

Building Australia's eResearch Capability: the challenge of data management

Adrian Burton
Australian National Data Service
adrian.burton@ands.org.au

Margaret Henty
Australian National Data Service
margaret.henty@ands.org.au

Building Australia's eResearch Capability: the challenge of data management

The creation of the Australian National Data Service (ANDS) provides an opportunity to devise a national approach to the provision of skills for improving both institutional and individual eResearch capability. One of the four proposed ANDS programs is called Building Capabilities. This will have three broad areas of activity: including curriculum development, stakeholder engagement and the development and implementation of an audit and certification framework. This paper discusses the first of these; its target groups and constituency, skills areas and levels, and delivery strategy.

1. INTRODUCTION

The importance of data in the scholarly communications cycle is becoming increasingly apparent. More data is being created, usually now in digital form, and our capacity to process and transform it has grown. Developments in information and communications technologies (ICT) make it possible to do different kinds of research. Bigger and more sophisticated computers and communications systems support measurement, analysis, modeling, simulation, collaboration and publishing. Data is valuable, not just because of the cost of collecting it, but because of its critical importance in providing solutions to many of the world's greatest problems such as global warming, pollution and alternative energy sources.

Universities and research institutions around the world are looking to improve the infrastructure that supports intensive data use in the eResearch environment. Obviously it is critical that this infrastructure includes information and communications technologies of high quality, such things as supercomputing facilities and network connectivity. But equally important are the capabilities of both institutions and individuals to provide the support necessary to exploit these facilities to the full and to ensure expert data curation for future access and use.

Recognition of the need to improve access to data, its management and use has led to the creation of the Australian National Data Service. The ANDS vision is to "transform the disparate collections of research data around Australia into a cohesive corpus of research resources. This transformation would assist the connection of Australian and international data centres, repositories and online collections to enable serendipitous discovery, cross-disciplinary research, and cross-repository workflows." (ANDS Technical Working Group, 2007) ANDS commenced in 2009 after extensive preparations in 2008. There are four programs of activity designed to provide data management support Australian researchers and to start building the Australian Research Data Commons. The programs will deal with issues of research data ownership, roles and responsibilities, access, co-ordination and curation. The four programs are Developing Frameworks, Providing Utilities, Seeding the Commons and Building Capabilities.

2. BUILDING CAPABILITIES

The Building Capabilities Program is designed to “improve the level of capability for research data creation and management as well as research access to data (and associated technologies) across Australia” (ANDS, 2008, p21). This covers three broad areas of activity: the development and delivery of a data management curriculum; engagement with key stakeholders through partnerships, a forum, reference groups, focus groups and so on; and the development and implementation of an audit and certification framework for staff, services and data facilities. Curriculum development and community engagement have priority commencing in 2009.

The Building Capabilities Program will assist research organizations, which includes universities and government-funded bodies, to develop their skills and capabilities in handling data. From the perspective of the organisations themselves, this is part of a broader framework to assess and improve their capabilities and capacity to support advanced ICT-enabled research. The activity of this ANDS program, therefore, sits within a broader capability maturity model for the Australian Research Data Commons. This model provides the context, framework, vocabulary and approach for the development and delivery of a data management curriculum.

The approach to the task is modified by some background assumptions. Resources will not permit ANDS, and nor does it have a brief, to deliver training for all research groups and data facilities around Australia. ANDS will, therefore, partner with organisations to coordinate the development of a formal set of training materials that any of these organisations can then make use of. ANDS will, however, do some “Train the Trainer” activities and will partner with strategic communities and organisations for the delivery of capability building programs. The formal training curriculum will be complemented by some ad hoc workshops, briefing sessions, symposia or other gatherings to provide information about matters of current importance or new developments. Overall we feel that a structured and nationally coordinated set of curriculum materials with a corresponding certification framework will make an enduring difference to Australia’s e-research infrastructure.

While ANDS is being funded by the Australian Government for the benefit of Australian researchers, we recognize that many, if not all, of the issues we face are being addressed internationally. ANDS will therefore take advantage of opportunities to partner with organizations overseas, either to exchange materials or to provide international access via online delivery if this proves feasible. This applies both to the development and delivery of curriculum, such as the Cornell University Digital Preservation Management Workshop now offered by ICPSR at the University of Michigan, and to the need for appropriate audit and certification frameworks, such as DRAMBORA (DCC and DPE, 2007) or the RLG/NARA Trustworthy Repositories Audit (RLG, 2007).

3. OUTCOMES

ANDS will provide a structured set of modules for delivery over the next two and a half years. These will target particular groups within our constituency, cater for different levels of expertise, target particular areas of skills and be designed to complement discipline- or institution-specific materials. ANDS will work with partners who have been associated with the development of the materials, who are committed to their maintenance and who are willing to take responsibility for their

delivery within their own institutions. The modules will be organized into a framework which will allow people to combine relevant modules into a certifiable program. At the time of writing, the precise nature of such a certificate has still to be determined.

4. THE ANDS CONSTITUENCY

The organizations which make up the ANDS constituency include all Australian universities, publicly-funded research organizations such as CSIRO, GeoScience Australia and the Australian Bureau of Statistics, and the cultural collections sector which includes galleries, libraries, archives and museums. The first two of these groups will be given priority in the first instance. The decision to give priority to universities and publicly funded research institutions is based on resource availability rather than on any question of the value of the organizations or the data they hold.

Within the constituency organizations, there are different groups which are seen as targets for training: research staff; IT staff such as data modelers and informatics specialists and sometimes referred to as data scientists; data facility and repository managers; librarians and archivists; data facility and repository IT developers and support staff; and those who are concerned with the policy framework and the strategic directions of data storage and management facilities. We recognize that different individuals and organizations are at different stages of capability, so it will be important to provide curriculum materials appropriate to those different levels. We also recognize that all groups need some skills in common but that each has specific requirements over and above those.

Still to be discussed is the issue of formal courses to be made available as part of undergraduate and postgraduate training. There are obvious advantages to training people at that early stage, especially those who are seeking a career in research. It is likely that training to be offered through university partners will include postgraduate students as part of the target audience, but it is unlikely that it would be regarded as part of their formal degree requirements.

5. SKILLS AND SKILLS LEVELS

The skills areas to be targeted have been identified through surveys and interviews with a wide variety of people associated with the conduct of data-intensive e-research and the stewardship of data (see for example, Henty, 2008).

The skills areas to be targeted include:

1. Legal & Regulatory issues such as ownership, copyright, licensing, meeting legal obligations, ethics and privacy, funding agreement obligations, and the public sector policies and guidelines which cover the conduct of responsible data-intensive research.
2. The Scholarly Communications Cycle, especially as it relates to data usage and including how to cite data, managing complex and compound publications and open access publishing.
3. Information Management, covering data storage and retention, discovery and access, persistent identification, curation and preservation, management platforms, and collection development/disposal schedules. The curriculum here will relate to some of the ANDS Utilities Services currently under development,

notably the persistent identifier service and a search service which will enable researchers to locate data sets in the Australian Data Commons.

4. Data Management Planning for research projects, both prospective and retrospective. There are suggestions that inclusion of a data management plan will be required by Australian research funding bodies, not to mention pressure from within organizations to undertake proper planning, so there is considerable interest in this topic. The need for a retrospective plan comes from researchers who now find themselves at the end of a project needing to create a plan retrospectively, or will be leaving the institution, because they are retiring, or have completed their higher degree training, or are moving on elsewhere. An important part of the retrospective plan in many cases is digitization; to convert data from analogue to digital form for preservation and sharing purposes.
5. Data Management Policy: this is a necessary corollary to requirements that research projects should have data management plans. The research institution therefore needs its own policies to support the implementation of these plans.
6. Informatics: this is a broad term to describe a number of skills of critical importance to the research team. It covers such topics as data modeling, data structures, metadata, vocabularies, taxonomies and ontologies.
7. Data Analysis: this is of most interest to the researcher and research team members and includes topics such as statistics, data mining, visualization and workflows. The extent to which this comes under the ANDS brief is still under discussion, as there is a fine line between data management and computation.

This list is neither exhaustive nor exclusive and will be refined on the basis of the availability of existing materials and expertise in partner organisations.

Several areas have already been identified as high priority. A recent study of researchers at three Australian universities (Henty et al, 2008, p17) showed the following preferences for training when asked "Would you be interested in training or advice on any of the following?"

Creating a research data management plan at the beginning of a project	52.0%
A data 'exit' plan (for retiring academics or departing academics and postgraduate students)	32.9%
Digitisation advice, tools and services	30.6%
Data 'rescue' for older digital materials, such as data on older media or migration of data from legacy systems	22.5%
Creating a research data management plan after a project has finished	22.4%

The need for training in data management planning is clear from this survey and ANDS will be able to build on work already done to provide a training module. During 2008, the Graduate Information Literacy Program of the Australian National University was commissioned by the Australian Partnership for Sustainable Repositories (APSR) to write a Data Management Planning Manual, on the proviso that this be made available for the use of others. (ANU GILP, 2008) This manual is

available to ANDS because of the relationship between APSR and ANDS. Staff of the APSR central office have been absorbed into ANDS, providing continuity of expertise and access to APSR resources. APSR has been engaged in training-related activities for some years, providing a substantial foundation for future ANDS activity. Further expertise among ANDS staff will come from the recent ARROW, DART and ARCHER projects.

Two topics given priority, persistent identifiers and collection description, relate to ANDS Utility Services currently under development. Other modules under development cover research data policy development for institutions, metadata and the scholarly communications lifecycle.

ANDS is establishing a forum to discuss and advise on the development and delivery of its curriculum. This will be made up of about twenty members from universities, research organizations and high performance computing facilities around Australia. The group will have face to face meetings as well as meeting via the access grid and using a web-based application for information sharing. The group will be consulted to locate existing resources which might be repackaged (with permission) for ANDS delivery, to suggest topics for curriculum development, to evaluate training materials, to suggest partners and potential contractors who may wish to be involved with curriculum development and to act as test beds for course delivery.

6. DELIVERY STRATEGY

There are many logistical issues to be worked through in the development of the proposed curriculum and some solutions and innovations should be apparent by the time of the conference. These issues include coping with different existing skills levels and finding suitable delivery mechanisms, given that ANDS funding does not extend to resourcing the delivery of capability building programs to all ANDS constituency groups. Training trainers will be a high priority.

7. CONCLUSION

The creation of ANDS provides an opportunity to address the issue of skills training for eResearch. The challenges are many. The target audience is wide and includes diverse perspectives on data creation, use and curation. Potential topics for curriculum development are many. The resources to be devoted to the task are of necessity limited, so managing expectations may be an issue. However there is a high level of good will in this new enterprise, with many showing readiness to take part in defining and refining a curriculum and developing and delivering materials. Work already carried out overseas provides excellent models and potential partnerships.

Perhaps the last word should go to the conclusion of the Australian E-Research Coordinating Committee (DEST, 2006). They identified the importance of skilled personnel in the eResearch environment in their final report (p3). This identified the outcomes of implementation of their recommended framework that relate to the benefit of improving the capability of those engaged in eResearch. These would include:

Australian researchers will have the necessary education, training and skills, and support from ICT and information management specialists, to use advanced ICT efficiently and effectively;
The implementation of e-Research capabilities across the Australian research sector will provide a leading influence on the uptake and enhancement of such technologies by Australian business and industry; and
The Australian community and economy will benefit from the advanced capability enabled by e-Research.

REFERENCES

ANDS 2008. *Australian National Data Services (ANDS) Interim Business Plan, 2008/2009*. Retrieved February 10, 2009, from <http://www.ands.org.au/andsinterimbusinessplan-final.pdf>.

ANDS TECHNICAL WORKING GROUP 2007. *Towards the Australian Data Commons: a proposal for an Australian National Data Service*. Canberra, Australian Government Department of Education, Employment, Science and Training, Retrieved February 10, 2009, from <http://www.pfc.org.au/pub/Main/Data/TowardstheAustralianDataCommons.pdf>.

Australian National University Graduate Information Literacy Program (ANU GILP) 2008. *ANU Data Management Manual*. Canberra, Australian National University. Retrieved February 10, 2009, from http://ilp.anu.edu.au/dm/ANU_DM_Manual_v1.03.pdf

DCC and DPE. (2007). *Digital Repository Audit Method Based on Risk Assessment*, Digital Curation Centre (DCC) and Digital Preservation Europe (DPE). Retrieved February 10, 2009, from <http://www.repositoryaudit.eu/>

Department of Education, Science and Training (DEST). (2006). *An Australian e-Research Strategy and Implementation Framework: Final Report of the e-Research Coordinating Committee*. Retrieved February 10, 2009, from http://www.dest.gov.au/sectors/research_sector/publications_resources/profiles/e_research_strat_imp_framework.htm#authors

Henty, M. (2008). "Developing the Capability and Skills to Support eResearch." *Ariadne* (55). Retrieved February 10, 2009, from <http://www.ariadne.ac.uk/issue55/henty/>

Henty, M., B. Weaver, S. Bradbury and S. Porter. (2008). *Investigating Data Management Practices in Australian Universities*. Canberra, Australian Partnership for Sustainable Repositories. Retrieved February 10, 2009, from http://www.apsr.edu.au/investigating_data_management

RLG and NARA. (2007). *Trustworthy Repositories Audit & Certification: Criteria and Checklist, v1.0*. Retrieved November 26, 2008, from <http://www.crl.edu/PDF/trac.pdf#page=7>

ANDS is supported by the Australian Government through the National Collaborative Research Infrastructure Strategy (NCRIS).