DREAMING OF DATA: THE LIBRARY'S ROLE IN SUPPORTING E-RESEARCH AND DATA MANAGEMENT

Margaret Henty

Australian Partnership for Sustainable Repositories
Australian National University
Canberra, ACT, 0200
margaret.henty@apsr.edu.au

ABSTRACT

The increasing focus on eResearch comes with an increasing focus on data management, data use, data sharing and data re-use. This presents a challenge to research institutions as they decide where the responsibility for all of this data lies. Libraries have long had a role in managing text-based materials in the digital environment, as can be seen by the rapid take up of institutional repositories and many library ventures into electronic publishing of different kinds. There are many who see that the library has a significant role to play here, and several Australian university libraries are already actively engaged in data management in its broadest sense, as are state libraries, the National Library and others.

This paper examines the capabilities and skills required within research institutions as they implement the organisational and cultural change which will be needed. It is based on two surveys conducted in 2007. The Skills for eResearch Project undertaken by APSR included interviews and surveys of a range of people associated with the eResearch agenda. The findings suggest that the provision of appropriately skilled personnel will be key factor in moving forward, and that the skills required go far beyond the technical. At the same time, there are currently many barriers to the conduct of eResearch and these will need to be addressed in order to develop new services supporting institutional capability. The Data Management Survey was conducted by three Australian universities in late 2007 and looked at the data management practices of researchers. The results show both positives and negatives, and point towards areas for improvement.

The eResearch environment

The growing capacity of information and communications technologies to contribute to research of all kinds has excited researchers the world over as they invent new ways of conducting research and enjoy the benefits of bigger and more sophisticated computers and communications systems to support measurement, analysis, modelling, simulation, collaboration and publishing. In the process they are collecting and creating data, masses of it, and increasingly more as the level of technological sophistication grows. The solutions to many of the world's greatest problems lie in our capacity to work across disciplines, to collect and analyse data, and to create models and simulations, as can be seen in areas such as global warming, the need for alternative power sources and maintaining political stability. As a consequence, governments are keen to support research, and at the same time keen to ensure that the products of research, which is to say data, are well managed, readily findable and, preferably, available on open access. If data sets can be re-used, they say, then they should be. If the government is paying for the research, then the public should be entitled to have access to all the products of that research (while recognising ethical and privacy issues are also important).

The issue of institutional capacity and individual capabilities required to support eResearch has received attention internationally. Technical developments and the take up of eResearch have been rapid, but have not always been matched by corresponding infrastructure support developments at the institutional or discipline levels. Similarly, the level of capability among those with support responsibilities has not been able to keep up. Two key questions in this discussion concern responsibility: who are the key players in ensuring that data is managed responsibly and sustained for later discovery and use and how are their roles to be defined?

What do we mean by data management?

Good data management is fundamental to an institution's capacity to provide the infrastructure to support eResearch. Data management means different things to different people, usually according to the part they play in managing some part of the data life cycle. Researchers are primarily engaged in data creation and analysis, but the decisions they take in deciding what formats to use to collect and store their data, what metadata they will use to describe it, who owns it, who has access, what software they will use to analyse it, what outputs there will be from the research, and countless other activities will have an impact further along the track. Data management for the person who then takes on responsibility for data stewardship will involve another set of activities, including organisation, preservation and the provision of access. Both data creators and stewards exist in a world where other components of data management come into play: systems architectures, policies and procedures which have to be specified and known to everyone engaged in the process.

Data stewardship, the longer term organisation, access provision and preservation, is a term which is used a lot in this paper. Sometimes this is referred to in the literature as data curation. I prefer not to use the term curation, as it suggests data locked away in vaults when that is the antithesis of what is required.

The scholarly communications cycle has not until now incorporated the importance of data management as part of that cycle. This is changing as data is seen as a research output of equal importance to publications.

A role for libraries

Libraries are already well engaged with the scholarly communications cycle through the organisation and provision of information resources in the form of publications, manuscripts, audiovisual materials and other formats. This has been accelerated recently with the rapid take up in academic libraries of services designed around the institutional repository, and there are few academic libraries in Australia now which do not manage a repository on behalf of their institution. The requirements for the Research Quality Framework (RQF) and the more recent Excellence in Research for Australia (ERA), while arduous, have delivered a bonus in bringing together repository managers and university research offices with a common purpose, aligning the library with the university's research reporting requirements.

Many academic libraries are now using their repositories for purposes beyond articles and other text-based publications. Some are being used to store image and sound files. We are starting to see repositories being used for the storage of other kinds of data. As the role of data in the scholarly communications cycle becomes more apparent, the institutional repository has an increasingly important role to play in the data lifecycle.

This does not necessarily mean, though, the institutional repository will stay in the same form as we currently know it.

It would probably be true to say that, in recent years, academic libraries have been more engaged with teaching than with research. But there is now a change of focus and this will more than likely develop further.

At the University of California in San Diego, there is an active partnership developing between the library and the Supercomputer Centre, to "build an intersect of personnel, expertise, and services to provide long-term preservation of and access to research data that enables domain scientists and researchers to carry-out longitudinal complex data analysis to support interdisciplinary research" (Schottlaender and McDonald 2007). At Purdue University, a Digital Data Curation Centre (D2C2) has been established within the University Library (Mullins 2007). Australian universities also have been quick to take up the challenge, and examples can be seen at The University of Queensland, The Queensland University of Technology, The University of Melbourne, Monash University and Swinburne University of Technology.

The changing role of libraries in this context has already been recognised by the Association of Research Libraries in the USA which established a task force in 2006 to investigate support for eScience. They reported:

> There is a perception that science librarians, more than ever before, need to be actively engaged with their user communities. They need to understand not only the concepts of the domain, but also the methodologies and norms of scholarly exchange. This level of understanding and engagement goes well beyond knowledge of the literature. It requires being a trusted member of the community with recognized authority in information related matters. This new paradigm suggests a shift in focus from managing specialized collections (the "branch library" model) to one that emphasizes outreach and engagement. (ARL Joint Task Force on Library Support for E-Science 2007)

I see no reason why these changes, however, should be limited to science.

The challenge of open access

Most librarians are already familiar with the open access movement. This developed in the late 1990s in response to the ever-increasing cost of journals and the realisation that information was being denied to anyone who could not afford to subscribe to journals or have access through their place of employment. It was argued, therefore, that publications should be deposited into repositories so that they could be available to anyone, anywhere, free of charge. This is a great idea, but has to take into account the need for refereeing of academic publications and the insistence on the part of many publishers that copyright of journal articles be handed over to the publisher rather than staying with the creator. Originally open access was seen as the preserve of research outputs in the form of text, which is to say journal articles and book chapters. As time has gone on, the idea has taken hold and has taken two main forms: encouraging researchers to make their publications available in institutional repositories set up for the purpose and when publishers will permit, or promoting alternate, ways of publishing which allow open access while maintaining peer review and scholarly integrity.

The notion that data should also be available on open access is becoming more prominent. This presents a different set of requirements for the researcher and for the body responsible for the provision of access and long term storage.

The challenge of data

While open access is generally agreed to be a good thing, it has not been as successful as hoped, with researchers slow to deposit in repositories or take to alternate publication methods. The result of this is that we still have “a system where gateways limit access to research results, and as a consequence only a small fraction of the world's research libraries subscribe to some journals. The gentleman's club survives, if only as metaphor.” (Swan 2007)

There are many reasons for this, not the least of which is that making publications accessible on open access is not well integrated into the scholarly communications cycle. In order for researchers to make their publications available on open access, they have to take steps unrelated to their usual publication processes. Even where deposit of publications has been mandated, deposit rates, while improved, do not reach 100% compliance. Making data available on open access presents a similar, and possibly more complex, challenge.

At present there are not the policies, attitudes, understanding, commitment or mechanisms in place to allow data deposit to occur as a matter of course and it is here that institutional policies and advocacy have a major role to play. Two other developments are having an impact: learned journals, especially in the sciences, are starting to insist that the data on which articles are based are made available together with the article and research funders are starting to insist that all the outputs of research are made more accessible.

The challenge for libraries

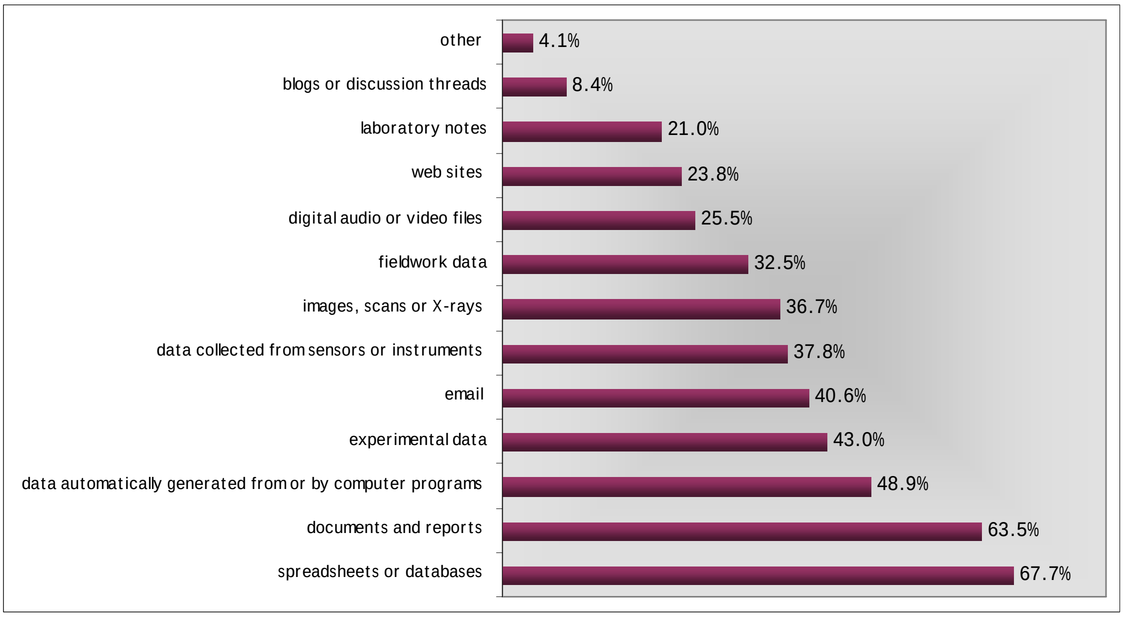
There is a challenge here for libraries to become part of the eResearch lifecycle by taking an active role in eResearch support and data stewardship. Libraries have a long history of preserving knowledge in various forms. No one questions this as a valid role for libraries and knowledge preservation is a key function for society as a whole. There is one major difference though in becoming engaged with eResearch support and data stewardship. Libraries have most often functioned as places where objects are stored, once they have been created by someone else. Creating the mechanisms required to support eResearch implies a different kind of service, one where the library is engaged in the research process.

Libraries are already used to offering services based on the objects which they acquire and organise. Many have become actively engaged in publishing and teaching as well as in storage and access. Libraries now need to define their role in providing eResearch support. This adds another dimension to library services as libraries continue to adapt to the rapidly changing technological and research environment. It will require the use of innovative tools, additional skills, changed organisational structures, a different set of partners and engagement in new collaborations.

Researcher data management practice

Three Australian universities surveyed their academics, post-graduate students and other researchers in late 2007 to get a better idea of what researchers are actually doing with their research data. Their aim was to use this information to improve support services. The three were The University of Queensland, The University of Melbourne and Queensland University of Technology. The Data Management Survey questionnaire was completed by a total of 879 academics, post-graduate students and adjunct appointees who ticked boxes and offered suggestions, comments and occasionally criticism. (Henty, Weaver et al. 2008) The full questionnaire included over 20 questions, but responses to only three are included here to provide some indication of the issues involved in data stewardship and support services.

Figure 1: Types of digital data held



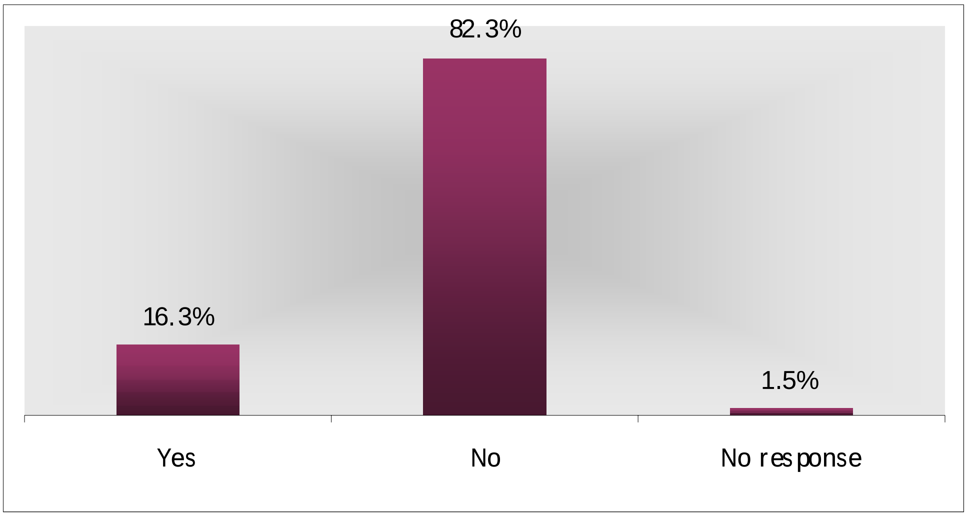
It's apparent from Figure 1 that there are all kinds of digital data being created. Spreadsheets and databases are the most common, with documents and reports not far behind. These are not necessarily published and refereed, and include drafts, survey responses, client files, transcripts, translations and more. More text can be found in emails, websites, laboratory notes, blogs and discussion groups. Raw data comes from sensors and instruments, from computer programs and from the recording of experiments. There are image files of various kinds, not to mention sound and video. Some researchers collect only one type of digital data: most collect many.

The category "other" referred to in Figure 1 revealed bibliographies, biographies, online surveys, secondary data analysis, questionnaires, bibliographic databases, mathematical models, simulations, interview transcripts, computer programs, satellite imagery, GIS data, CAD models, genotyping and sequencing data, electronic health records, music scores, podcasts, laser scanning imagery, GPS measurements, mind maps, flow cytometry and spectral data, and computational fluid dynamic codes. This all provides an idea of the kinds of challenges ahead for data stewardship.

One way for data to be better organised and managed from the outset is for researchers to have individual data management plans. The survey asked the researchers whether they currently have one. The answer can be seen in Figure 2, and is overwhelmingly no.

Data management planning sits within institutional, disciplinary and regulatory contexts of policies and expectations. The Australian Research Council (ARC) and the National Health and Medical Research Council (NH&MRC) have regulations covering the ongoing storage of data, and currently there is a requirement that all data should be kept for at least five years. This is expressed in the *Australian Code for the Responsible Conduct of Research*. (Australian Government 2007) Most Australian universities reflect this requirement in their internal policies, but it is questionable whether these are sufficient to ensure long term data management and sharing.

Figure 2: Do you have a formal research data management plan?



When asked whether they would be interested in training or advice on a range of topics, about one-half said that they would like training on how to develop a data management plan before starting a project. About one-half said that they would like training on developing a data management plan after completing a project and about one-third said they would like to develop a data "exit" plan (offered for retiring or departing academics). About one-quarter asked for help with data rescue for older digital materials such as data on older media or data from legacy systems. About one-third asked for help with digitisation.

Capabilities

APSR conducted another survey in 2007 which gives us some insight into the gap which exists between the potential for eResearch and current practice and data infrastructure support in terms of the needs for the development of skills and capability. (Henty 2008) The scope of the study went beyond the role of the library to the way in which the whole institution in involved in eResearch support.

The report's conclusions were based on a number of sources. Interviews were conducted with twelve key established researchers in six Australian institutions, with a focus on academics engaged in data-intensive research. In addition, we conducted a questionnaire survey of those who attended eResearch Australasia 2007, and used the results of a small questionnaire conducted by SAPAC (the South Australian Partnership for Advanced Computing). A fourth source of information was an informal survey conducted at a workshop held in association with eResearch Australasia 2007: ‘The Researcher/Librarian Nexus: The challenges of research data management in institutional repositories’ where a group of librarians identified the skills which they saw as being needed to take on a role in repository management and data stewardship.

Two themes emerged from the various surveys and interviews. One related to the development of institutional capability to overcome current barriers to eResearch and to institute cultural and organisational change. The other concerned skills and the need for training and staff development.

By institutional capability we mean everything which might contribute to the successful conduct of research. This includes having the right people, expert and available knowledge, well-defined business processes, excellent facilities and equipment, functional information and communication technologies, and well-established accountability and governance.

The study showed that there is a lot of work to be done on all of these fronts and that some organizations are working hard to overcome gaps. The most commonly mentioned barrier to eResearch was that of skills, not just the existence of the skills, but knowing how to locate them within the university’s organizational structure. As someone commented: ‘It always seems to be “somebody else” who will provide the ground level support.’

Skills

The list of skills identified for the support of eResearch was huge. The skills were both technical and non-technical. At the technical end there was mention of high-performance computing (HPC) and the access grid, data (and database) management, data curation, information engineering, information modelling, software development, remote communications, distributed processing, informatics, portal design, computational fluid dynamics, database integration, metadata, visualisation and programming of all kinds. At the non-technical end, which is to say the non-ICT end, were data analysis (including the use of statistical packages and other techniques such as data mining), information seeking, project management, business analysis, communications, negotiation, intellectual property, team building and train the trainer. It is interesting here that few of the skills mentioned correspond with those selected by respondents to the Data Management Survey. There are several reasons for this. The Data Management Survey was addressed only to researchers and provided options to tick, rather than asking an open ended question of a more general nature.

Libraries have a long and honourable tradition of being involved with information literacy. Perhaps the skills associated with data management planning could also been seen as part of information literacy. Like information literacy though, the disciplinary context is all important. At the post-graduate level especially, there is a strong need. The Australian National University has a Graduate Information Literacy Program, for which participants receive a qualification if they complete a set number of modules. A recent addition has been a module on How to Write a Data Management Plan which is being received well, with comments along the lines of "I have a problem I didn't know I had". Creation of this module has been funded by APSR and it will be available for download from the APSR website at www.apsr.edu.au for local adaptation.

There are different groups of people engaged in the eResearch process, and hence a range of knowledge associated with the provision of support. At one end of the spectrum there are the researchers, who know about the discipline, but should not need to know everything about data management and stewardship. Researchers are often supported within their research teams by data experts who are able to contribute their analytical, disciplinary, computing and systems design skills to the design of the research project. This group are sometimes referred to as data scientists, although not all are happy with that term. Some might come from a background in disciplinary based informatics – bio-informatics or chemo-informatics, for example. Then there are those who have a part to play in the data life cycle by offering longer term access, preservation and curation skills. There is still work to be done in defining the skills needed by each group and the best way of providing them.

Skills are a key issue for libraries in this context and this has already been recognised at the 2008 ALIA Skills Summit:

> Information management and information technology skills. As academic libraries assume greater responsibility in their institutions for management of repositories and research data, it is clear that there is a dearth of people who have the skill to work in these areas. (Wells 2008)

The issue is one both for library schools which are training new librarians and also for existing staff who need to update their skills.

ANDS

Research institutions in Australia will not have to face all of these issues alone, as the proposed Building Capabilities program within the Australian National Data Service (ANDS) will be able to provide some guidance.

The Australian government has shown significant interest in developing research infrastructure in recent years, with investments in a range of projects. Notable among these were the projects funded under the Strategic Infrastructure Initiative announced in 2003. These included the Australian Partnership for Sustainable Repositories (APSR) and the ARROW Project, both with a focus on repositories. At the same time, we have the National Collaborative Research Infrastructure Strategy, designed to "provide researchers with major research facilities, supporting infrastructure and networks necessary for world-class research" (NCRIS 2008). Through NCRIS, the Government has committed $542 million over the six years, 2005-2011.

Part of the funding provided to NCRIS will be used to fund ANDS. This is described in *Towards the Australian Data Commons* as "the essential meeting place where the Australian path forward for research data management can evolve and where a vision can be achieved". (ANDS Technical Working Group 2007)

ANDS will be launched during the second half of 2008 with initial funding allocated to cover three years, but this is only the beginning. ANDS will have its headquarters at

Monash University and will work with the Australian National University in delivering its services. At the same time, ANDS will be a partnership, not just between Monash and ANU, but including researchers, disciplines and institutions that are ready, willing and able to contribute. Institutions that work with ANDS will be able to gain access to outside datasets, services, tools, guidelines, best practice, additional staff expertise thereby increasing their profile through greater data and research visibility. ANDS will facilitate access to their data, in order to learn from their best practice, to benefit from their expertise and experiences in the various forums and to commission specific pieces of activity.

Australian researchers are keen to have greater access to data and to see data managed with greater efficiency. The establishment of the Australian Research Data Commons is therefore a high priority. Around half of the money in the ANDS budget will be spent outside the core participants, funding institutional activities of various kinds.

ANDS will have four programs: Developing Frameworks, Providing Utilities, Seeding the Commons and Building Capabilities.

A major task of the Capabilities Program will be to develop a Capability Maturity Model for eResearch infrastructure support. This will enable ANDS to provide measures whereby a research institution will be able to measure its own performance against guidelines set for each level of maturity. The measure will cover all the aspects of capability mentioned above: people, knowledge, business processes, facilities and equipment, information and communication technologies, and accountability and governance. Ultimately we envisage being able to audit and certify institutions which are part of the ANDS community.

In its first year of operation, the Capabilities Program will have much to do. It will establish the Australian Data Commons Content Forum, it will engage the community on capacity and capability constraints and the means of overcoming them, it will build communities of best practice and start to address the skills shortage by offering training and by working with the higher education sector to upgrade and improve courses.

Conclusion

The research landscape is changing rapidly in response to the possibilities offered by improved communication and technical infrastructure, leading to what is now generally referred to as eResearch. Libraries are well situated to take a part in providing some of the infrastructure required to support eResearch, especially access, use and reuse of data, and storage and preservation. This is an exciting time to be working in libraries as they take up the challenge of new roles and responsibilities.

REFERENCES

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, http://www.pfc.org.au/pub/Main/Data/TowardstheAustralianDataCommons.pdf.

ARL Joint Task Force on Library Support for E-Science (2007). Agenda for Developing E-Science in Research Libraries: Final Report and Recommendations. Washington DC, Association of Research Libraries, http://www.arl.org/bm~doc/ARL_EScience_final.pdf.

Australian Government (2007). Australian Code for the Responsible Conduct of Research. Canberra, http://www.nhmrc.gov.au/publications/synopses/_files/r39.pdf.

Henty, M. (2008). "Developing the Capability and Skills to Support eResearch." Ariadne(55). http://www.ariadne.ac.uk/issue55/henty/

Henty, M., B. Weaver, et al. (2008). Investigating Data Management Practices in Australian Universities. Canberra, Australian Partnership for Sustainable Repositories, http://www.apsr.edu.au/investigating_data_management.

Mullins, J. L. (2007). Enabling International Access to Scientific Data Sets: Creation of the Distributed Data Curation Center (D2C2). International Association of Technological University Libraries, Stockholm, Sweden, IATUL. http://www.iatul.org/doclibrary/public/Conf_Proceedings/2007/Mullins_J_full.pdf

NCRIS. (2008). "National Collaborative Research Infrastructure Strategy (NCRIS)." http://www.ncris.dest.gov.au/.

Schottlaender, B. and R. H. McDonald (2007). Data-Cyberinfrastructure Collaboration at the University of California, San Diego Project Briefing CNI Fall 2007 Meeting, Washington DC, CNI. http://www.cni.org/tfms/2007b.fall/Abstracts/PB-data-schottlaender.html.

Swan, A. (2007). "Open Access and the Progress of Science." American Scientist **95**(3): 197-199

Wells, A. (2008). ALIA Education and Workforce Summit 2008: Submission from the Council of Australian University Librarians. Canberra, ALIA http://www.alia.org.au/education/summit08/caul.pdf.