on the information needs that we have collected in our efforts in establishing the Australian National Data Service. These needs cover much more than data – there are needs for information about the data, their creators, needs for overviews, and deeper needs to do with proof, collaboration, and innovation. We have provided an analysis of those needs, and a set of conclusions that has led to some implementation decisions for ANDS. We introduced this work by noting that it is important not only to establish information needs but to evaluate the outcomes of basing discovery services on these needs and their context. As we establish the ANDS services, we are also establishing evaluation of these services against the needs so as to track our progress.

References

Association of Research Libraries. (2006). *To stand the test of time: Long-term stewardship of digital data sets in science and engineering*, final report, Association of Research Libraries. Retrieved July 30, 2008, from http://www.arl.org/pp/access/nsfworkshop.shtml

Bailey, C. (2008). *Scholarly electronic publishing bibliography,* Version 72: 6/9/2008. Retrieved July 30, 2008, from http://www.digital-scholarship.org/sepb/sepb.html

DeLone, W. H., & McLean E. R. (1992). Information systems success: The quest for the dependent variable. *Information Systems Research*, *Volume 3, Issue 1*, 60-96.

Hey, T., & Trefethen, A. (2003). The data deluge: An e-science perspective in F. Berman, G. Fox and T. Hey (Eds.), *Grid computing: Making the global infrastructure a reality*. (Chichester: Wiley, 2003), pp. 809-24. Retrieved July 30, 2008, from http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/research/esci/datadeluge.pdf

Ingwersen, P., & Järvelin, K. (2007). *The turn: Integration* (The Information Retrieval Series). Springer-Verlag New York, Inc.

ISO 2146. (2008). *Information and documentation - Registry services for libraries and related organisations* (ISO TC46 SC4 Working Draft, 13 December 2005). Retrieved July 30, 2008, from http://www.nla.gov.au/wgroups/ISO2146/n197.doc

Lynch, C.A. (2003) *Institutional repositories: Essential infrastructure for scholarship in the digital age*, ARL Bimonthly Report, Vol. 226, 1-7. Retrieved July 30, 2008, from http://www.arl.org/resources/pubs/br/br226/

National Science Board. (2005). *Long-lived digital data collections: Enabling research and education in the 21st century,* National Science Foundation. Retrieved July 30, 2008, from http://www.nsf.gov/pubs/2005/nsb0540/start.jsp

Treloar, A., & Wilkinson, R. (In press). Rethinking metadata creation and management in a data-driven research world. *Proceedings of 4th IEEE International Conference on e-Science*, Indianopolis, December 2008.