Sustainability Issues for Australian Research Data

The Report of the Australian e-Research Sustainability Survey Project

Dr. Markus Buchhorn, The Australian National University

Paul McNamara, The Australian National University

APSR
AUSTRALIAN PARTNERSHIP FOR
SUSTAINABLE REPOSITORIES
http://www.apsr.edu.au

Document Version 1.1 (October 2006)

Dr. Markus Buchhorn undertook the survey through his participation in the Australian Partnership for Advanced Computing.

Paul McNamara undertook the survey on secondment to the Australian Partnership for Sustainable Repositories.

This survey and its report were commissioned by the Australian Partnership for Sustainable Repositories (APSR), a broad coalition of institutions promoting excellence in digital research collections for the higher education sector. APSR is considering the issues raised by the report, and is keen to pursue them with interested readers.

Contact:

Dr Adrian Burton

Leader, Australian Partnership for Sustainable Repositories

enquiries@apsr.edu.au

(02) 6125 6659

Document URL: http://www.apsr.edu.au/aeres/

"Due consideration should be given to the sustainability of access to publicly funded research data as a key element of the research infrastructure. This means taking administrative responsibility for the measures to guarantee permanent access to data that have been determined to require long-term retention. This can be a difficult task, given that most research projects, and the public funding provided, have a limited duration, whereas ensuring access to the data produced is a long-term undertaking. Research funding agencies and research institutions therefore should consider the long-term preservation of data at the outset of each new project, and in particular, determine the most appropriate archival facilities for the data."

Draft OECD recommendation concerning access to research data from public funding, DSTI/STP (2006) 18, Consultation version May 2006.

Table of Contents

Acronyms

ACSPRI	Australian Consortium for Social and Political Research Incorporated
AERES	Australian e-Research Sustainability Survey
AHDS	Arts and Humanities Data Service (UK)
AHRC	Arts and Humanities Research Council (UK)
ANU	Australian National University
ARC	Australian Research Council
APAC	Australian Partnership for Advanced Computing
APSR	Australian Partnership for Sustainable Repositories
ARROW	Australian Research Repositories Online to the World
ASSDA	Australian Social Sciences Data Archive
AUQA	Australian Universities Quality Agency
AVCC	Australian Vice-Chancellors' Committee
CRC	Cooperative Research Centre
CSIRO	Commonwealth Scientific and Industrial Research Organisation
DART	Dataset Acquisition, Accessibility and Annotation e-Research Technology Project
DEST	Department of Education Science and Training
FRDC	Fisheries Research & Development Corporation
IGS	Institutional Grants Scheme
LIEF	Linkage Infrastructure Equipment and Facilities
MAMS	Meta-Access Management System
MNRF	Major National Research Facility
MRC	Medical Research Council (UK)
NARA	National Archives and Records Administration (USA)
NCGP	National Competitive Grants Program
NCRIS	National Collaborative Research Infrastructure Strategy
NERC	National Environment Research Council (NERC)
NHMRC	National Health and Medical Research Council
NIH	National Institutes of Health (USA)
NSB	National Science Board (USA)
NSF	National Science Foundation (USA)
OECD	Organisation for Economic Cooperation and Development
RCUK	Research Councils United Kingdom
RIBG	Research Infrastructure Block Grant
RLG	Research Libraries Group (USA)
RQF	Research Quality Framework
RUBRIC	Rural Universities Building Research Infrastructure Collaboratively
SII	Systemic Infrastructure Initiative

EXECUTIVE SUMMARY

The Australian e-Research Sustainability Survey (AERES) project was undertaken by the Australian Partnership for Sustainable Repositories (APSR) and the Australian Partnership for Advanced Computing (APAC) to survey the sustainability issues for data-intensive research projects, including the capabilities and demands of research groups and institutions for the storage, access, and long-term management of research data.

The immediate and critical issue for the stewardship of research data in Australia is the lack of administrative responsibility for the task.

The current policy framework for research data in Australia is provided by the funding rules of the Australian Research Council (ARC) and the National Health and Medical Research Council (NHMRC), by Records and Archives legislation and by the *Joint NHMRC/AVCC Statement and Guidelines on Research Practice*[1], currently under revision. This framework is currently lacking guidelines for clear administrative responsibility for data stewardship.

The survey found that researchers are both providers and consumers of data and have a broad range of needs for research data and its management. There are strong disincentives for researchers to engage with long-term data management. There is little recognition for good data management. Researchers do not see a national data management system with which they can work.

The current data management infrastructure is in general decentralised and uncoordinated. This data management infrastructure needs to be more closely aligned and coupled with the evolving policy framework for data stewardship. Moreover, this infrastructure would benefit from greater systematic recognition in the policies of institutions, government and funding agencies.

A mature data stewardship system, interlinking policy and infrastructure could address the needs of researchers and improve the quality and efficiency of Australian innovation and research. A successful data stewardship system needs to:

- identify administrative responsibility;
- address disincentives for researchers to manage data for the future;
- strengthen the engagement of researchers, universities and funding agencies; and
- encourage the development and sharing of skills

The technological challenges of data management are also significant and ongoing. Work funded by initiatives such as *Backing Australia's Ability* has begun to address some of these challenges. The potential of these technological solutions can best be realised within an appropriate policy environment.

Government, policy-creators, funding bodies, and research institutions have an opportunity, and perhaps an obligation, to assist in the development of a coherent data stewardship system.

1. INTRODUCTION

The Australian Partnership for Sustainable Repositories (APSR) has established a centre of excellence for the management of scholarly assets in digital format, focusing on the critical issues of access continuity and the sustainability of digital collections. APSR consists of partner institutions using their repository facilities to develop projects in digital continuity and sustainability. By providing national coordination throughout the sector, APSR is developing skills and expertise, strengthening this area.

The Australian Partnership for Advanced Computing (APAC) was established in 1998 to lead the development of an Australia-wide advanced computing infrastructure, supported by coordinated programs in research, education and technology diffusion. The role of APAC is to contribute to the advanced computing, data management and grid services for Australia's research communities.

The Australian e-Research Sustainability Survey (AERES) Project was undertaken by APSR and APAC to survey the sustainability issues for data intensive research projects, including the capabilities and demands of research groups and institutions for the storage, access, and long-term management of research data.

The survey authors were mindful of recent and ongoing developments in Australian research infrastructure resulting from the *Backing Australia's Ability* program.

Backing Australia's Ability – Building our Future through Science and Innovation[2] is an Australian Government package delivering $5.3 billion over seven years from 2004-05 to pursue excellence in research, science and technology, through three key themes:

- the generation of new ideas (research and development);
- the commercial application of ideas; and
- developing and retaining skills.

The Systemic Infrastructure Initiative (SII) is part of *Backing Australia's Ability*, funding required upgrades to the systemic infrastructure of universities. Innovative approaches include linking or expanding access to shared facilities - such as libraries, information and communications technologies, and providing specialised equipment, technical and administrative assistance.

SII has funded a number of significant projects for the development of data annotation, preservation and sharing at the level of technology and practice:

- Australian Partnership for Advanced Computing (APAC)
- Australian Partnership for Sustainable Repositories (APSR);
- Australian Research Repositories Online to the World (ARROW);
- Dataset Acquisition, Accessibility and Annotation e-Research Technology project (DART);
- Meta-Access Management System (MAMS); and
- Rural Universities Building Research Infrastructure Collaboratively (RUBRIC).

Collectively, these projects are strongly focussed on developing infrastructure and capability within universities to enable the preservation of, and access to, digital data. They are capable of evolving into a repository framework to support the preservation of Australian digital research data.

The AERES results can be used to refine development paths for Australian institutional repositories, identifying threats to the sustainability of research data and demonstrating how repositories may respond to these.

This project follows on from two reports issued by APSR in 2005. The *APSR Sustainability Issues Discussion Paper*[3], issued in January 2005, examined technology, risk management and economic aspects of the sustainability of digital assets. The *Survey of Data Collections*[4], issued in December 2005, provides data for the design of software tools, development of repositories, assessment of risk and the implementation of preservation metadata. The survey includes an indicative table for use in assigning data collections between institutional repositories and specialised repositories.

The AERES methodology included a consultation stage and a survey stage. The consultation stage involved talking with a number of interested parties and senior academics as background to ascertain or clarify their views of the National Collaborative Research Infrastructure Strategy (NCRIS) and the e-Research Coordinating Committee agendas. The Academies, the Federation of Australian Scientific and Technological Societies, CSIRO, the Australian Research Council (ARC), NCRIS, APAC and the National Health and Medical Research Council (NHMRC) were contacted in this stage. Other inputs included the second consultation draft of the *Australian Code for the Responsible Conduct of Research*[5], the *NCRIS Strategic Roadmap*[6] and the *Interim Report of the e-Research Coordinating Committee*[7].

The survey stage involved interviews with individual academics, heads of research groups, administrators, managers and IT staff from approximately 40 groups or service providers. Dr. Markus Buchhorn (APAC) and Paul McNamara (APSR) conducted these interviews. Some of the interviews for this stage were conducted as part of an associated APSR project: *Sustainable Paths for Data-intensive Research Communities* implemented at the University of Melbourne. Eleven diverse research communities were audited by the AERES and Melbourne project teams and provided with some local follow-up support consultations to address key support issues raised during the audit.

The survey stage revealed that the lack of an appropriate infrastructure is an issue for the long-term retention of research data. Accordingly an additional survey was incorporated into the original project. This element was intended to clarify the extent of planning and policy development for the retention of research data within a sample of the university sector. Deputy Vice-Chancellors, Pro Vice-Chancellors, Chief Information Officers, Librarians and other senior staff holding information portfolio positions were interviewed by Margaret Henty (APSR) and Paul McNamara (APSR).

The initial survey stage also revealed that there are compliance requirements for the retention of research data within funding guidelines, but it is unclear how well these are observed.

2. THE SURVEY

2.1 RESEARCHERS

The authors talked with 80 research or research support staff in 40 groups across a number of research organisations and research service providers. The staff included individual academics, heads of research groups, Deans and IT officers. The organisational units included departmental and faculty groupings, as well as Cooperative Research Centres, an ARC Centre of Excellence, a Major National Research Facility (MNRF) and a number of designated university centres. Information was also obtained from staff associated with ARC Special Initiative E-Research projects.

Most of the groups we approached viewed the issue as sufficiently important for the head of the group to speak with us. This meant we spoke to more senior researchers than early career researchers, and reflects the importance they place on e-Research.

More science groups were surveyed than non-science because of a desire to understand the management of complex data sources. This meant we spoke to more male than female researchers. Science units are also typically aggregated into groups or clusters, which meant we were given a community view of data management. Social science and humanities researchers were more likely to be undertaking individual research.

All units were receiving combinations of grant, university and external funding, and all groups surveyed were involved in collaborative research with other researchers and research units. Most groups were providing advice to government, agencies and commercial enterprises, with a minority having collaborative links with industry.

Attempts to develop a representative survey group were limited on several counts including the practicalities of coordinating visits with busy researchers. We also took advantage of APSR partners who identified target groups. Some groups we approached responded that their data practices were unworthy of investigation – typically, 'We all keep it on our own machines'.

Data service providers were also included in the survey, to determine their experiences in data management and what part they might play in data preservation, as well as any lessons they had for repositories.

Awareness of e-Research issues

The surveyed researchers had a strong awareness of the e-Research agenda and issues. This awareness extends to the specific issue of the enduring value of research data. All groups were tracking NCRIS activities and, where applicable, NCRIS timelines in their respective disciplines.

There is a general belief that publicly funded data should be made available to all, within appropriate access and security guidelines. This belief is more strongly held by researchers with management responsibilities than at the practitioner level.

Motivation

It is clear from the survey that the research world is strongly focussed on publication as an outcome, as it is seen as providing certification of quality. Publications, citation rates and impact factors are the traditional research quality indicators. The Research Quality

Framework (RQF) exercise has stimulated a strong debate on quality versus impact. In the words of one of our interviewees*:

> "Data is not the object of the research so it is not valued in the same way as the outcome – the journal article. All recognition is about the journal article."

Researchers are concerned about the rewards and anxieties normally associated with having and keeping any job. They are working in a highly competitive environment that uses qualifications for certification and relies on publication and citation for career progress and peer recognition.

The research environment is also heavily based on a funding cycle that perpetuates a focus on short-term results and the generation of outputs that will assist in obtaining the next grant.

The publicly funded research environment is a refined system clearly understood by funding bodies, researchers and administering institutions. Researchers indicated that elements of this environment are tolerated because they are regarded as essential to their core business. Researchers have learnt to engage with the system in a way that maximises both their chances of success and the time they can actually spend conducting research. The question arises as to the best way of encouraging changes to data management practices, either from within the existing system or introduced elsewhere in the work environment.

Policies

There is great diversity in the organisation, governance and funding of research groups. There can be considerable autonomy for researchers and projects within larger structures.

The survey found that research groups and organisations rarely have formal policies for the management of data. They usually have a set of practices that may or may not be adhered to at the project level. These practices relate to the storage, backup and security of the data and how long the data must be kept. Implicit in these practices is the understanding that research data arising from a project belongs to those undertaking the research and is their responsibility. They make appropriate arrangements for data security depending on an assigned value of the data, given the project.

> "I tell the students that your data is like your children – it's your responsibility."

All researchers surveyed understood that research data must be retained for a nominated period of time. Most thought this period was five years. This figure comes from the *Joint NHMRC/AVCC Statement and Guidelines on Research Practice,* which has been widely adopted by universities, and recommends a minimum period for retention of at least five years from the date of publication. Researchers are not sure whether the five years is a university requirement or a funding body requirement.

* Throughout the document, verbatim quotes from interviews are formatted with lines above and below the text.

> "Who will store data for post docs or retired staff after they go, out to the duration of the five years?"

Project planning

The survey found that individual researchers and research groups do not include data management as an element when planning research projects. It is not considered in the project planning phase beyond what is necessary for the relatively brief duration of the project. Researchers are focussed on the research activity and the outcomes. It is typical for there to be little consideration given to format, storage, standards or metadata. Past practice, group software and knowledge standards are more important in determining data practices than considerations for future use. The data is an input to, or a product of research, but it is generally not considered as an outcome.

Funding

Grants do not fund the creation of datasets as an end in itself, nor are funds provided explicitly for the management of data. Researchers saw the requesting of funds for data management as potentially jeopardising a grant application. Researchers are concerned that emphasising the data aspects of projects may lead to a reduction in the funds provided for 'real' research.

> "It will always be difficult because infrastructure reduces the dollars available for primary research."
>
> "Funds for this sort of infrastructure should be separate from and in addition to research funds."

The grant funding cycle is a key driver for the organisation of research work; short-term project grants do not fit with the creation of systemic infrastructure within current policy frameworks. Funding is not provided for the long-term storage and use of research data in the current system.

Data ownership

Ownership of research data completed from projects funded by the ARC and the NHMRC rests with the organisation with which the Funding Agreement has been completed. This was not understood by most researchers.

ARC Funding Agreements for Discovery, Linkage, Centre and Linkage Infrastructure Equipment and Facilities (LIEF) all require the organisation with which the Agreement has been completed to, *"…establish and comply with its own procedures and arrangements for the ownership of all Material produced as a result of any Project funded under this agreement."*

The NHMRC Funding Agreements for Researcher Support Schemes and Research Funding Schemes stipulate that the institution with whom the agreement is made owns the 'project material' and 'award material'.

The draft Australian Code for the Responsible Conduct of Research states that the Institution must, *"Develop policies and procedures for research data ownership and storage…."*

Researchers, particularly individual researchers, see research data as belonging to them. There is no organisation that manages their data that reinforces researchers' belief that they own the data. Experienced researchers often have collections going back to the beginning of their career, which may or may not be readable, may require obsolescent equipment to act as a 'translator' and may be stored in unsuitable conditions. Experienced researchers have been managing data all their careers. They ask the question, 'Who else could this data belong to?'

The life cycle of data

Researchers distinguish between a number of stages in the life cycle of data. Data is collected/generated and analysed. The data then serves as a source for scholarly output. There may be a considerable gap between data collection and publication. Once publication starts to occur, the data may be needed because there may be queries relating to the data or requests for collaboration as a result of scholarly publication. Data is then of historical interest to the researcher.

The life cycle for data collected for, or generated by, research projects is linked to its use by the researcher or group and/or by the nature and scope of the data and its potential use to others.

Most researchers surveyed had not used historical digital data collections at an organisation or group level. They do have experience with these collections at an individual level. Most reported that they rarely used their own older data sets or were asked by others for them. Some maintained older equipment and software for the rare occasions when they needed to access this data. Even when possible, migration to newer formats and mediums was not seen as worth the trouble. Researchers were storing data collections without any way of accessing the data, in the hope that they might get around to doing something one day or that the department or the university might take it on when they retire.

> "I download to floppies and then to CD's and I keep an old MAC as a translator for the older stuff."

Researchers who believe data should be available for others distinguish between data that is 'live' to their project and data which has reached an historical stage for the group. This has implications for any organisation with the task of managing research data.

There were differences of opinion about when data should be lodged with a repository. The majority of groups consulted saw the data as belonging with the group for the life of the project. A minority thought that data lodgement for management and preservation needed to take place during the life of a project not at the end of the project. This view was more prevalent for groups in organisations with data centres. The reasoning was that lodging data at the end of the project reduced the chances of the data being lodged at all because the researchers would be focussed on the next project.

Data management

Research groups surveyed were found to be undertaking their own management of the research data generated for, or created by, their research. Researchers see the benefits of this situation but are less aware of the disadvantages.

If data is important to a project, it is managed sufficiently for the needs of the project, but not necessarily managed to the standards that professional IT staff would use. Most groups manage the data themselves using research group staff or students (frequently PhD students). Some research groups manage data with group or faculty IT staff, and some with university IT bodies or external groups.

Most of the researchers surveyed believe that for data management to be undertaken in a competent fashion there needs to be discipline or project knowledge.

Good research depends upon an understanding of the data collection processes and the data. Manipulation of data, either by high level use of off-the-shelf software, or by writing applications, is part of the process by which PhD students and early career researchers obtain this understanding. The role of data interpretation blends into and with that of de-facto data manager in most research groups. There is some inefficiency in this unavoidable situation, which is necessary whilst the research group is actively using the data.

Researchers wish to know where their data is stored, who is looking after it (if not them) and what understanding of the data these people have. They want to know what the security and backup arrangements are for their data. Ease of access and assured control often determine the location of data. A trust relationship with local support staff also often determines the location of data. Data management is influenced by volume, size, complexity and data formats. Raw, processed and reduced data may be managed differently. We found instances of data being saved as snapshots from critical points because of the amount of raw data being generated.

Researchers were unaware to a significant extent of the skills and capabilities within their discipline or within their required technology, even where those skills were world-class and within the same institution. Many groups indicated that finding required skills and services was a major issue, sometimes leading to the reinvention of wheels. There do not appear to be mechanisms for universities to register the rich skill sets within the university for their own internal use. It was frequently noted that it would be beneficial to the work of the university to make better use of hard-won and expensive skill sets within research groups. It would also be beneficial to look at coordinating the data management skills required by the institution. At present these skills reside in the research groups, data centres, IT centres and repositories.

Time, funds and a competitive environment encourage researchers to do that which is necessary for data management and no more.

Awareness and usage of central services

There is little awareness of central IT services within the university that might be of use to researchers in managing data. This is not surprising given their beliefs and practices in data management. These practices are inwardly focussed for the reasons discussed under ‘Data management’.

The general belief is that it is the researchers’ data and the university, faculty or department expects the researchers to solve their data problems when they deviate from the norm in format, volume, size or application. This belief is equally matched by a general understanding by the providers of IT services that they supply an infrastructure and a base level service which is then built on by faculties, schools and departments. Central areas of universities expect that specialised needs are met by those creating the need. Specialised needs are typically met by the purchase of equipment and software as well as the short-term hiring of necessary skills and/or the acquisition of the skills by those researchers and students associated with the specialised need.

> "If it was important enough to a researcher or to a critical mass of them their own department would do something and go off and buy a server of which there are an enormous number and probably buy some software and hire someone to look after it."
>
> "The review asked us why we were storing and maintaining the data behind the datasets and the answer was who else could have done this in the university at the time."

The level of knowledge about central services is directly related to whether researchers trusted them. If researchers did not know of any services that would meet their data management needs they would then be reluctant to use them until a level of knowledge and trust was established. Where services did exist, the fact that researchers did not know about them was considered to be an indication that the services were unreliable.

> "I don't know of any [service] and since I don't know I doubt that I would be happy to use it."
>
> "There's no mass storage that I know of on campus."
>
> "If we used the University's services we would have to pay, we would have less control – doing it ourselves we save time and we have guaranteed access without sharing facilities."

A notable exception to the awareness and use of central services (at a local, regional and national level) were groups that needed specialised advice in supercomputing services and to some extent, mass storage. This advice was less about the management of the data than about manipulating it around a grid infrastructure to bring data and computing power together. The use of this advice and these services was related to the level of development of supercomputing activities at the institution. The use of this infrastructure and services across a network is informative for the practices that might be developed for a decentralised national repository service.

Data sharing

There is a general belief that publicly funded data should be made available to other researchers within appropriate access permission regimes. This belief was more strongly held by researchers with administrative responsibilities than individual researchers.

> "It's public funds and belongs to the taxpayer, we just have a licence to work with it for a while."

Researchers are often unwilling or reluctant to share data generated for or from their research. This may be because they believe the data is 'their' data, because they believe the data is not of consequence beyond their project or because of practical difficulties in accessing the data. This difficulty may include their recognition that there is no system or infrastructure for the sharing of data. Researchers may also be reluctant to share data because they have been trained in a discipline that historically does not do this.

> "We generally don't share - there is competition to be the first to publish. Data is seldom presented in our field, but people report when and how they collected it."
>
> "The Centre has a limited tenure so we looked to the Centre partners for archiving but with little success – we also asked government but were told it's our contract and our business."

Data is shared where there is a need to do so. Large and commonly used datasets such as astronomical, ocean and climate data, or genetic information, are shared as a matter of community interest. Communities that depend upon scarce or expensive instrumentation are prepared to share the gathered data.

> "The scientists who are tied into international programs know and value data sharing."
>
> "Communities never exist at the publication level, where they compete; they may exist at the analysis level and often exist at the collection level where they cannot individually afford the data collection."

The development of standards is critical for the sharing of data. Research communities with effective data sharing have developed standards based structures for sharing data. Standards do not exist for some domains, whilst in others there is more than one relevant standard and no advice is available for appropriate choices. Communities without standards need assistance in this process and can be guided by intermediaries such as APSR or APAC, who can bring together researchers who have been through this experience with those still to do so. Such standardisation of formats and metadata can take a very long time to achieve as it relies on consensus within a research community or national and international bodies.

> "We don't have agreed formats…they are on the way but assisting the discipline comes second to personal/group/project survival, which means real research and publishing. The time is right when you have a number of people in different institutions with the same interests…it moves from a cottage to an industry base."

Data held by individuals or small groups is often only discoverable through other researchers reading journal articles or personal contact. There is little chance of researchers outside the domain discovering this data.

> "A researcher writing a grant application contacted me in 2005 about data I produced in 1982 that he found out about through one of my articles."

Researchers are approached by others asking to see their data or use their data. These requests are based on other researchers wanting to check the published results of research or more commonly by researchers with a similar interest and/or similar data who are seeking collaboration. Collaboration is a much more common result of data being available than authentication of research.

> "We were approached by a group who had conducted a broader survey in another region – their survey included our speciality, so we will be able to

> work together. We will be able to do deeper interpretation for them of some aspects of their data that we are familiar with and we can use other aspects of their data to build new correlations for our work."

Most researchers reported that they did not use data from other researchers and that others did not use theirs. Where researchers were aware of and interested in the data of others they were prepared to recreate it if they thought it was too difficult to acquire. They were also prepared to recreate it if they were concerned about the reliability of the data or of the methodology behind the data collection.

> "It's probably not cost effective to recover my older data – if I needed it I would redo the experiments and with the current equipment get better results."
>
> "We need a database of the data that is available to save us hunting around or spending money and time doing it again."

Some researchers reported that carrying out quality control on their own data had made them aware of the possibilities for error in data collections. Researchers did not know how others might use their data. They were concerned about other researchers drawing erroneous conclusions from it, using it 'out of context' and not having regard or understanding for how the data was derived.

> "We did data checking on a set that had been collected and entered manually and picked up a surprising amount of errors – this raised the question of errors in other people's data and whether we could trust it."
>
> "I prefer to handle requests on a one-by-one basis rather than making the data generally available and letting people anonymously use it. The data may be corrupted in some way if that were to happen – they would wrongly sample it or draw erroneous conclusions."

All researchers who worked with data collected from survey communities expressed strong concern about the importance of maintaining a trust relationship with that community. This trust may include an agreement or understanding about how the data might be used and who might have access to it. There was recognition that research data collected from survey communities may inform policy development, but that the policy may not be what the survey community desired.

> "It is important for the people creating the data to be sensitive to the needs of the groups that it is collected from and what sort of repository they might need as future clients for this rare material."
>
> "The study will continue for years so it's very important that we look after the group and not risk losing their goodwill and cooperation."
>
> "We work with rural communities so even de-identified data could probably be used to identify individuals – we won't risk the survey by doing this."
>
> "The object is to build a resource that the community can maintain because they have the knowledge – we have to be sensitive to their cultural constraints and to a technology that they are comfortable with."

> “They were happy to work with us once they saw what we did with the information and as we gave them feedback that helped them run their business better. It was a slow job to build this relationship.”

Researchers are often reluctant to share their data. Data is viewed as a personal good, created by researchers and to be exploited by them. While many researchers feel data collections should be available to the research community there is a very strong and unanimous view that researchers should be able to exclusively exploit ’their’ data for a period of time before it becomes available to others. Some domains, such as those with publicly-funded instruments, have established policies for time limited data exploitation.

Younger researchers were happier with the idea of sharing data. It is not clear whether this attitude is related to their generation or their career stage. The age profile of the academic workforce means that generational change factors will not be as strong as in other occupations. The age profile also means that there is a looming wave of retirements with associated requests for institutions to deal with real and digital collections.

There was concern that shared data may allow others to gain from the work of a group without recognition of that group. There was also a concern that further analysis of data by others may diminish the impact of research already reported.

A significant barrier to sharing is the perceived lack of security and trust frameworks for controlled access to data, with user-managed policy frameworks. Awareness of available technologies and methods was low. Once they were informed of them, many groups indicated an increased level of comfort with depositing data elsewhere.

> “We wouldn’t accept that it should be available to just anyone, there would have to be appropriate frameworks.”

It was clear from the survey that the culture for data sharing is growing with the advent of the Internet and associated technologies - there appears to be a cultural change amongst research groups.

> “It is a cultural change – for so long we have not exposed methods and data. The amounts were so small compared with what is now generated.”

Data preservation

Researchers believe that not all data should be kept and a value decision is required in deciding what should be retained. Researchers currently make value decisions about data. Value is based on uniqueness, time slice, cost and ease of reproduction, but not always with an appreciation of the opportunities for re-use outside of their project or even their domain. Value decisions are also made by omission as data collections are stored but not migrated across media or software.

> “There is a lot of data across the combined units but only material recorded from the sites over 20 years is genuinely important as it is unique and is potentially part of commercial relationships. It would be a bugger if we lost all the experimental data from the industry group, but we would move on, the same for the database that has been built and the individual researchers’ results.”

The decreasing cost of storage and basic services for data management and access will allow for significantly larger proportions of data to be preserved. This will influence value decisions about what should be kept.

> "Data rates are enormous – we will not be able to archive all that will be produced."
>
> "A new staff member will be doing retinal imaging – there are huge space implications."
>
> "Storage will be a problem – we will have conformal microscopy images of 50MB each."
>
> "The 'mapper' will generate gigabytes per hour."

Once a project is completed, data is rarely cleaned up, quality controlled or metadata added – there is no time or money to do so and the focus is on the next project.

> "The researchers are massively resistant to completing metadata."
>
> "Hoarded data may not be written up for years, you get told 'it's not quality controlled yet.'"

The timing of data deposition, if there was to be somewhere to deposit it, was an issue for some researchers. Some research groups were comfortable with the concept of depositing data as a project proceeded; some indicated this was crucial to effectively capturing research data. The distinction between live access storage and long-term repository was clear in some but not all cases. Others felt that the deposit should only occur once a project had been completed.

> "Data more logically belongs with a repository when time has passed and it is no longer useful to the group."

There was a general view that changing practices and workflows today would be cost/labour effective for data collected from today. There is a significantly higher cost in recovering old data. However, there was an appreciation by a minority of those consulted of the potential value of existing non-digital data sets and a desire to be able to discover and access these.

Preserving data for the future is not part of the core business of researchers in the same way as intellectual property, copyright, publication, ethics, mentoring and training. There is no recognition of the preservation of data in the researchers' reward system.

There is a low-level 'market' view operating in data preservation with the premise that a way will be found for data of sufficient use and value to be preserved.

Data deposition

There is no evidence that there is any significant deposition of data beyond those disciplines where it is considered a norm or beyond those disciplines where the publication of an article is dependent upon lodgement of supporting data.

The existing compliance environment for the preservation of data was not seen as strong and the current requirements were typically not being observed. This is related to a number of factors for researchers – funding, time, culture, rewards and infrastructure.

> "There are format issues because they have the attitude that 'it's managed by me for me'; they don't have the time or inclination; they don't have resources to do it or they want someone to simply tell them what they have to do [for lodgement]."

There is no clear responsibility for the enforcement and auditing of compliance. This weakens the existing system and reduces the chances of success of a new system.

There is an interesting analogy for data deposition with the voluntary lodgement in repositories of e-prints and this is discussed in the section on Institutional Repositories. There is commonality in some of the reasons why lodgement should take place and also in some of the reasons why it does not.

Incentives

The survey found that researchers do recognise that data has a value and that there may be a value for data in the future that cannot be foreseen today. There are few incentives for researchers to manage data for the future.

> "You need to make it worthwhile for scientists if you want them to do it."
>
> "It's a very competitive field; time is a problem; producing papers is critical if you want funding for future work."

Researchers who manage data for the future know there may be a benefit for future researchers and hence for society.

Researchers have a view that exposed data allows a level of certification for the published research conclusions.

Disincentives

It was clear from the survey that there are no rewards for researchers for good data management in terms of their intrinsic or extrinsic motivations. Researchers are interested in having and retaining a job. They are interested in solving problems and in making a difference, in being recognised as an expert in a field and as being good at research. They want to conduct the maximum amount of research for the least amount of system work, and need funding to continue to achieve these rewards.

Researchers receive no funding for long-term data management and have little time for it. They receive no peer recognition or career advancement for managing their data and do not receive credits for data management in the metrics used to measure success. Nor do they receive recognition if someone else uses their data. There are no rewards from within their organisations for data management and rather than undertaking data management, deans and group leaders expect their staff to get on with teaching and research.

> "Recognition is more important than a framework – you would still need to resource it but the issue would largely go away if researchers received recognition for managing their data."

Managing data well for the future currently means that researchers would have to be prepared to spend less time on research and accept the career risks associated with this.

2.2 UNIVERSITIES

The university support structure

The survey found that there are a number of structures within universities that are, or could be, involved in efforts to preserve research data collections.

Units within universities with an interest in the preservation of data include the research office, which administers the grants received from funding bodies and is involved in the submission of both progress and final reports for the administered projects.

University IT units are usually concerned with supplying infrastructure and a common level of service across the campus, although areas requiring a variation to what is commonly supplied usually meet the need themselves. IT units are not usually involved in the long-term storage of data or in making data accessible.

An interesting insight into the place and meaning of data management for IT managers is provided by the UK-based University Colleges Information Systems Association. The Association surveys its membership to determine those IT issues that are of most concern. In 2004-2005 'Data management' ranked tenth out of 11 issues.[8]

> "We keep data for seven years and for ten plus for some of the commercial stuff, but IT don't want the volume we generate nor do they want to keep it for that long."

Some universities run mass storage services. These services provide large-scale storage for units and projects within the university. The emphasis is on size and whether the data is on line or near line. The services typically do not extend to accessibility or to many of the activities associated with data management and curation. There is an emphasis on the technical aspects of the activities undertaken.

> "We talked with the mass storage group but they were not interested in sustainability and long-term preservation."

Universities run supercomputing or research computing facilities. The brief of these services is typically to support computing activities in teaching and research that cannot be provided by the central IT services or by faculty or department IT staff. These services may run to additional computing power, grid computing, development systems, visualisation facilities and the provision of advice on the best use of these services. These services may include advice on storage and they may be linked to central IT services and mass storage systems. The emphasis is on technical solutions and computing solutions for complex or especially demanding activities. These do not include long-term storage, data preservation and curation.

University records management units are responsible for the preservation of university records, including the capture, use, storage and disposal of the records. The records in questions have typically been the records of the university as a business, so corporate records are available for consultation in the future. This is a legislative requirement. The task has changed in what is now a digital age and records units face particular problems with the amount of business that is undertaken through email. Records units have not traditionally seen research outcomes as university records they capture. Instead they have provided advice to the university community on the retention and disposal of research-related records.

University archives are usually well integrated with university records units and share the legislative basis for their activities. The archive remit is broader than that of the records units. They seek to retain more than just the official records relating to the history and mission of the organisation. Digital records are still new to the archives units of universities and where they have been encountered, the archives units have sought the cooperation of other parts of the university.

University libraries have long had the mission of acquiring, organising and making appropriate teaching and research materials available. These materials have included a range of formats and increasingly include digital collections. They have not included, to date, data collections other than those that have been organised as coherent collections and published, such as statistical data collections.

Institutional repositories are relatively new in the university landscape. They are responsible for the acquisition and management of digital collections. To date they have focussed on textual material, with some collections of images and/or sound. There is a strong emphasis on scholarly outputs in their textual collections and this is linked in some cases to current agendas for open access publishing, RQF assessment and Australian Universities Quality Agency (AUQA) exercises. Institutional repositories are also thought of as a way of showcasing the intellectual output of the University.

> "In this area you've got to put resources into developing the actual facilities, into having the support staff, but also educating the academic staff and having real research projects to work on with outcomes and impact."
>
> "The industry has approached us about taking their data. It is a great opportunity but we can't embark on it with ad-hoc processes or an agreement with the Centre – it would have to be with the University – we don't know how to handle this with the structure that is around."

There is potential for organisational conflict within the university and therefore a need to look at where data best belongs in the university structure for policy, funding and administration. It is important that this be done in light of national (and in some cases international) efforts to make data available for the future. There could be a data management group in Australian universities in the same way as there are audit, risk management and records groups.

University policies

University policies for the treatment of research data are shaped by the funding rules of the granting bodies and to a lesser extent by the relevant legislation. The activities of university records, archive units and the legislation are not connected. The strongest guide for policies would appear to be the *Joint NHMRC/AVCC Statement and Guidelines on Research Practice*, which is concerned with research data and records management, supervision of students and research trainees, publication and dissemination of research findings, authorship, peer review, conflicts of interest, collaborative research and research misconduct.

Universities have clear policies on all of the above issues which are core business for a university engaged in research activities. The policies universities use to address these issues form part of their risk management strategy: funding, legal action and reputation are all at stake if they are not followed.

Data management policies, however, are less evolved than these other areas of research. There are no rewards for universities to manage research data well, nor significant consequences for failing to do so.

The NHMRC/AVCC Statement is being revised and will be issued as the *Australian Code for the Responsible Conduct of Research*[9].

The response from senior university information officers

Deputy Vice-Chancellors, Pro Vice-Chancellors and University Librarians from selected universities were interviewed as part of AERES. They were asked about the development of their institutional repositories, as well as the current and the likely future role for the repository. They were also asked about the status of planning and policy development for the management of research data across the university and to identify the challenges in developing capacity in this area.

Among those interviewed, there was a strong awareness of research data management as an emerging issue in e-Research, of the capacity development being undertaken by NCRIS and their recommended 'Platforms for collaboration'.

Institutional repositories were, on the whole, seen as belonging with the university's information portfolio rather than with the research portfolio. Library and IT staff have a closer engagement with the management and development of repositories and a greater awareness of repository strengths and potential. Research staff, by contrast, are focussed on grant success and RQF. The linking of repositories and research offices for the RQF would appear to be a rare overlap between the information and research portfolios for many universities. These portfolios need to be better integrated to address the preservation of research data at the institutional level.

> "It belongs with the library, in partnership with the research portfolio. How successfully the library does it and how much backing it gets from the research office will work differently from place to place."
>
> "I only see us [library] as being able to take on the small to medium size data collections. Big collections like astronomy and so forth have got to be handled somewhere else and that's what I see the NCRIS type facilities hopefully taking on."
>
> "We are designing our ePrint service to integrate with our research administration system…to capture once and use many times."

Most of the universities surveyed recognised the need to look at the issue of managing research data and most were in the process of setting up committees to investigate the broader issues associated with e-Research. This has not yet resulted in a coordinated institutional response in most of the universities but there was a wide understanding of the need for action. The larger, research-focussed universities show a much greater awareness of the importance and complexity of data management issues than those universities without the same research involvement.

> "The data issue will be discussed at the eResearch group and they will report to the Research Committee."
>
> "Nobody is focussing on it here given the structural issues we are going through."
>
> "You need some mechanism to ask the researcher what is going to happen to the data."

The major engagements by universities with e-Research are supercomputing, and through research groups, involvement with the NCRIS agenda. There is some order to this in that these activities are seen as key enablers of current and future research, requiring the integration of research agendas and national research priorities. This reflects the researchers' own practices of looking to the next project, rather than thinking about how today's activities might assist projects in the future.

Institutional repositories have, for the most part, developed from the open access movement, where libraries in particular have attempted to provide alternative means of access to a university's published research output, to publicise research and provide access to articles only available through expensive subscription journals. More recently repositories have become linked with RQF and AUQA agendas.

> "For data most of the cost is in staff and their knowledge – you have to preposition the repository and work with communities that value data. The data has to be secure, accessible, access controlled by academics, owned by and if necessary returned to the academics. The repository has to be stable – no test beds, no beta versions, and must be easy to contribute to by academics."
>
> "Data storage/retention is patchy and still evolving. Sustainability is ultimately with the community to decide."

The complexity of the response required by universities to the data challenge was well illustrated by the APSR project at the University of Melbourne. This project, associated with the current AERES survey, demonstrated the effectiveness of establishing a team of Library and ICT staff as a consultation model for working with diverse data-intensive research groups. The team was able to profile their data collections, practices and needs in great detail and the project has informed the development of an eResearch strategy at the University. The *Sustainable Paths for Data Intensive Research Communities at the University of Melbourne*[10] report notes the need for an institution wide strategy for eResearch, including governance, guidelines, policies, and infrastructure. It notes that lack of governance and policy leads to duplication of effort, 'reinvention of the wheel' and poor use of rich skill sets within the University. It also notes the need to establish an information exchange strategy around eResearch as well as a registry of eResearch expertise within the University.

3. THE INFRASTRUCTURE

3.1 INSTITUTIONAL REPOSITORIES

The NCRIS Strategic Roadmap noted that, "*Repositories have the potential to move beyond the traditional approaches, e.g. just for storing publications, to support innovative new forms of research data, collections and research output*."[11]

The advent of open source and proprietary softwares since 2000 has enabled universities to develop web-based repositories, and most Australian universities now have one or more. Those without repositories are planning to implement them, whether in their own right or as part of a consortium. Repositories offer a means of actively participating in the scholarly communication debate that favours open access, where the leverage of the publishers of scholarly journals is reduced.

Recent debate has been focussed on the citation rate for the pre- or post-prints lodged in repositories and whether institutions with repositories should compel researchers to deposit. The record for voluntary deposition of pre-prints has so far not been good overall, although at least one university has had marked success with voluntary deposit. The reasons researchers are reluctant to deposit are primarily about work practices for researchers – lodgement is seen as too difficult, time consuming and without reward, the same reasons as are given for the non-lodgement of data.

Recent research has been focussed on the economic benefits of increasing exposure of scholarly output through deposition in repositories. Houghton and Sheehan note *"While it is impossible to calculate the quantum of benefits with certainty…simple estimates of the potential impacts of enhanced access on returns to R&D suggest that a move towards more open access may have substantial positive impacts"*.[12]

The survey authors found no use-case scenarios and business models available for institutional repositories. Data centres that have been established longer within the university structure can offer some business models.

Some universities have recognised the potential for extending the functionality of their repository. In some cases this has meant increasing the range of data formats held, particularly to image and sound formats, using the repository as a basis for the electronic publishing of books and journals, or meeting the wide-ranging need of e-Research.

Institutional repositories are part of the information landscape of most universities and are most commonly linked in that landscape to university libraries. This is logical whilst the repository content is largely textual but there are issues that need to be considered if the role of the repository is to include sustaining other formats. These issues include the technical issues associated with the wide range of research data formats and media, access to and use of the data, and the development of the skills needed to manage and support these. Some of these issues are being addressed by the work of the SII funded projects. Addressing these difficult issues will provide a policy and infrastructure framework.

> "We needed content management and the repository can't do this…it is fine as a stack but won't support our teaching needs. We chose this

> software because it allows the teaching staff to work the way they need to."

Researchers indicated that bodies responsible for research data have to be enduring. Repositories, if they are seen to need to exist, have to be treated in the same manner as other collecting organisations such as art galleries, archives, libraries and museums. Collecting organisations are there for the long haul and must be funded in a sustainable manner if they are to fulfil their mission. Failure to view repositories as long haul collecting agencies with enduring responsibilities will see a failure in the necessary policy development and critically, a failure to build trust with the research community.

> "What is the role of the institutional repository compared to advanced computing, particularly if the role of the institutional repository is not understood? We need to build expertise, as well as management, policies and procedures."
>
> "The trust comes from knowing they are competent and that long term archiving is their business…we'd be happy if the library was doing it because that's their job."
>
> "There have been a number of 'bee in bonnet' approaches in the Centre over the years from well intentioned amateurs…it is half baked and doesn't survive."

The notion of a repository being accredited and being trusted and certified is important to any system that asks or compels researchers to deposit research data. There has been considerable interest in the world of institutional repositories in recent years in building links between certification and deposition.

The Research Libraries Group (RLG) is a not-for-profit organisation of over 150 research libraries, archives, museums, and other cultural memory institutions. RLG was founded in 1974 by The New York Public Library and Columbia, Harvard, and Yale universities. Australian members include the National Library of Australia, the University of Melbourne and the University of Sydney.

The goal of RLG is, "*To increase online discovery and delivery of research resources, enable global resource sharing, and foster digital preservation for long-term access*."[13]

RLG and the United States National Archives and Records Administration (NARA) have developed a draft set of guidelines for certified repositories, *Audit Checklist for Certifying Digital Repositories*[14]. These are the most developed guidelines available but it is not yet clear how widely they will be adopted. These guidelines cover most aspects of certified and trusted repositories.

RLG states that producing certification requirements for establishing and selecting reliable digital information repositories is, *"…part of ongoing work with the Open Archival Information System model, and RLG and NARA intend the results to go into the standardization process through the International Organization of Standardization Archiving Series."*[15]

The RLG checklist is comprehensive, with sections on organisation; repository functions, processes and procedures; the designated community and intended uses of the information; and technology and technical infrastructure.

The checklist may be too comprehensive for a developing repository system. The notion of certification is more important than exact specifications. It is equally important that certification is adopted, publicised to researchers and reviewed and changed as appropriate. The key element of certification is that there is formal agreement between an agency with administrative responsibility for the management of data and enduring institutions.

The United Kingdom's Natural Environment Research Council (NERC) provides an example of certification through policy, rather than regulation. The *NERC Data Policy Handbook*[16] specifies the responsibilities of the Council's designated data centres and lists the Council's audit procedures.

3.2 DATA CENTRES

Data centres are formed to manage the data generated by a research unit. Data centres provide worthwhile lessons for those with an interest in the sustainability of, and long-term access to, research data.

A number of data centres and central IT service providers were contacted for their view of data practices and of policies linking their operations to research groups. These included the APAC National Facility, the Australian Social Sciences Data Centre (ASSDA), the Australian Antarctic Data Centre and CSIRO Marine and Atmospheric Research.

Data centre managers were aware of the need to be involved at the project planning stage to maximise the chances of preserving data, and making it accessible to others in the most efficient and cost effective way. They are also aware that this involvement requires the project planners to make an early decision on the likely value of the data to be generated or collected. This value may change as the project evolves. In any case there are difficulties in determining the future value of the data, particularly if it is considered that future use may be from outside the discipline in which the data was first generated.

Even in organisations with a deposition policy, data centres have experienced difficulties obtaining research data collections from research groups. These data collections were 'lost' if they were not managed.

> "Datasets that belong to the organisation are walking with the scientists who collected it when they leave."
>
> "Not everyone gives us their data – we have no mandate or policy framework which allows us to go and get it."

The earlier in the research project process data centres were consulted, the better the chance of having data deposited and having it in a form requiring minimum intervention. Metadata and standards are very important, as poor metadata reduces the chance of the data being discovered, and home grown standards minimise the chances of interoperability.

> "When we say we can't do it, it's because of the quality of the data and the quality, or lack, of metadata."

Some data centres encourage the deposition of data through a 'throw it over the fence' approach. This means there is no onus on the research groups to do anything other than pass on the data, regardless of the state of the data or the metadata. This method of data collection is not ideal as it is inefficient.

> "Project submissions should include detailed statements on data including down to the level of questionnaires; good quality data needs preparation; there is no point in storing rubbish data."
>
> "If you are involved in the grant phase you can talk to people about the life of a dataset and what they have to do."

Research groups in an organisation will take on data management to differing levels depending on its value to them and their resources. Resource-poor groups are worse at managing and using data. Organisational research groups sometimes ask data centres to write applications for them to use deposited data, and while this is a cost to data centres, it gives them a 'hook' to attract data collections.

> "Writing applications to use the data does help bring the data in, particularly as we have no policy framework to have the groups lodge the data…trying to get it built into the project cycle."

A data miner employed by a data centre is in demand from attached research groups. Again, while this is a cost, it offers the benefit of giving the data centre additional credibility with the users.

Staff in data centres are more likely to have a background in information or IT than in a specific subject discipline. Research projects often poach data centre staff after they gain disciplinary knowledge from working on a project, causing staff retention issues for data centres.

Data centres are geared to working with 'live' data rather than historical collections. There are rarely visited archaeological mounds in data centres. The cost of data rescue is significantly higher than that of taking in new data, making going back to rescue data a low priority.

Successful data management includes preservation, discoverability and accessibility. This is more likely to occur within an organisation with a unit whose business is to manage data. This is less successful when data management is undertaken by a single group or a cluster of groups, whose focus is on collection and analysis rather than long term sustainability. The profile and the role of the data centre in the organisation increases with its length of operation and the increased provision of services.

Data centres that are not well supported financially are at risk of not being seen to be able to provide services so are less likely to be used.

Australian Social Science Data Archive (ASSDA) and Australian Consortium for Social and Political Research Incorporated (ACSPRI) are mentioned in ARC and NHMRC funding agreements respectively. ASSDA was set up in 1981 with a brief to collect and preserve computer-readable data relating to social, political and economic affairs and to make the data available for further analysis.

In 2001, ASSDA was incorporated into the Australian Consortium for Social and Political Research Incorporated (ACSPRI) Centre for Social Research (ACSR), established through a joint initiative by the Australian National University (ANU) Research School of Social Sciences and the ACSPRI.

ASSDA collects data files from all parts of Australia, and from many different types of organisations, including universities, market research companies and government organisations. Since its establishment, ASSDA has collected over 1,050 datasets from Australian surveys and opinion polls. ASSDA also holds Australian population Census data and data from other countries within the Asia Pacific region. The uniqueness of ASSDA as a repository for machine-readable data makes it an attractive storage place for many important national surveys. Data stored in the archives can usually be made available for secondary analysis, depending on any access restrictions set by the depositor.

ASSDA currently acquires quantitative data and is working with the University of New South Wales and the University of Queensland, through a Linkage Infrastructure Equipment and Facilities (LIEF) grant, to develop a capability to work with qualitative data. ASSDA encourages the deposition of data and has developed a web page that lists reasons why researchers may be reluctant to lodge data and details why they should. ASSDA considers financial inducements or penalties are more effective than self-interest in promoting the lodgement of data.

ACSPRI members fund ADSSA. While both the ARC and NHMRC require data sets to be lodged with it, neither contributes any funding for its ongoing management.

Involvement with researchers at the grant development/application stage helps determine what can be done to make the acquisition of data easier. Poor metadata and unusual data formats are a greater problem than the volume of the data deposited.

ASSDA believes that researchers are not generally aware of the ARC and NHMRC requirements for data lodgement.

3.3 CSIRO

CSIRO is an Australian Government statutory authority constituted and operating under the provisions of the Science and Industry Research Act 1949.

The functions of CSIRO are to carry out scientific research and to encourage and facilitate the application and use of the results of its scientific research. CSIRO reported revenues of $925 million in 2005.[17]

CSIRO is examining the need for corporate data management across the enterprise. The management of research data is still at the divisional or lower level. This reflects the universities' approach where it is expected that those who have specialised needs are in the best position to determine how to meet those needs.

The data problems being faced by CSIRO are the same as those of the universities; space, volume, media, format and timing of deposit.

Policy making is a challenge due to the cultural issues associated with discipline- and project-focussed research groups across a large organisation, many of whom are working in collaborations outside the organisation.

There are opportunities within CSIRO to develop a repository framework, including working through their collaborations to institutional repositories. The CSIRO experience with supercomputing and its engagement with the grid means it is well placed to manipulate data across its network. The organisation also has considerable experience with geospatial, satellite-derived, and sensor-derived data.

CSIRO is an enduring institution and an icon for excellence in science, the largest research and development organisation in the country. There is a natural expectation that CSIRO will be very active in e-Research and that this will extend to providing data infrastructure to allow future researchers to benefit from today's research efforts.

CSIRO is in a more favourable position than the higher education sector to respond to its data stewardship responsibilities by developing and implementing divisional, if not enterprise wide, solutions. The solutions will be informed by the organisation's recent experiences in bringing together its administrative and corporate IT functions and from the bringing together of staff with information roles.

4. THE ENVIRONMENT

4.1 THE AUSTRALIAN RESEARCH ENVIRONMENT

Research and development in Australia is big business. Public funds were the source of approximately 45 per cent of the $12.25 billion expended on research and development in 2002-2003.[18]

The higher education sector expenditure on Research and Development was $4.3 billion in 2004. Ninety per cent of these funds came from government sources. General university funds were the source for $3 billion of these funds whilst Australian competitive research grants provided $740 million. Labour costs make up the largest component of higher education research and development expenditure.[19]

The major funding mechanisms for publicly funded higher education research in Australia, outside general university funds, are the ARC with the National Competitive Grants Program (NCGP) and the NHMRC. Significant blocks of funds are delivered by the Institutional Grants Scheme (IGS), the Research Infrastructure Block Grants (RIBG) Scheme, with performance based formulae and the SII.

In 2004-2005 the funds provided through these mechanisms for research or research support were:

ARC – NCGP (2004-2005)[20]	$481,406,000
NHMRC21	$479,125,000
IGS (2005)	$290,591,000
RIBG (2005)	$182,982,000
SII (2004)	$51,452,000

In addition to these figures, $202 million was allocated to Cooperative Research Centres (CRC) and $38.5 million to MNRFs in 2003-2004.[22]

A number of research-intensive universities account for the majority of the grant funds that go to higher education.

Higher education sector expenditure on data management can be conservatively estimated at $100 million per annum, based on 1 per cent of capital expenditure and 5 per cent of labour expenditure.

Funding agencies place clear requirements on researchers and institutions about how the funded research activities are to be managed and reported. There is also a detailed document prescribing standards for the conduct of publicly funded research. The *Joint NHMRC/AVCC Statement and Guidelines on Research Practice*[23] provides a comprehensive framework of minimum acceptable standards to guide institutions in developing their own procedures.

Australian research activity generates and uses large amounts of data. Much of this activity is focussed on specific national priorities by way of increased collaboration amongst researchers and through the development and better use of infrastructure. The e-Research agenda has a focus for collaboration and infrastructure and includes elements related to research data.

The focus on research priorities means that some of the data being generated by the national research activity can be reused by researchers in the domain in which it was generated. Some of this data may in turn be of use to future generations of researchers in the domain. It is also reasonable to assume that large datasets may be of use to researchers outside the domain in the future and that future data mining exercises may be able to harvest results from seemingly disparate datasets, provided they are sustained.

A new characteristic of Australian and international research is the development of e-Research agendas. The Department of Education Science and Training (DEST) web site defines e-Research as follows:

"The term 'e-Research' encapsulates research activities that use a spectrum of advanced ICT capabilities and embraces new research methodologies emerging from increasing access to:

- Broadband communications networks, research instruments and facilities, sensor networks and data repositories;
- Software and infrastructure services that enable secure connectivity and interoperability;
- Application tools that encompass discipline-specific tools and interaction tools."[24]

There are two new bodies of significance in the Australian e-Research setting. They are the e-Research Coordinating Committee and the NCRIS.

e-Research Coordinating Committee

The e-Research Coordinating Committee was appointed in April 2005 by the Minister for Communications, Information Technology and the Arts and the Minister for Education, Science and Training. The Committee serves primarily as an expert advisory group and consists of e-Research experts and key stakeholders, including representatives for the two Departments, the ARC, the Australian Vice-Chancellors' Committee (AVCC), the Council of Australian University Directors of Information Technology, the Council of Australian University Librarians, CSIRO, the National Academies Forum, NHMRC and National ICT Australia.

The objectives of the e-Research Coordinating Committee are to:

- Engage stakeholder groups in the identification of key policy issues and strategic directions in developing a national e-Research agenda;
- Recommend to the Australian Government an overarching strategic policy framework and implementation strategy.

The e-Research Coordinating Committee issued an interim report in September 2005, *An e-Research Strategic Framework*[25]. The interim report sets out the policy issues pertinent to Australia securing maximum benefit from the use of e-Research techniques; proposes strategic directions that should be pursued; and proposes further steps that would allow generation of an implementation plan.

The report lists access to data as an issue.

> *"Digital data in all its manifestations is now the core of modern research and knowledge generation in all fields. The volume and importance of such data is increasing rapidly through conversion of existing information in other forms and generation of new information through research. Stakeholders need to recognise the speed of this change and must adopt as quickly as possible best*

> *practice in digital data management and curation, standards and common practices, and security issues. These matters are fundamental to the provision of efficient and transparent access to, and dissemination of, knowledge and information."*

Chapter 5 of the report deals in detail with 'Better access to data'. It describes the essential issues of researchers' needs for data as; data management, curation, storage, security, access and standardisation. The nature and elements of data curation are described in detail.

> *"Data curation includes (but is not limited to):*
>
> - *data collection design for ease of retrieval and archival in compliance with data standards;*
> - *data collection standards;*
> - *data indexing, meta-tagging and cataloguing;*
> - *data certification, authentication and validation;*
> - *data quality assurance and data quality control;*
> - *management of data transmission and storage protocols;*
> - *digital rights, content and intellectual property management;*
> - *access control and security frameworks for data sets;*
> - *disposal schedules and procedures;*
> - *archival and preservation."*

The National Collaborative Research Infrastructure Strategy (NCRIS)

The Federal Government announced the NCRIS in the 2004-05 Budget as part of *Backing Australia's Ability: Building Our Future Through Science and Innovation*. It is an initiative that aims to provide researchers with access to the infrastructure and networks necessary to undertake world-class research.

NCRIS issued a *Strategic Roadmap*[26] in February 2006. It states:

> *"The purpose of the Strategic Roadmap is to inform decisions on where Australia should make strategic infrastructure investments to further develop its research capacity. It is intended to facilitate a coordinated approach to infrastructure investment across governments and agencies that:*
>
> - *Concentrates effort nationally on areas of greatest strategic impact;*
> - *Increases collaboration within the research system, and between it and the wider community; and*
> - *Reduces the duplication and sub-optimal use of resources arising from lack of co-ordination.*
>
> *In developing the Roadmap, the NCRIS Committee has drawn on expert advice and consultation with the research and wider communities. Development proceeded through several steps: consultation on an initial concept; more comprehensive scoping of the options; an expert advisory process; and further*

> *consultation on an exposure draft. 192 submissions were received on the exposure draft and considered in drafting this final version.*
>
> *The Roadmap provides a framework of capabilities, prioritised on the basis of the NCRIS principles, that represents the Committee's view as to where medium to large-scale research infrastructure investment should be focused over the next 10 years. It identifies the capabilities that Australia should strive to develop, rather than specific infrastructure, and also make some recommendations on the appropriate means to support them.*
>
> *More specifically, the Roadmap will provide a framework for the allocation of the NCRIS programme funding available from 2006-07 onwards."*

The Roadmap provides an overview of 16 priority capability areas that may be funded by NCRIS. One of these areas is 'Platforms for collaboration'. A section of the document dealing with the platforms for collaboration discusses data access and discovery, storage and management. The Roadmap also comments on data:

> *"In addition to new sets of data, some identified capabilities will depend for their utility and success upon curation of and access to large collections of existing information resources, in a variety of formats e.g. print publications, databases, sound recordings, images, (photographs, paintings, x-rays) and repositories of non-bibliographic information.*
>
> *Ideally, investment in platforms for collaboration should provide researchers with the ability to:*
>
> - *gain access to information relevant to their field from a variety of sources seamlessly;*
> - *exchange information collaboratively with colleagues; annotate their datasets or publications; and to*
> - *manage and disseminate the results of their research through supported repositories.*
>
> *Repositories have the potential to move beyond the traditional approaches, e.g. just for storing publications, to support innovative new forms of research data, collections and research output. Some possibilities include:*
>
> - *Life cycle management of research and research results;*
> - *Smart publications that link experiments, results and a range of documents that shorten and change the "publication cycle" (time to release new research);*
> - *The ability to validate not only research conclusions but also research results; and*
> - *The ability to allow other researchers access to original raw data – even for different purposes – or to provide stronger support for authenticity, authority and integrity of research.*
>
> *In order to manage research outputs, many elements need to be in place. These include: appropriate hardware and software (the technology); supporting workflows, policy and regulatory frameworks and administrative*

> *arrangements; and resources, especially staff resources. In addition, there are copyright and other legal considerations, together with technical standards issues, including sustainability, that need to be considered.*
>
> *In order to be exploited by search engines and data mining software tools much of the data, including experimental data, that will be exposed through the linkage of databases, needs to be annotated with relevant metadata providing information on provenance, content, conditions of use and so on."*

The *NCRIS Investment Framework*[27] was published in April 2006. This document describes the processes to be used to facilitate the development of investments relating to the capability areas. It notes that*: "Platforms for collaboration enable the research community to collect, share, analyse, store and retrieve information."* The document further notes that one of the components of infrastructure that supports collaboration is, *"Data storage management, access, discovery and curation to improve interaction and collaboration".*

It is clear from the statements of the e-Research Coordinating Committee and NCRIS that data preservation and access is recognised by government as being important to the current and future research endeavours of the nation. This recognition is equally clear in other countries with an interest in e-Research.

4.2 FUNDING AGREEMENTS AND FUNDING RULES

Australian Research Council (ARC)

The ARC provides research funds through grants for Discovery Projects, Linkage Projects, Centres and LIEF. Discovery Projects support three types of research to acquire new knowledge. *Pure basic research* is undertaken simply to acquire new knowledge. The findings of *strategic basic research* are directed into specified broad areas with the expectation of useful discoveries. *Applied research* is original work undertaken with a specific application in view.

The funding for Discovery Projects and Linkage Projects can range from $20,000 to $500,000 per annum and can cover a period of from one to five years. Centres of Excellence can expect funding at a level of between $1 and $3 million per annum for a period of five years.

LIEF provides funding for research infrastructure, equipment and facilities that will be used to support high-quality research projects. The minimum level of funding that will be provided by the ARC for a project is $100,000 per annum. Funding is normally for one year.

Despite these investments, the specifications and requirements for data management under the grants are uneven.

The Discovery Project funding rules lists areas of investigation not supported, including *"...compilation of data, unless this is an integral part of a project, in which case applicants must provide a statement indicating the research objectives to which the data would contribute."*[28]

Section 18 of the ARC Funding Agreement for Discovery Projects covers material produced under the agreement:

> *"18.1 The Organisation shall establish and comply with its own procedures and arrangements for the ownership of all Material produced as a result of any Project under this Agreement.*
>
> *18.2 For any Material produced under this Agreement, the Organisation shall ensure that all Specified Personnel:*
>
> > *(a) take reasonable care of, and safely store any data or specimens or samples collected during, or resulting from the conduct of the Project;*
> >
> > *(b) make arrangements acceptable to the ARC for lodgement with an appropriate museum or archive in Australia of data or specimens or samples collected during, or resulting from their Project; and*
> >
> > *(c) include details of the lodgement or reasons for non-lodgement in the Final Report for the Project."*

Under the Agreement, *"Material includes documents, equipment, software, goods, information and data stored by any means".*

Schedule C of the Funding Agreement for Discovery Projects lists "Research Special Conditions"[29]. Schedule C8 reads:

> "*Social Science Data Sets: Any digital data arising from a Project involving research relating to the social sciences should be lodged with the Australian Social Science Data Archive (ASSDA) for secondary use by other investigators. This should normally be done within two years of the conclusion of any fieldwork relating to the Project research. If a Chief Investigator is not intending to do so within the two-year period, s/he should include the reasons in the Project's Final Report*."

The ARC's pro forma for the Final Report does not list data as an item for reporting.

There are similar clauses to those quoted above in the Funding Agreements for Linkage, Centre and LIEF grants.

National Health and Medical Research Council (NHMRC)

All grants funded by the NHMRC are offered in accordance with the Deed of Agreement between the NHMRC and the Administering Institution. The funding includes a requirement for regular administrative activities and reporting.

Funding for Project Grants may be up to $300,000 per annum. The duration of the grants is normally three years.

The NHMRC Deed of Agreement[30] for NHMRC Researcher Support Schemes defines "Award Material" as

> "*…all material created, provided or required to be provided as part of, or for the purposes of the Award, and included (without limitation) any material derived from such material and any documents, equipment, information or data stored by any means*".

The Deed of Agreement specifies that funds cannot be used for any purpose other than funding the Award.

Section 2 of the Agreement deals with the administration of the award and includes requiring the Institution to:

> "*…ensure that any machine-readable data arising from an Award involving research relating to the social sciences should be lodged with the Australian Consortium for Social and Political Research Inc. (ACSPRI) or any other appropriate archive for secondary use by other investigators. This should normally be done within two years of the conclusion of any fieldwork elating to the Award research. If an Award recipient is not intending to do so within the two-year period, s/he should include the reasons in the Award's Final Report*."

A similar definition and stipulation is included in the Deed of Agreement for NHMRC Research Funding Schemes.

Both Agreements stipulate that ownership of Project Material and Award Material is with the institution with whom the agreement is made. The Agreements also note that the institution must ensure that research is conducted in accordance with principles outlined in NHMRC guidelines, including the *Joint NHMRC/AVCC Statement and Guidelines on Research Practice*.

Fisheries Research and Development Corporation (FRDC)

Several of the groups surveyed mentioned the FRDC's requirements for obtaining funds for research.

The Evaluation Criteria for the obtaining of funds notes that, *"In evaluating applications, the Board may also consider questions such as…Is there a strategy for managing data arising from the project so that it will be easily accessible by others in future?"*[31]

4.3 THE AUSTRALIAN CODE FOR THE RESPONSIBLE CONDUCT OF RESEARCH

The NHMRC, the ARC and the AVCC are revising the *Joint NHMRC/AVCC Statement and Guidelines on Research Practice (1997)*. The second stage of public consultation on the revised draft, now called the *Australian Code for the Responsible Conduct of Research - February 2006* has closed.

Joint NHMRC/AVCC Statement and Guidelines on Research Practice (1997)

Section 2 of the Statement deals with 'Data storage and retention'. Requirements in this section include recording data in a durable and appropriately referenced form, as well as the establishment by the department or research group of procedures for the retention of data and for the keeping of records of data held.

There are other requirements about how long data must be kept (at least five years), where it must be retained (the department or research unit whenever possible). It is also required that data related to publications be available for discussion with other researchers.

Australian Code for the Responsible Conduct of Research (February 2006)

The Code is in draft form and has been through two rounds of public consultation. It is a very significant document for Australian research as it requires institutions receiving funding to adhere to the code.

> "*In issuing this code, the ARC, the AVCC and the NHMRC assert that they expect high standards in the conduct of research in Australia, and specify how this should be achieved. All institutions that receive funding from the ARC or the NHMRC to support their research are required to adhere to this code.*"

The Code deals with important matters. It has sections on general principles of responsible research, research data and records management, supervision of students and research trainees, publication and dissemination of research findings, authorship, peer review, conflicts of interest, collaborative research and research misconduct.

Section 3 of the Code deals with 'Research data and records management'. The introduction to Section 3 outlines some of the difficulties in defining data and some of the reasons why it might be kept.

> *"The responsible conduct of research includes the proper handling and storage of research data… It is not possible to list comprehensively all the material that should be kept. In general, the researcher is responsible for this decision and should take into account:*
>
> - *the adequacy of the research data and records to justify the conclusions made, and to be useful to other researchers wishing to extend the research*
> - *the possibility of challenges to research data*
> - *whether the original material has heritage or other community value*."

The Introduction further notes:

> *"The research data retained should allow others to confirm that published findings are genuine, analysed appropriately, and not fabricated. In some cases, retention of the original material is required by law; retention of original material for other researchers to use is increasingly required by a funding agency or convention in the discipline."*

The section on managing the storage of research data and records note:

> "*Researchers have primary responsibility for the appropriate and secure management of research data, original material and records, according to the institution's policies."*

4.4 ARCHIVES AND RECORDS LEGISLATION

There is Commonwealth, State and Territory legislation in Australia governing the records of Commonwealth, State and Territory organisations, including universities. This legislation needs to be considered in any discussion of the preservation and/or disposal of research data.

The *Commonwealth Archives Act* (Archives Act 1983) deems that records which are the property of a Commonwealth institution (including both hard-copy and digital records) to be Commonwealth records. There is therefore a responsibility on the part of such institutions to comply with the requirements of the Act.

The Archives Act defines a record as "*…a document (including any written or printed material) or object (including a sound recording, coded storage device, magnetic tape or disc, microform, photograph, film, map, plan or model or a painting or other pictorial or graphic work) that is, or has been, kept by reason of any information or matter that it contains or can be obtained from it or by reason of its connection with any event, person, circumstance or thing.*"

State legislation exists governing the preservation of state records, which includes the records of state funded universities. This legislation may cover matters such as the preservation, disposal, access (including technology related matters such as format and media) and control of records. Typically the legislation goes to a statement such as, "A state authority is responsible for ensuring the safe custody and preservation of records in its possession."

The relevant legislation is:

NSW State Records Act 1998

Public Records Act 2002 (Queensland)

Public Records Act 1973 (Victoria)

State Records Act 1997 (South Australia)

State Records Act 2000 (Western Australia)

Archives Act 1983 (Tasmania)

Research data funded by Commonwealth and State Governments may fall under the jurisdiction of both State and Commonwealth legislation.

At least one university with significant research interests has developed, in cooperation with the appropriate records authority, a detailed disposal schedule. This schedule includes research data collections. Another university is investigating its responsibilities for data management under state records legislation.

4.5 COMPLIANCE

As noted above, the draft Australian Code for the Responsible Conduct of Research notes in its introduction:

> "*The research data retained should allow others to confirm that published findings are genuine, analysed appropriately, and not fabricated. In some cases, retention of the original material is required by law; retention of original material for other researchers to use is increasingly required by a funding agency or convention in the discipline*."

ARC and NHMRC Funding Agreements

There is little evidence that the requirements of the ARC and NHMRC Funding Agreements for external data deposition are complied with by the research community.

The ASSDA acquired eleven new datasets in 2004. There is no information available as to the number of datasets created.

It is possible to say that one proviso of the ARC Funding Agreements is being observed. Most researchers, *"…take reasonable care of, and safely store any data or specimens or samples collected during, or resulting from the conduct of the Project*." However, 'reasonable care' is not defined.

NHMRC Agreements stipulate that ownership of Project Material and Award Material is with the institution with whom the agreement is made. The Agreements also note that the institution must ensure that research is conducted in accordance with principles outlined in NHMRC guidelines, including the *Joint NHMRC/AVCC Statement and Guidelines on Research Practice.*

Data and its management and preservation are not reported on in the Annual Reports of the ARC or the NHMRC.

Joint NHMRC/AVCC Statement and Guidelines on Research Practice

The Guidelines state that, "*Data must be recorded in a durable and appropriately referenced form*", and "*The department or research unit must establish procedures for the retention of data and for the keeping of records of data held*".

The draft *Australian Code for the Conduct of Responsible Research* is more rigorous in its stipulation that institutions are responsible for policies and facilities for storage of research data, as well as durable records of the location of stored research data. Institutions should, *"Wherever possible, retain research data in the researchers' department(s), research unit(s), institutional repository or other multisite storage facility".*

There are no guidelines for what constitutes an appropriate repository.

The new code for the conduct of research will be a significant element in the framework for the retention of data. The code will be released after widespread consultation. It contains the requirement that institutions receiving funding from the ARC or the NHMRC must adhere to the code. It would be a welcome additional step to have greater clarity about how the code will be monitored.

Most people didn't know what the guidelines are for putting data into secondary repositories, let alone comply with it. It was never checked up on and most people didn't even know what it was that was being referred to.

Archives and Records Legislation

The Australian National Audit Office has conducted audits of organisations in recent years to establish their compliance with records legislation. Audit Office concerns have included records not being entered into the record-keeping systems, limited controls in place over electronic records, especially for those saved to shared network drives or personal workspaces, and lack of long-term sentencing programs for the disposal of records.

4.6 INTERNATIONAL DEVELOPMENTS

Australia's contribution to world scientific publication output for 1997 to 2001 has been calculated at around 3 per cent, while that of the United States, the United Kingdom and European Union are 35 per cent, 9 per cent and 37 per cent respectively.[32]

There are differences in the detail of the funding arrangements for these areas but publicly-funded research is common to them all. It is useful to look to their situation for comparison with Australia and as a learning opportunity.

United States

National Institutes of Health (NIH) provides an example of data sharing expectations and policy. The NIH promulgated a data sharing policy in 2003.

Under 'Goals of Data' the policy notes, "*Data sharing promotes many goals of the NIH research endeavor. It is particularly important for unique data that cannot be readily replicated. Data sharing allows scientists to expedite the translation of research results into knowledge, products, and procedures to improve human health.*"

The policy requires a plan for sharing data:

> *"In NIH's view, all data should be considered for data sharing. Data should be made as widely and freely available as possible while safeguarding the privacy of participants, and protecting confidential and proprietary data. To facilitate data sharing, investigators submitting a research application requesting $500,000 or more of direct costs in any single year to NIH on or after October 1, 2003 are expected to include a plan for sharing final research data for research purposes, or state why data sharing is not possible."*[33]

NIH FAQs on data sharing address the cost issue:

> *"NIH recognizes that it takes time and money to prepare data for sharing. You can request funds for data archiving and sharing as part of your grant application for collecting the data. If you have already collected the data, you may want to ask your NIH Project Officer about a competitive or administrative supplement. NIH recommends that you consider procedures and costs for data sharing during the application process rather than after the data have been collected."*

The National Science Board (NSB) provides insight into its role in the data management issue in the United States. The National Science Foundation (NSF) published *NSB-05-40, Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century*, in 2005. This comprehensive report makes recommendations on technical and financial issues and on policy matters related to these issues. The development process for the report included a series of consultative workshops. One of the workshop outcomes was the awareness that:

> "*The National Science Board and the National Science Foundation are uniquely positioned to take leadership roles in developing a comprehensive strategy for long-lived digital data collections and translating this strategy into a consistent policy framework to govern such collections.*"[34]

The Report concludes, "*…discussions should be designed to examine in both the national and global contexts how the investment, the policies, and inter-agency management can provide cost-effective, high-quality digital data collections. The need to address these issues is urgent. The opportunities are substantial*."

The NSF also developed the *NSF's Cyberinfrastructure Vision for 21st Century Discovery*[35] which elaborates Strategic Plans for: High Performance Computing; Data, Data Analysis and Visualization; Collaboratories, Observatories and Virtual Organizations; and Education and the Workforce.

The strategic plan for data, data analysis, and visualisation for the period 2006-2010 notes that, "*The enormous growth in the availability and utility of scientific data is increasing scholarly research productivity…*" and "*…U.S. international leadership in science and engineering will increasingly depend upon our ability to leverage this reservoir of scientific data captured in digital form….*"

One of the goals listed by the NSF's Vision document is, "*Support state-of-the-art innovation in data management and distribution systems, including digital libraries…*"

The Vision statement says the following about culture change:

> *"NSF's actions will promote a change in culture such that the collection and deposition of all appropriate digital data and associated metadata become a matter of routine for investigators in all fields. This change will be encouraged through an NSF-wide requirement for data management plans in all proposals. These plans will be considered in the merit review process, and will be actively monitored post-award."*

At a practical level the NSF, under its current Grant General Conditions, "…expects investigators to share with other researchers, at no more than incremental cost and within a reasonable time, the data, samples, physical collections and other supporting materials created or gathered in the course of the work."[36]

Developments by universities are reported on by Scott Carlson in *The Chronicle of Higher Education*. In an article entitled 'Dealing with data deluge' Carlson writes:

> "Dealing with the data deluge…will be among the great challenges for science in the 21st century. Many in the field say scientists should not be left to manage the data on their own. Instead, librarians will have to step forward to define, categorise and archive the voluminous and detailed streams of data generated in experiments. Already, librarians on some campuses – among them Purdue, the Johns Hopkins University, Maryland, and the University of California at San Diego – are beginning to take on that role."[37]

United Kingdom

Research Councils UK (RCUK) is a strategic partnership through which the UK's eight Research Councils work together to champion the research, training and innovation they support. The Councils are the main public investors in fundamental research in the UK with interests ranging from bio-medicine and particle physics to the environment, engineering and economic research. RCUK works alongside the Office of Science and Innovation to support the UK's academic researchers and to ensure the best investment of public money in research.

The Councils have published a position paper on access to research outputs.[38] The paper, *"reaffirms the Research Councils' commitment to the guiding principles that publicly funded research must be made available and accessible for public examination as rapidly as practical; published research outputs should be effectively peer-reviewed; this must be a cost effective use of public funds; and outputs must be preserved and remain accessible for future generations"*.[39]

The policies and practices of three of the Research Councils illustrate the place of research data in the national science endeavour.

The Medical Research Council (MRC) provides advice to MRC award applicants through its MRC Guidance on open and unrestricted access to published research. This includes a requirement that, *"…all applicants submitting funding proposals to the MRC are expected to include a statement explaining their strategy for data preservation and sharing.."*.[40]

The MRC policy on data sharing and preservation states that, *"Publicly-funded research data are a public good, produced in the public interest…should be openly available to the maximum extent possible".*

The following relevant extracts from the MRC Policy highlight the policy direction of the Council.[41]

> *"MRC expects that the valuable data arising from the research it supports will be made available to the scientific community to enable new research with as few restrictions as possible. Such data must be shared in a timely and responsible manner.*
>
> *MRC supports the view that those enabling sharing should receive full and appropriate recognition by funders, their academic institutions and new users for promoting secondary research.*
>
> *A limited, defined period of exclusive use of data for primary research is reasonable: different disciplines require different approaches, reflecting the nature and value of the data and the way they are generated and used.*
>
> *To enable effective sharing, data must be properly curated over its life-cycle and released with the appropriate high-quality metadata. This is the responsibility of the data owners who are usually those individuals or institutes that have received funding to create or collect the primary data."*

The National Environment Research Council (NERC) is another of the eight UK Research Councils that fund and manage scientific research and training in the UK. NERC uses a budget of about £220 million a year to fund scientific research in universities and at its own sites, covering the full range of atmospheric, earth, terrestrial and aquatic sciences.

NERC has developed a Science Information Strategy which, "...addresses the acquisition, management and dissemination of the data and information deriving from NERC-funded activities..." The Strategy includes a section on delivery and monitoring of progress of data deposition:

> *"Data centres will be provided with accurate information on data-generating awards made by NERC and NERC will ensure that a lead data centre is identified for each data-generating award made.*

> *NERC will actively enforce its Data Policy. Failure to meet policy obligations to manage data effectively and to provide copies of data to the data centre will result in final grant payments being withheld and future grant applications being withheld."*[42]

NERC's draft policy statement on data management rights and responsibilities makes it clear that researchers receiving public funds through NERC must manage data for the lifetime of the project and that in return for this NERC will manage the data for the long term.

The NERC Data Policy handbook[43], currently being reviewed, provides a comprehensive guide to the obligations and responsibilities of all NERC research participants with a relationship to data.

The Arts and Humanities Research Council (AHRC) supports research within a huge subject domain from traditional humanities subjects, such as history, modern languages and English literature, to the creative and performing arts.

The AHRC funds research and postgraduate study within the UK's higher education institutions. In addition, on behalf of the Higher Education Funding Council for England, it provides funding for museums, galleries and collections that are based in, or attached to, higher education institutions in England.

The Council has a partnership with the Arts and Humanities Data Service (AHDS) *"…to promote shared aims with regard to the application of information and communications technologies (ICT) in the arts and humanities."* This extends to a requirement for some awards for a data plan.

> *"The AHDS must be consulted within three months of the start of the proposed research to discuss and agree the form and extent of electronic materials to be deposited with the AHDS.*
>
> *The Research Organisation must ensure that any significant electronic resources or datasets created as a result of research funded by the Council, together with documentation, are offered for deposit at the AHDS within three months of the end of the project. Resources or datasets offered for deposit with the AHDS will be considered for deposit according to criteria designed to assess their fit within the AHDS collections.*
>
> *If the offer of deposit is not accepted by the AHDS or the Research Organisation is prevented from depositing for a specified reason it is the Research Organisation's responsibility to ensure that a waiver of deposit is agreed by the AHDS."*[44]

Organisation for Economic Cooperation and Development (OECD)

The OECD has produced a draft recommendation for publicly funded research data. It notes, "*Research policies, practices, support systems and cultural values all affect the nature of new discoveries, the rate at which they are made, and the degree to which they are made accessible and used*".[45]

The document comments, "*A lack of planning for and execution of the proper documentation and archiving of data sets is one of the key impediments to realising maximum value from the investment in research data…Attention should be paid to* *incentives, funding and the development of professional expertise in all areas of research data management.*"

The document makes the following key statement, highly relevant to the Australian situation, under the heading of 'Sustainability':

> *"Due consideration should be given to the sustainability of access to publicly funded research data as a key element of the research infrastructure. This means taking administrative responsibility for the measures to guarantee permanent access to data that have been determined to require long-term retention. This can be a difficult task, given that most research projects, and the public funding provided, have a limited duration, whereas ensuring access to the data produced is a long-term undertaking. Research funding agencies and research institutions therefore should consider the long-term preservation of data at the outset of each new project, and in particular, determine the most appropriate archival facilities for the data."*

5. THE ISSUES

Current data practices generally see data:

- managed sufficiently for research needs, but not professionally;
- discoverable through scholarly publication, but not otherwise;
- having a value placed on it for present needs, but not for the future;
- lost through commission and omission; and
- accessible only via approach to the author of the related publication.

The research community is producing more and more data and will continue to do so. An unknown amount of this data will have value for the future. Whilst researchers can place value on data collections for their domain it is less certain that they can predict how data might be used in the future, either inside or outside their domain.

The immediate problem is how to manage the data which should be preserved for the future without:

- creating more organisational bodies;
- creating more rules;
- keeping all data;
- spending less money on research; and
- doing less research.

How does the research sector achieve this in a way that is acceptable and workable for researchers?

Can the Australian research sector have research and data sustainability, instead of research or data sustainability?

Researchers

Researchers manage research data collections for the short-term. They are not funded to do otherwise, nor do they have time to do so. There is no recognition of the management and sustainability of data in the researchers' reward system.

Researchers and research groups managing data for the future may never be the beneficiaries of data sustainability.

Research groups change and may not endure. Individual researchers move from institution to institution.

The strengths of the current data management practices in Australia are:

- data resides with the individual or group to whom it is most valuable – data management efforts reflect a value decision by the researcher or group;
- data is discoverable through publication;
- data discovery leads to collaborations which provide tangible and intangible rewards;
- access to data via the researcher resolves many of the access and control concerns researchers have about data being misused; and
- researchers make the priority decision between data access and research.

The weaknesses of the system are that there are more and stronger disincentives for managing data than incentives, and that whilst there is an incentive to manage data for the life of a project there are no incentives for long-term management. Data discoverability is difficult under the present system. Data may be lost. Those making decisions about data sustainability are unlikely to see themselves as the beneficiaries of this activity. They will not be rewarded for data sustainability, and can see themselves as being disadvantaged by undertaking data sustainability.

Researchers receive a clear message that data sustainability is not important because no funds are provided to undertake it, because they receive no rewards or recognition for doing it, because they see no supporting infrastructure or services, and because there are unenforced/unaudited requirements for data deposition.

Institutions

Institutions, in general, do not see their job as being the management of research data for the future. They receive no funds for this activity nor do they provide facilities for the long term. They have no mandate or administrative framework for this activity other than as an add-on to the grant funding cycle.

Institutions are responsible for establishing and complying with their own procedures and arrangements for the ownership of data and for ensuring that researchers take care of and lodge data, as required by funding agreements. They have an implicit role in determining what research will be supported.

Institutions are the major recipients of public funds for the conduct of research and they are the major generators of research data. Their future researchers would be the major beneficiaries from the sustainability of research data.

Institutions are one of the few enduring features of the research landscape.

Researchers bring reputation and funds to the universities. Universities are keen to see a regulatory environment that maximises research outcomes, which are currently established by publishing and citation metrics. It is not in the interests of the university to impose irksome regulation that does not improve research outcomes.

Access management

A significant cultural impediment to data sharing is the ability to control access to data, by identifying users and uses in a trustworthy framework, and by data providers being able to stipulate specific and complex access requirements. Technology is being developed to allow this framework to be built. Organisations also need to develop appropriate security and identity policies and structures. This is generally seen as outside the scope of data management and sustainability, but it is a key enabler that needs to be addressed.

Funding agencies

The business of funding agencies is to advance research, administer grant programs and provide advice to government. Funding agencies are locked into the provision of short-term project funds. They have no brief for the development and sustaining of infrastructure. They are in the situation where they can encourage and mandate the sustainability of data but cannot supply or maintain any of the requisite infrastructure. It is not the business of the funding agencies to take on the issue of data sustainability.

Funding agencies are interested in maximising the research outcomes from the public dollar. Essential to this is giving researchers maximum freedom consonant with prudential regulation. A strength of the funding agencies is that they are strong enablers, not regulators.

Funding agencies are the primary source of funds for research. They are a logical mechanistic point from which to drive change. They are the quality audit point for grant applications and the logical quality audit point for providing assessment of data management plans, were grants and data management to be formally linked.

The ARC reported that approximately 25 per cent of Discovery Grant applications were successful in 2005.[46] The figure for NHMRC is approximately 20 per cent.[47] Allocation of funds to data sustainability would reduce these figures unless additional funds were allocated to the agencies.

Funding agencies, in one form or another, can be expected to endure.

Costs

Data sustainability will have significant costs, but will also enable significant research benefits. The costs of highly centralised data management can be estimated as some fraction of the funds being expended on labour and equipment. The costs of data management are currently being covered by grant and university funds, for the benefit of individual research groups. There is no administrative mechanism for bringing cost benefit analyses to this situation and comparing it with a funding stream that preserves and exposes data for broader use and public benefit. There does not appear to be a common mechanism for valuing data as an asset. The activities required for data stewardship can be costed and can be built into the research financial cycle.

How should we provide for future research from today's funds? Preliminary work on repositories indicates that there is a positive cost benefit to allowing open access to scholarly outputs. The nation funds research because it is of benefit to all. It is logical to extend the provision of funds to infrastructure that will benefit future generations and enable future research. Research is needed on the costs and benefits of data sustainability.

> "It's like museums – no glitz, glamour or votes but when you need data it is incredibly important to your work."

Capabilities

There are boundaries between research groups, data providers, repositories and data centres. These boundaries lead to duplication or capability gaps. It is important to identify responsibilities and opportunities across these groups where possible. Data management requires greater cooperation between the players. One of the problems is that there is no identified occupation or structure around which the issues that have to be decided can be discussed. This problem would be assisted if there were people and structures whose known job was data stewardship.

Organisations involved in significant preservation and access, like the National Library, have a mandate and a place within a government framework.

Infrastructure

There is no systemic sustainable infrastructure available to broadly support research data management, sustainability and curation. There are some existing domain-specific or technology-specific entities that provide opportunities, which are themselves on soft (short-term) money.

The available infrastructure is generally geared to short-term data storage not to long-term preservation. Infrastructure within universities that is geared to longer-term preservation, such as institutional repositories, currently tends to address scholarly output and formats such as text and images. It has the potential to evolve into a data management role.

Ease of use and trust are key issues for researchers in choosing to use infrastructure. There is little infrastructure with a track record.

Providing recognition of researcher effort in depositing data is seen as crucial. Any recognition would need to include the quality of the data (is it well-described?), the quality of the deposit (has everything relevant been provided?), and the quality of the repository where the deposit is occurring (is it sustainable, well-managed and accessible?). This leads to the concept of certification or accreditation for some of these elements.

Skills

Researchers and research groups are, for better or worse, self-reliant in their management of data. This situation does not support the sustainability and sharing of research data. There exists a cycle of learning which sees skills acquired for projects but not shared outside the project and not retained within the group as PhD students and researchers move.

Institutional repositories may be funded on soft monies. They are staffed by a small number of skilled specialists. The skills of these staff make them attractive to other employers. Some staff are employed on a contract basis because the funds to pay them are short-term. The number of staff and the basis of their employment are both threats to sustainability.

During 2005 and 2006, APSR has offered both formal training in select repository applications and informal training through conferences, symposiums and forums on the management and support of repositories. It has a unique position in this field and its training events have been heavily subscribed, indicating the great need for these services.

At the same time, there are no programs for training repository staff in existing tertiary courses for archives, library, information and ICT staff. The e-Research Coordinating Committee lists the nature and elements of data curation. Referral to this list makes it clear that specialised training is required if repositories are to develop in a sustainable manner and broaden their current role.

Policy and administration

The current policy framework for data management consists of a number of clauses in funding agreements. There is no administrative unit that has responsibility for data sustainability such that policies can be developed, cost benefit analyses undertaken, funding provided and the benefits of the system harvested.

Discipline Communities

Discipline-based communities need to be aware of their roles and responsibilities for data management and of the value of their data to other disciplines. The lack of community in some disciplines is a threat to sustainability.

6. THE FUTURE

Four systemic threats to sustainability for research data emerged from the survey results.

1. A lack of administrative responsibility for the task, which means there is:
   - a lack of mature and enforceable policy;
   - an absence of costing and cost-benefit information on which to make decisions;
   - an absence of 'ways and means' between data and infrastructure in research activities;
2. Strong disincentives for researchers to change their data practices;
3. Inconsistent engagement with the issue of long-term data management by researchers, universities and the funding agencies; and
4. A lack and/or poor use of data skills.

The survey results point to a number of possible scenarios that may be of assistance in responding to the identified issues.

6.1 SCENARIOS

A number of scenarios potentially address these systemic threats with the minimum disruption to the positive features of the existing national research effort.

The evolutionary marketplace

One scenario is to leave the problem to those who will benefit from solving it. Universities are in the best position to identify what strategic steps will maximise their position in a competitive environment. The existing policy framework can be left in place, allowing a market (needs) driven evolution of the repository movement, mass storage and data centres. The University of Melbourne project demonstrates the complexity of the needs to be met, as well as the richness of the skills that are available. A market driven response will develop innovative and appropriate solutions at the local level.

The funding agencies, with external assistance, can provide advice through the grant program to researchers on how to store their data in a manner such that discoverability is maximised, where researchers see that as desirable.

Information about the work undertaken by APSR, ARROW, DART, MAMS and RUBRIC can be referred to through the grants program. This could include statements on how these projects can assist research staff.

Researchers could be provided with information and guidelines about why research data should be preserved, including how current researchers have benefited from data having been preserved. Information about simple measures to make data better suited for preservation, such as metadata and where to go for assistance, would be available.

The regulated model (sticks)

Another scenario is to leave the existing system in place and allow a market (needs) driven evolution of the repository movement, but add a minimum regulatory change. This would see the addition of an emphasis on audit and compliance by the universities, including the provision of an annual data report to the funding agencies, describing their data management requirements, procedures and audit arrangements. The universities, as part of their responsibilities as an 'Eligible Organisation', would report on infrastructure developments that are available for data sustainability by its researchers or external researchers and on the data management skills (and development) within the university. The annual report would list collections, together with statements about their discoverability, access conditions, value and where appropriate, sustainability plans.

The annual reports will be used by the funding agencies to build an Australian research data collections database, a research infrastructure database and a research support skills database. These databases would be pushed to researchers, via the grant program, with the request that applications indicate what use they will make of the existing resources.

A coalition of repository, data centre and IT staff could be used to build the reporting framework and to assist the funding agencies in database development and maintenance.

The regulated model (sticks and carrots)

A variation to the sticks model would see funds awarded to the universities for data infrastructure on the basis of their reported need and demonstrated success in meeting that need. The more data collections that the universities preserve, the more a simple formula would provide funds to them. This would reward success in the same way that the IGS and the RIBG reward success.

It would be desirable to see a 'Research Data Sustainability Grant Scheme' as funds in addition to current sources. The sustainability of data is a new need, providing for the future and should be funded as such.

A National Research Data Stewardship Framework

The above scenarios will not address the systemic threats to sustainability. They could be considered as stopgap arrangements if cost benefit research indicated a more formal approach for preserving research data collections is required.

A more formal approach for the sustainability of data would need to address the threats to sustainability and to deliver benefits for future research. Such an approach needs to provide cohesion, coordination, collaboration and compliance. The key features of a national data stewardship landscape would be:

- a distinct administrative home for the task of data sustainability;
- the use of an existing layer of repositories and data centres for the provision of data storage and sustainability services for data no longer actively required by those who generated it;
- data reviewed for sustainability by the appropriate research community;
- a level of certification for the repository structure which allows clear understanding by all parties of the range and depth of services to be provided by individual repositories;
- the requirement that institutions receiving significant research grants develop data management plans which include certified repositories;
- a level of certification linked to research funds received – the more research funds received the higher the level of certification required;
- the linking of certification with the provision of advisory services by repositories about data management;

- the mandating of a data management plan for grant applications;
- the review of data management plans by grant review panels;
- the provision of funds in addition to the research funds, where appropriate, for data management;
- consideration of data creation and management as some part of research metrics for researchers;
- the addition to these metrics if data is used by others;
- the provision of funding to existing repositories by institutional grants, based on research quantum and later to be based on research data preserved, for use in skill development, maintenance and projection;
- ongoing external assessment of system costs and benefits; and
- an administrative apparatus that audits and amends the national system.

A national data stewardship framework requires acceptance of a new role for repositories and data centres by:

- Government, which would establish administrative responsibility for the task of data sustainability;
- universities, as the primary providers of the services within the system;
- researchers, who would use the system to lodge their research data;
- researchers, who would use the system looking for data;
- funding bodies, who would formally encourage and support researchers to use the system; and
- Government, which would provide enabling and developmental funds for the system through the funding agencies or other mechanisms.

It is relevant that NCRIS has a commitment to data storage management, access, discovery and curation through the 'Platforms for collaboration' plan. The NCRIS role will evolve in response to the investment plans for the research capability areas, together with the, *"…infrastructure requirements…identified by the e-Research Coordinating Committee and other bodies examining infrastructure needs"*.[48]

A formal system for data sustainability would overcome the threats to sustainability, providing administrative responsibility for the task of data sustainability and rewards for researchers and universities to engage with the issue. It would provide an audit point involvement by the funding agencies.

Skills

The issue of skills needs to be considered regardless of the scenario. There are significant skills in the research sector for data management and data manipulation, while there are fewer skills across the sector for data sustainability. The data management and manipulation skills need to be better utilised across the sector, while data sustainability skills need to be developed and the staff that acquire them need to be retained.

Universities can begin to address the organisational aspects of how to make best use of skills they already have. They need to link information and research portfolios around the data issue. This would be a model for the development of virtual groupings around data, such as geospatial, simulation and modelling, visualisation, sensor collected data and images.

The acquisition of skills could be investigated by DEST by undertaking discussions with groups such as the Australian Computer Society and Higher Education and VET bodies involved in IT training about the development of a register of competencies and training courses. Discussions with providers of ICT training have made it clear that courses can be provided, at different levels, for repository staff or others engaged in the tasks required for data curation. Courses should also be developed for the next generation of repository staff. The levels suggested were:

- elements within undergraduate degrees;
- graduate certificates or diplomas; and
- short term intensive bridging courses.

Courses developed should be designed such that they assist students and early career researchers in research groups, allowing them to focus better on research. Consideration needs to be given to the development by universities of graduate course awards in data management to reduce ‘on-the-job’ training and as a vehicle to introduce data sustainability as an issue.

The attracting, training and retaining of staff for repositories and data centres needs to be considered in light of sustainable funding and institutional policy settings. The skills requirements for data management should become clearer following the submission of investment plans for the NCRIS capability areas.

6.2 ACTIONS

A number of preliminary actions can be undertaken by the research sector to assist the development of data stewardship in Australia.

The ARC, CSIRO, DEST, the NHMRC and Pro Vice-Chancellors/Deputy Vice-Chancellors Research can:

- commission economic modelling for data sustainability, using cost-benefit analyses from existing repositories and data centres;
- ascertain from existing data centres and agencies what capacity they have for data sustainability, for discipline and/or data types;
- ask the National Archives of Australia to undertake an investigation, with State records authorities, of the status of research data as records;
- ask APAC to report on its capacity to undertake data management for data intensive research projects and its ability to provide support for large-scale science research;
- ask the National Library of Australia what role it might play as a repository for Humanities and Arts research data collections;
- ask ACSPRI to report on what is needed for ASSDA to continue and to develop, as well as what is required to ensure qualitative social science data can be preserved;
- investigate development of courses required for data stewardship; and
- investigate the experience of the FRDC and of UK funding agencies, where these agencies have requested data management plans.

NCRIS will be involved in research sector discussions of data stewardship. An element of the NCRIS platforms for collaboration investment plan may be the facilitation of administrative responsibility for the development of data stewardship in Australia.

Stewardship

There is a growing interest in the stewardship of the nation's research data. This stewardship is not the prime interest or core business of any of the players in the Australian research sector. It is unreasonable to expect the matter of building future benefit to be taken up by those engaged in or funding today's research. Nor does it make sense for those who might provide a service for the future, through repositories and data centres, to determine how a systemic solution should be developed.

The stewardship of research data can best be accomplished by making it someone's business. Concerned stakeholders, such as the ARC, CSIRO, DEST, NHMRC, the research community and the universities, will need to work as a coalition to assume the role of steward of Australia's research heritage.

> "For a lot of academics this is just too hard, they'll do what they have to do today and somebody else will have to worry about that bigger position."

Appendix

GROUPS & INDIVIDUALS CONSULTED

Researchers/Research Groups

Antarctic Climate and Ecosystems CRC
University of Tasmania
Professor Bruce Mapstone

Applied Economic and Social Research (HILDA)
University of Melbourne
Ms. Nicole Watson
Mr. Simon Freidin

Archaeological Computing Laboratory
University of Sydney
Mr. Andrew Wilson

Art History and Theory
University of Sydney
Mr. Tony Green

Astronomy
University of Queensland
Dr. Kevin Pimbblet

Australian Antarctic Data Centre
Australian Antarctic Division
Ms. Kim Finney

Australian Paediatric Surveillance Unit
University of Sydney
Professor Elizabeth Elliot
Dr. Yvonne Zurnyski

Australian Phenomics Facility
Dr. Edward Bertram
Mr. David Porter

Australian Public Sound Installation Archive
The Australia Centre
University of Melbourne
Dr. Ros Bandt
Mr. Iain Mott

Australian Social Sciences Data Archive
Ms. Sophie Holloway

Australian Virtual Observatory
University of Melbourne
Dr. Kathie Manson
Professor Rachel Webster
Dr. Randall Wayth

Bioinformatics – MMIM
Bio21 Institute
University of Melbourne
Dr. Marienne Hibbert
Mrs. Naomi Rafael

BlueNet
University of Tasmania
Ms. Kate Roberts

Capital Markets CRC
Professor Michael Aitken

Centre for Cross Cultural Research
Australian National University
Ms. Katie Hayne
Mr. Kim McKenzie

Centre for Health Informatics
University of New South Wales
Professor Johanna Westbrook

Centre for Mental Health Research
Australian National University
Ms. Trish Jacomb

Centre for Pensions and Superannuation
University of New South Wales
Professor John Evans

Civil Engineering
University of Queensland
Dr. Hubert Chanson

Classroom Education Research
University of Melbourne
Professor David Clarke
Mr. Cameron Mitchell

Cultural Informatics and Humanities Computing
Australian Science and Technology Heritage Centre
University of Melbourne
Mr. Gavan McCarthy

CSIRO
Mr. Ewan Perrin

Ecological Impacts of Coastal Cities
University of Sydney
Professor Tony Underwood

Environmental Sensing, Prediction and Reporting
CSIRO Land and Water

Dr. David Lemon

Dr. Stuart Minchin

Dr. Joel Rahmann

Faculty of Agriculture, Food and Natural Resources
University of Sydney

Dr. Dhia Al Bakri

Professor Les Copeland

Dr. Budiman Minasny

Dr. IAO Odeh

Faculty of Medicine IT
University of Sydney

Mr. Daniel Burn

Geoscience Australia
Department of Industry, Tourism and Resources

Mr. Frank Brassil

Mr. Paul Trezise

Dr. Lesley Wyborn

High Energy Physics
University of Melbourne

Associate Professor Martin Sevior

Dr. Glen Moloney

Dr. Marco La Rosa

Hydrology
University of Melbourne

Dr. Jeffrey Walker

Mr. Rodger Young

Information Policy and Practice Research Group
University of Sydney

Professor Sue Williams

Integrated Catchment Assessment and
Management Centre
Australian National University

Professor Tony Jakeman

Ms. Susan Cuddy

Intelligent Transport Systems
University of Queensland

Dr. Hussein Dia

Marine and Atmospheric Research
CSIRO

Mr. Paul Tildesley

Mechanical Engineering
University of Queensland

Dr. Michael Macrossan

Menzies Research Institute
University of Tasmania

Mr. Tim Albion

Professor Simon Foote

National Centre for Epidemiology and Population Health
Australian National University

Ms. Melissa Goodwin

Professor Terry Hull

Mr. Colin McCulloch

National Centre for Social and Economic Modelling
University of Canberra

Ms. Jeannie McLellan

National Centre in HIV Social Research
University of New South Wales

Ms. Maude Frances

Professor Sue Kippax

Neuroscience - MRI
University of Melbourne

Associate Professor Gary Egan

Dr. Neil Killeen

Pathology
University of Sydney

Professor Nicholas King

PARADISEC
University of Melbourne

Dr. Nick Thieberger

Associate Professor Steven Bird

Pharmacy
University of Queensland

Dr. Carl Kirkpatrick

Physiome - Kidneyome project
University of Melbourne

Professor Peter Harris

Dr. Andrew Lonie

Securities Industry Research Centre of Asia-Pacific (SIRCA)

Dr. Michael Briers

Supercomputing
University of Queensland

Professor Bernard Pailthorpe

Sustainable Minerals Institute
University of Queensland

Mr. Robin Evans

Dr. Guldidar Kizil

Professor Don McKee

Ms. Cathie Mortimer

Sydney University Biological Informatics & Technology Centre

University of Sydney

Dr. Lars Jermiin

Domain information

Dr. Paul Cooper

School of Botany and Zoology, ANU

Dr. Patrick De Deckker

School of Earth and Marine Sciences, ANU

Professor Adrienne Hardham

Research School of Biological Sciences, ANU

Professor Anthony Hill

Research School of Chemistry, ANU

Professor John Hutchinson

Mathematical Sciences Institute, ANU

Professor Brenton Lewis

Research School of Physical Sciences and Engineeering, ANU

Dr. Ian Morgan

Research School of Biological Sciences, ANU

Organisation information

Australian Academy of Science

Dr. Sophia Dimitriadis

Australian Partnership for Advanced Computing

Professor John O'Callaghan

Australian Research Council

Professor Elim Papadakis

Federation of Australian Scientific and Technological Societies

Dr. Ken Baldwin

National Archives of Australia

Mr. Stephen Pearson

National Collaborative Research Infrastructure Strategy

Dr. Mike Sargent

National Library of Australia

Dr. Warwick Cathro

The Australian Academy of the Humanities

Dr. John Byron

Dr. Kate Fullagar

Institutional data-infrastructure information

Australian National University

Professor Lawrence Cramm

Mr. Vic Elliott

Dr. Peter Raftos

Professor Robin Stanton

Flinders University

Mr. Bill Cations

Professor Ian Gibbins

Professor Chris Marlin

Griffith University

Ms. Janice Rickards

Queensland University of Technology

Ms. Paula Callan

Mr. Tom Cochrane

South Australian Partnership for Advanced Computing (SAPAC)

Professor Paul Coddington

Mr. Craig Hill

Professor Tony Williams

Swinburne University of Technology

Mr. Derek Whitehead

University of Adelaide

Mr. Ray Choate

Ms. Patricia Scott

Mr. Stephen Thomas

University of Canberra

Ms. Anita Crotty

University of Melbourne

Ms Linda O'Brien

Dr Glenn Swafford

Dr. Angela Bridgland

University of New South Wales

Mr. Andrew Wells

University of Queensland

Mr. Andrew Bennett

Ms. Belinda Weaver

Mr. Keith Webster

University of South Australia

Ms. Helen Livingstone

Mr. Stephen Parnell

Ms. Jenny Quilliam

University of Sydney

Mr. Ross Coleman

Mr. John Shipp

University of Technology Sydney

Ms. Fides Datu Lawton

REFERENCES

[1] National Health and Medical Research Council (1997) Joint NHMRC/AVCC Statement and Guidelines on Research Practice, [Online], 18/08/06, Available at:
http://www7.health.gov.au/nhmrc/funding/policy/researchprac.htm

[2] Department of Education, Science and Training (2006), Backing Australia's Ability, [Online], Available at:
http://backingaus.innovation.gov.au/

[3] Australian Partnership for Sustainable Repositories (2005), APSR Sustainability Issues Discussion Paper, [Online], 18/08/06, Available at:
http://www.apsr.edu.au/documents/APSR_Sustainability_Issues_Paper.pdf

[4] Australian Partnership for Sustainable Repositories (2005), Survey of Data Collections, [Online], 18/08/06, Available at:
http://www.apsr.edu.au/publications/data_collections.htm

[5] National Health and Medical Research Council (2006), Australian code for the responsible conduct of research, [Online], 18/08/06, Available at:
http://www.nhmrc.gov.au/funding/policy/code.htm

[6] National Collaborative Research Infrastructure Strategy (2006), Strategic Roadmap, [Online], 18/09/06, Available at:
http://www.dest.gov.au/sectors/research_sector/policies_issues_reviews/key_issues/ncris/

[7] An e-Research Strategic Framework: Interim Report of the e-Research Coordinating Committee (2005), [Online], 18/09/06, Available at:
http://www.dest.gov.au/NR/rdonlyres/B6F765A7-DD2C-432B-9064-2F9CD4E17E66/10518/InterimReport2.doc

[8] UCISA (2005), UCISA top concerns 2004/2005, [Online], 18/08/06, Available at:
http://www.ucisa.ac.uk/activities/surveys/tc/2005

[10] APSR, (2006), Sustainable Paths for Data Intensive Research Communities at the University of Melbourne, [Online], Forthcoming to:
http://www.apsr.edu.au

[11] National Collaborative Research Infrastructure Strategy (2006), Strategic Roadmap, [Online], 18/08/06, Available at:
http://www.dest.gov.au/sectors/research_sector/policies_issues_reviews/key_issues/ncris/

[12] Houghton, John and Sheehan, Peter (2006), The economic impact of enhanced access to research findings, CSES Working Paper 23, [Online], 18/08/06, Available at:
http://www.cfses.com/documents/wp23.pdf

[13] Research Libraries Group (2006), Mission and Goals, [Online], 18/08/06, Available at:
http://www.rlg.org/en/page.php?Page_ID=362

[14] Research Libraries Group (2006), Audit Checklist for Certifying Digital Repositories, [Online], 18/08/06, Available at:
http://www.rlg.org/en/page.php?Page_ID=20769

[15] Research Libraries Group (2006), Digital Repository Certification Task Force, [Online], 18/08/06, Available at:
http://www.rlg.org/en/page.php?Page_ID=367

[16] Natural Environment Research Council 2006, NERC Data Policy Handbook, [Online], 18/08/06, Available at:
http://www.nerc.ac.uk/data/documents/datahandbook.pdf

[17] CSIRO (2005) Annual Report 2004-05

[18] Australian Bureau of Statistics (2004) 8112.0 – Research and Experimental Development, All Sector Summary, Australia, 2002-03

[19] Australian Bureau of Statistics (2006) 8111.0 – Research and Experimental Development, Higher Education Organisations

[20] Australian Research Council (2005) Annual Report 2004-2005

[21] National Health and Medical Research Council (2005) Annual Report 2004

[22] Australian Research Council (2005) Annual Report 2004-05

[23] National Health and Medical Research Council (1997), Joint NHMRC/AVCC Statement and Guidelines on Research Practice, [Online], 18/08/06, Available at:
http://www7.health.gov.au/nhmrc/funding/policy/researchprac.htm

[24] Department of Education, Science and Training (2005), e-Research, [Online], 18/08/06, Available at:
http://www.dest.gov.au/sectors/research_sector/policies_issues_reviews/key_issues/e_research_consult/

[25] Department of Education, Science and Training (2005), Interim Report, [Online], 18/08/06, Available at:
http://www.dest.gov.au/sectors/research_sector/policies_issues_reviews/key_issues/e_research_consult/interim_report.htm

[26] National Collaborative Research Infrastructure Strategy (2006), Strategic Roadmap, [Online], 18/08/06, Available at:
http://www.dest.gov.au/sectors/research_sector/policies_issues_reviews/key_issues/ncris/

[27] National Collaborative Research Infrastructure Strategy (2006), Investment Framework, [Online], 18/08/06, Available at:
http://www.dest.gov.au/sectors/research_sector/policies_issues_reviews/key_issues/ncris/

[28] Australian Research Council (2005) Discovery, [Online], 18/08/06, Available at:
http://www.arc.gov.au/funded_grants/management_fund_discov.htm

[29] Australian Research Council (2005) Discovery, [Online], 18/08/06, Available at:
http://www.arc.gov.au/funded_grants/management_fund_discov.htm

[30] National Health and Medical Research Council (2005), Deeds of Agreement and Conditions of Award, [Online], 18/08/06, Available at:
http://www7.health.gov.au/nhmrc/funding/funded/manage/policy/deeds.htm

[31] Fisheries Research and Development Corporation 2006, Evaluation criteria, [Online], 18/08/06, Available at:
http://www.frdc.com.au/research/applicants/evaluation.php

[32] King, David A. (2004), The scientific impact of nations, Nature, 15 July 2004

[33] National Institutes of Health 2003, NIH Data Sharing Policy and Implementation Guidance, [Online], 18/08/06, Available at:

http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm#archive

[34] National Science Board 2006, NSB-05-40, Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century, [Online], 18/08/06, Available at: http://www.nsf.gov/pubs/2005/nsb0540/start.jsp

[35] National Science Foundation (2006), NSF's Cyberinfrastructure Vision for the 21st Century, [Online], 18/08/06, Available at:
http://www.nsf.gov/od/oci/ci_v5.pdf

[36] National Science Foundation (2006), Grant General Conditions (GC-1), [Online], 18/08/06, Available at:
http://www.nsf.gov/awards/managing/general_conditions.jsp

[37] Carlson, Scott (2006), Dealing with the data deluge, The Chronicle of Higher Education, v 52, 42

[38] Research Councils UK 2006, Research Councils UK' updated position statement on access to research outputs, [Online], 18/08/06, Available at:
http://www.rcuk.ac.uk/access/2006statement.pdf

[39] Research Councils UK 2006, News release 28 June 2006, [Online], 18/08/06, Available at:
http://www.rcuk.ac.uk/press/20060628openaccess.asp

[40] Medical Research Council 2006, MRC Open access guidance, [Online], 18/08/06, Available at:
http://www.mrc.ac.uk/index/public-interest/public-consultation/open_access-2/open_access.htm

[41] Medical Research Council 2006, MRC statement on data sharing and preservation policy, [Online], 18/08/06, Available at:
http://www.mrc.ac.uk/strategy-data_sharing_policy.htm

[42] National Environment Research Council, NERC Science Information Strategy v10.1

[43] Natural Environment Research Council 2006, NERC Data Policy handbook, [Online], 18/08/06, Available at:
http://nerc.ac.uk/data/documents/datahandbook.pdf

[44] Arts and Humanities Research Council 2006, Research Grants, [Online], 18/08/06, Available at:
http://www.ahrc.ac.uk/ahrb/website/apply/research/research_grants.asp

[45] OECD 2006, Draft OECD recommendation concerning access to research data from public funding, [Online], 18/08/06, Available at: http://www7.nationalacademies.org/usnc-codata/OECD_Principles_and_Guidelines.pdf

[46] ARC 2006, *Discovery Projects* Selection Report for funding Commencing in 2006, [Online], 18/08/06, Available at:
http://www.arc.gov.au/funded_grants/DP06_SelectionReport.htm

[47] NHMRC 2006, NHMRC Project Grant Funding Commencing in 2006 – Statistics, [Online], 18/08/06, Available at:
http://www.nhmrc.gov.au/publications/_files/projgr06.pdf - search="NHMRC Project Grant Funding Commencing in 2006"

[48] National Collaborative Research Infrastructure Strategy (2006), Investment Framework, [Online], 18/08/06, Available at:
http://www.dest.gov.au/sectors/research_sector/policies_issues_reviews/key_issues/ncris/