



**FROM DATA TO WISDOM:
Pathways to Successful Data Management
for Australian Science**

**Working Group on Data for Science
Report to PMSEIC
December 2006**

A paper prepared by an independent Working Group for the Prime Minister's Science, Engineering and Innovation Council (PMSEIC). Its views are those of the Group, not necessarily those of the Australian Government.

Table of Contents

Terms of Reference	7
Working Group Membership	7
Overview	9
Recommendations	11
Terminology	15
Acronyms	19
Chapter 1: Setting the Scene	21
The Importance of Data to Science	21
Urgent Demand for Data for Today’s Complex Societal Problems	22
The Data Explosion	23
Data Access, Discovery and Linkage	24
Doing More with Data	24
Stewardship of Data	25
Incentives and Disincentives	25
International Activities	26
Case Studies of Interest	26
Chapter 2: The Emerging Research Information Environment	27
Making Data Available	27
ICT Infrastructure and Databases	28
Some Challenging Questions for Australia	30
What Role for Industry?	34
The Impact of Privacy on Research	34
Other Regulatory Factors	35
Data Analysis and Data Management Skills	36
Importance of International Research Collaborations	36
Case Studies of Interest	38
Chapter 3: The Way Forward	39
A. A National Strategic Framework for Scientific Data	40
B. The National Network of Digital Repositories	41
C. Data Management, Access, Sharing and Collaboration – Changing the Culture	44
D. Ensuring there are no Regulatory Impediments	47
E. Skills for Data Management	51
Comment on a National Centre for Data for Science	53
Chapter 4: Possible National Initiatives	55
Opportunities for Change	55
The Case for a National Drug Research and Policy Initiative (Pharmaco-epidemiology/ Pharmaco-vigilance)	55
The Case for a National Environmental Reporting Initiative	60
Appendix A: Case Studies	67
Case Study 1: Avian Influenza Gene Sequence Sharing	67
Case Study 2: Australian Ocean Data Centre Joint Facility	68
Case Study 3: PARADISEC – Pacific and Regional Archive for Digital Sources in Endangered Cultures	69
Case Study 4: Australian Digital Thesis Project	70
Case Study 5: Species 2000 and ITIS Catalogue of Life	72
Case Study 6: Biometric Technologies	75
Case Study 7: ICTVdB, Universal Virus Database built for the International Committee on Taxonomy of Viruses	76
Appendix B: Consultations	79
Appendix C: Australian and International Initiatives	81
References	91

Tables and Figures

Figure 1:	A diagram showing a scenario projecting an exponential growth in data.....	23
Figure 2:	Population health data linkage systems around the globe.....	56
Figure 3:	National data coverage across 263 indicators (SOE 2006)	60
Figure 4:	Indicator coverage (SOE 2006) in seven of eight themes.....	61
Figure 5:	WRON web services.....	63
Figure 6:	Progress with compiling the <i>Catalogue of Life</i>	74

Terms of Reference

The PMSEIC Working Group will:

1. Provide an overview of current approaches to the management of large amounts of scientific information and data for research;
2. Outline the issues surrounding the nature of data and data repositories and libraries, including access, storage, authentication, sustainability, protection and standards for interoperability;
3. Examine whether there are advantages in Australia for a single virtual repository or key multiple domain-specific repositories of scientific information covering all research institutions, both Universities and Government (and, if they wanted to, private companies);
4. Identify issues relating to the development of an Australian virtual repository and infrastructure mediating access to the repository;
5. Identify issues relating to industry participation in the development of, and access to, an Australian virtual repository or repositories;
6. Identify strategies that could be introduced to improve access to research undertaken by publicly funded research agencies (e.g. protocols around repositories), including international access and collaboration;
7. Identify a data management strategy to ensure Australia's scientific sector is globally competitive and provides benefits to the Australian economy, environment, and society;
8. Take into account the conclusions and recommendations of the e-Research Coordinating Committee.

Working Group Membership

- Professor Robin Batterham AO, Chief Technologist, Rio Tinto (Co-Chair)
- Professor Fiona Stanley AC, Director, Telethon Institute for Child Health Research and Executive Director of the Australian Research Alliance for Children and Youth (Co-Chair)
- Ms Rozanne Frost, Chief Information Officer, CSIRO
- Professor Ah Chung Tsoi, Director, e-Research Centre, Monash University
- Ms Kim Finney, Manager, Australian Antarctic Data Centre, Department of the Environment and Heritage (DEH)
- Dr Warwick Cathro, Assistant Director-General, Innovation, National Library of Australia
- Professor Ramamohanarao Kotagiri, Head of Computer Science and Software Engineering, University of Melbourne
- Dr Evan Arthur, Group Manager, Innovation and Research Systems Group, Department of Education, Science and Training (DEST)

The Working Group was supported by advisers from the Australian Bureau of Statistics (ABS) and the Department of Communications, Information Technology and the Arts (DCITA).

Overview

The PMSEIC Data for Science Working Group was brought together during 2006 to examine and advise on directions for managing the vast amounts of data now being generated from scientific research, observational projects, instruments, national and international collaborations, data mining and analysis.

There has been considerable investment by Commonwealth and State/Territory governments in science (e.g. Backing Australia's Ability; Health and Medical Research funding; National Collaborative Research Infrastructure Strategy (NCRIS), Centres of Excellence; Cooperative Research Centres, State Biotechnology and other investments) and this investment needs to be matched by a commitment to ensuring that data are regarded as a vital asset to be used and managed for the greatest economic and social benefit.

This report examines evolving issues and research practices and their implications for Australia and the future development of research infrastructure. It proposes a comprehensive agenda for change and in particular aims to address generic issues to ensure that Australia has a systemic approach to the major challenges and makes the most of our opportunities to enable the best use of data for science.

The recommendations are aimed mostly at governments at all levels, and their agencies which fund or produce data. But they are also relevant to the community of scientists and researchers in universities, institutions and centres, and to those in non-government and business environments who are also committed to Australia being more effective economically and socially at home in our region and as part of the global picture.

This report is timely as a variety of other national and international activities are occurring to address the issues identified in this report. It contains the findings and conclusions of Australian experts in the field of digital data management.

Our hope is that this report stimulates debate within governments, research funding agencies and research communities. It is intended to inform science policy thinking and move the agenda on from a focus on the science community to one which looks at the importance of data management to achieving national objectives.

Influences on the Working Group

The Working Group was influenced by other recent work which is generally aimed at better collaborations and more effective use of information nationally. In particular the Working Group was impressed with the work of the e-Research Coordinating Committee, the National Collaborative Research Infrastructure Strategy and the report of the Australian e-Research Sustainability Survey Project.

The Working Group was also influenced by a number of major international reports recently released on this topic. Those from the International Council for Science (ICSU) and the Organisation for Economic Cooperation and Development (OECD) have produced excellent recommendations similar to those in this report. These recognise the value of ensuring full and open access to data for scientific research and education purposes. In these reports the issues of data and information rescue, appropriate investment in data archiving and management, equitable access to both scientific data and publications and how intellectual property is managed are thoughtfully and fully examined. From these reports and from many written submissions, the Working Group was able to consider how to use data more creatively to best serve a research and development agenda.

Many other countries have had similar Working Groups to our own and we have also been guided by them. Like Australia, the United States, the United Kingdom, Canada, the European Union, and many countries in Asia are investing heavily in the development of national research infrastructures. While the bulk of overseas investments are building significant high performance

computing resources and advanced networks, they also identify end-to-end scientific data management, from data acquisition and integration, to data treatment, provenance and persistence as immediate and important challenges. How well Australia is performing with respect to the support required for this whole life cycle of data was an important aspect of our deliberations.

In addition to reading relevant reports from Australia and overseas, the Working Group has sought and received submissions and presentations from a broad group of national and some international experts, organisations and bodies (list at Appendix B: Consultations). Some case-studies of real life examples identified by the Working Group, consultants and from the submissions and which illustrate some of the problems and solutions are included in Appendix A.

Final comments

We hope that our recommendations support the other national activities and accurately represent the concerns and suggested solutions of those who contacted us. The challenge to us in making over-arching generic recommendations is that the scientific disciplines have different needs and we know there is considerable diversity in data management capacity and quality in different areas of science.

Many researchers, across a range of disciplines, are collaborating well and making major international contributions in important scientific areas, supported by grants and infrastructure with well-managed repositories, clear data management plans and allowing access to and use of data for relevant science and knowledge creation. Yet, in some areas of research, data management strategies are much more vulnerable or even non-existent.

We hope to change the national culture to ensure data are valued, kept, discovered, described, archived adequately, protected, shared and used again and again. We want to raise awareness about the advantages to science and the nation of changing our approaches to data and the dangers of not taking appropriate action now. The effort we have outlined will need adequate resources from government and the private sector and will need to be undertaken over the next five to seven years if we are to make any real headway.

Acknowledgements

The PMSEIC Working Group would like to acknowledge the valuable contributions by all consulted, especially Associate Professor Bob Beeton, Chair of the State of the Environment Committee, and Dennis Trewin, Australian Statistician.

Recommendations

A. A National Strategic Framework for Scientific Data

Recommendation 1:

That Australia's government, science, research and business communities establish a nationally supported long-term strategic framework for scientific data management, including guiding principles, policies, best practices and infrastructure.

Recommendation 2:

That a high-level expert committee be established to provide the leadership role in progressing the formation of the long-term strategic framework for scientific data management.

B. The National Network of Digital Repositories

Recommendation 3:

That the necessary policy and programmes be implemented with a view to establishing a sustainable publicly funded national network of federated digital repositories.

Recommendation 4:

That the expert committee consider the development of a strategic roadmap for the implementation and evolution of the national network of federated digital repositories.

C. Data Management, Access, Sharing and Collaboration – Changing the Culture

Recommendation 5:

That standards and standards-based technologies be adopted and that their use be widely promoted to ensure interoperability between data, metadata, and data management systems, providing authentic users of the data with appropriate processes and safeguards.

Recommendation 6:

That the principle of open equitable access to publicly-funded scientific data be adopted wherever possible and that this principle be taken into consideration in the development of data for science policy and programmes.

As part of this strategy, and to enable current and future data and information resources to be shared, mechanisms to enable the discovery of, and access to, data and information resources must be encouraged.

Recommendation 7:

That funding agencies offer incentives to encourage researchers and institutions to:

- ***develop data management plans for each research grant application involving data collection and generation, and that standards be made freely available and widely disseminated so as to encourage best practice in data management;***
- ***introduce policies and practices to encourage collaboration and sharing of data across Australia's scientific research institutions and across agencies; and***
- ***analyse and re-use existing data.***

D. Ensuring there are no Regulatory Impediments

Recommendation 8:

That funding agencies such as the NHMRC and ARC ensure that best practices and policies are developed and followed that allow bona-fide researchers to access individual population data, including the integration and linking of data from multiple sources, whilst protecting privacy, and ensuring that ethics committees fully understand these policies and their rationale.

Recommendation 9:

That in the context of developing the strategic framework for scientific data management, Australia's intellectual property approaches be checked to ensure they do not impede the sharing of data.

In particular, it should take into account the OECD Committee for Scientific and Technological Policy guidelines on access to research data and the International Council for Science statements about the benefits of sharing data.

E. Skills for Data Management

Recommendation 10:

That data management expertise becomes a core skill for researchers, including graduate and postgraduate science students across all disciplines, and that they receive data management training as part of their education.

Recommendation 11:

That the Australian Government give early consideration to the findings of the e-Research Coordinating Committee regarding changing research behaviour, practices and skills.

Terminology

Data

Data are defined (OED) as things known or assumed as facts; facts collected together for reference or information, reasoning or calculation.

For the purpose of this report our definition is broad and includes data from the social sciences and humanities as well as other scientific disciplines such as astrophysics, mathematics and biology, and information collected not just by scientists or researchers but by agencies for administrative purposes such as health, welfare, population, education, employment and crime.

Data exist in a variety of formats including any information that can be stored in digital form, such as text, numbers, images, audio, video, software, algorithms, equations, animations, and model simulations.

Data Archiving

A curation activity which ensures that data are properly selected and stored, can be accessed, that their logical and physical integrity are maintained over time, and that concerns about security and authenticity continue to be addressed and monitored.

Data Preservation

An activity within archiving in which specific items of data are maintained over time so that they can still be accessed and understood through changes in technology.

Data Grid

A grid computing system that creates a virtual collaborative environment through which many users in many places can access and share large quantities of distributed data, for their own sake, or to support large scale computational applications. Its functions may include data aggregation and mirroring, data search, query access, storage resource brokering and database management.

Data Management

Managing the storage and use of data from the time they are generated or collected, maintaining their integrity, security and useability, and ensuring that it can be discovered and re-used by others for as long as it is required. The term is taken to mean all of the actions needed to maintain data over their entire lifecycle and over time for current and future users. Data management encompasses both data archiving and data preservation.

Data Mining

Data mining is the field that includes machine learning and statistical inference techniques that can be applied to large volumes of data with the goal of finding interesting patterns which can be used for classification, clustering, prediction, explaining underlying processes, trends and causes of problems.

Digital Standards

Specified protocols, guidelines, criteria or procedures to ensure that data are useable by a community of users, or across a range of systems and applications. Standards have been defined and promulgated for many aspects of data, including logical structure, organisation and file formatting, data interchange, records management, storage media, indexing, meta-tagging (the incorporation of metadata) and ontological organisation, protocols for network transmission, security, privacy and database accessibility.

Digital Data

Information created in, or converted to, a digital format for storage, transmission, processing and/or logical control. Such information can include text, numbers, images, audio, video, indexing metadata, telecommunications signals, instrument telemetry, control and sensor signals, biometric information, software, algorithms, equations, animations, inputs and outputs for computer analysis, modelling or simulation.

Digitisation

The process of conversion or reproduction of 'analogue' or 'physical' data objects to digital data objects. In this process, it is critical that provenance data is also captured and recorded as metadata to accompany the new digital object, to enable subsequent indexing, discoverability, authentication and tracing back to its non-digital origin. While many steps of digitisation can now be automated, much of the effort required in digitisation relates to individual expert decisions entailed in the creation of useful metadata; to date, this activity is only partly automated.

eResearch

Research activities that use a spectrum of advanced ICT capabilities and embrace 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; and
- application tools that encompass discipline-specific tools and interaction tools.

Federated Repository

A cross-index of a variety of repositories maintained by discrete bodies or institutions, that may allow users to link, for example, to digital content, research data, research outputs such as theses and papers, or computer software and associated documentation in a variety of formats.

Within a federated repository, each individual repository may have been developed or established independently (possibly for different purposes). For the contents to be discoverable and accessible, the owners of the constituent repositories index and package the stored objects according to a mutually agreed set of standards, particularly regarding metadata.

This can also be referred to as a virtual repository.

Grid Computing

Aggregating computing resources and dynamically allocating processing tasks to multiple computers, data repositories or instruments across a networked infrastructure. Grid computing uses a set of open standards and protocols, to undertake major and otherwise computationally intensive tasks and to link incompatible or isolated resources.

Interoperability

The ability of different information technology systems and software applications to communicate, to exchange and integrate data accurately, effectively and consistently, and to use the information that has been exchanged.

Life Cycle of Data

The complete existence of data, from their creation or collection and initial storage, through maintenance and sustainability, to the time when they become obsolete and are deleted. The life cycle of data includes:

- data acquisition, collection and generation;
- data storage and management;
- standards to enable data to be used and interpreted; and
- access to data.

Metadata

Structured data that describe a data resource, analogous to cataloguing data held by libraries, museums and archives. Metadata aids classification, management, discovery, and use of data by people or by automated processes. Metadata may include data attributes such as type, structure, size, title, content, provenance, creation date, author or location.

Repository

A central place where data are stored and maintained. A repository can be a place where multiple databases or files are located for distribution over a network, or can be a location that is directly accessible to the user without having to travel across a network.

A digital repository is either a local, institutional, or central (e.g. subject-based or discipline-based) digital archive for depositing and providing access to digital contents.

Science

The systematic observation of natural events and conditions in order to discover facts about them and to formulate laws and principles based on these facts; the organised body of knowledge that is derived from such observations and that can be verified or tested by further investigation; any specific branch of this general body of knowledge.

Scientific research is not an isolated activity, but interacts strongly with other disciplines. Issues of relevance in management of data for science are relevant to many other fields that generate or use data.

Semantic Web

A project that intends to create a universal medium for information exchange by putting documents with computer-processable meaning (semantics) on the World Wide Web. Currently under the direction of the World Wide Web Consortium (w3C), the semantic web extends the web through the use of standards, markup languages and related processing tools.

Virtual Repository

Refer to *Federated Repository*.

Acronyms

ABS	Australian Bureau of Statistics
ADT	Australian Digital Thesis
ANU	Australian National University
AODCJF	Australian Ocean Data Centre Joint Facility
AONS	Automated Obsolescence Notification System
APSR	Australian Partnership for Sustainable Repositories
ARC	Australian Research Council
ASIBA	Australian Spatial Information Business Association
ASSDA	Australian Social Science Data Archive
ATCC	American Type Culture Collection
AVCC	Australian Vice-Chancellors' Committee
BAA	Backing Australia's Ability
CAUL	Council of Australian University Librarians
CSIRO	Commonwealth Scientific and Industrial Research Organisation
CSPR	Committee for Scientific Planning and Review
CSTP	Committee for Scientific and Technological Policy
DAFF	Department of Agriculture, Fisheries and Forestry
DCWG	Digital Content Working Group
DEH	Department of the Environment and Heritage
DEST	Department of Education, Science and Training
DLU	Data Linkage Unit
DOHA	Department of Health and Ageing
eRCC	e-Research Coordinating Committee
GP	General Practitioner
GOSC	Grid Operations Support Centre
GSD	Global Species Database
ICSU	International Council for Science
ICT	Information and Communications Technology
ICTV	International Committee on Taxonomy of Viruses
ICTVdB	International Committee on Taxonomy of Viruses Database
IP	Intellectual Property
ITIS	Integrated Taxonomic Information System
MNRF	Major National Research Facility
NASA	National Aeronautics and Space Administration
NBS	National Broadband Strategy
NBSIG	National Broadband Strategy Implementation Group
NCBI	National Center for Biotechnology Information
NGS	National Grid Service
NHMRC	National Health and Medical Research Council
NIH	National Institutes of Health
NCRIS	National Collaborative Research Infrastructure Strategy
NDN	National Data Network
NRM	Natural Resources Management
NRP	National Research Priority
NSB	National Science Board (US)
NSF	National Science Foundation
NSSDC	National Science Space Data Centre
OECD	Organisation for Economic Co-operation and Development
OED	Oxford English Dictionary
PBS	Pharmaceutical Benefits Scheme
PMSEIC	Prime Minister's Science, Engineering and Innovation Council
SciDIF	Scientific Data and Information Forum

SET	Science, Engineering and Technology
SII	Systemic Infrastructure Initiative
SKA	Square Kilometre Array
SOE	State of the Environment
UNESCO	United Nations Educational, Scientific and Cultural Organisation
USDA	United States Department of Agriculture
VIDE	Virus Identification Data Exchange
WADLS	Western Australian Data Linkage System
WADLU	Western Australian Data Linkage Unit
WHO	World Health Organisation
WRON	Water Resource Observation Network

Chapter 1: Setting the Scene

Chapter Overview

This Chapter introduces the problem addressed in this report: that the amount of scientific data being generated from multiple, often unrelated sources is vast, varied and needs to be managed in a way that provides for its ready discovery, easy access, long-term storage and most importantly its re-use. Analysis of these data, when brought together, can provide answers to some of the most complex problems we face today – such as how to manage our environmental resources.

The term “data for science” in this report applies to data and datasets of immediate and long-term value to a wide range of academic, industry, government and community users. This report is relevant particularly to the management of data used in any rigorous and methodical process of research, investigation, invention and innovation, generation of knowledge or scholarly analysis and documentation of information, but also to a wider range of creative endeavours and society activities that use, generate or archive data.

In the current world of scientific research, the collection, analysis and interpretation of large quantities of data are crucial components that drive and determine the success of research projects. Scientists in many fields now produce datasets some of which are made accessible, via high-speed network links, to colleagues and collaborators around the world. Technologies supported by the Internet also provide the means for scientists, irrespective of their location, to take data and manipulate and combine it into new datasets for further analysis.

It is clear that global collaboration is becoming a wide-spread phenomenon that often relies on large, complex and highly distributed databases.

Scientists’ ability to analyse and interpret data to obtain a better understanding of scientific phenomena is dependent upon:

- improved ways to manage massive amounts of data from observatories, satellites, sensors and scientific simulations,
- advanced integration of powerful analysis tools into databases to enable intelligent interaction and interrogation,
- improved forms of scientist-computer-data interaction that support visualisation, annotation and interactivity,
- enhanced collaboration and cooperation among geographically dispersed teams of scientists, and
- the transformation of scientific communication and publishing.

Significant data management, analysis and visualisation challenges are similar across many different scientific disciplines.

To empower scientists whose research relies heavily on data we need to remove any major obstacles and provide solutions in a concerted and systematic way in order to maximise their capabilities to contribute to the national scientific endeavour.

The Importance of Data to Science

We live in the age of the knowledge economy. Knowledge is not only a driver of new creative industries and high technology businesses, it is also relevant to traditional manufacturing, mining, primary and service industries and to predicting our future infrastructure needs. It is equally crucial to our health and wellbeing, to understanding our past and our plans for the future.

One of the basic building blocks of knowledge is data – increasingly available, more diverse than ever before, with fantastically enhanced capacities for linkage, analysis, transportation and sharing. All disciplines, from the physical and chemical, engineering and agricultural, biological and medical to the social sciences and humanities, depend upon data. This is particularly critical for research projects requiring data from around the world, notably in biological, earth and space sciences, such as the Human Genome Project, the International Geosphere-Biosphere Programme and the Hubble Telescope.

Urgent Demand for Data for Today's Complex Societal Problems

Our ability to manage the dramatic increase in scientific data and to store, link and use it has come at a time of an increase in complex problems which are challenging many societies. Examples of these problems are environmental degradation, climate change, water shortages, mental ill-health, child and youth problems, epidemics of new or known infections spread by human/agricultural activities, social unrest and terrorism. All of these challenges demand collection of and access to data across disciplines, jurisdictions and time periods for monitoring and for the development of effective solutions.

Also important is the increasing ability to store and manipulate a wide range of data formats, including image, video and audio files, in areas as diverse as astronomy, medicine and music. As analyses become more elaborate, advanced techniques are required to manipulate, visualise and interpret scientific data. The ability to aggregate disparate datasets has the potential to lead to new, often serendipitous discoveries that have the potential to answer complex questions, such as those around land use in the context of the current drought.

In discussions around the current drought in Australia, interested parties may want researchers to consider issues such as ecological sustainability and the social and economic impacts of land degradation. In doing so researchers may pose questions such as:

- *Where are current land use practices sustainable?*
- *What management actions are appropriate across industries and commodities to minimise off-site impacts?*
- *What are the social and economic effects of land degradation on rural Australia?*
- *What are the long-term climatic models in Australia?*

To answer these questions researchers need data about agricultural land-use practices and the social and economic factors that influence them, land tenure, land degradation patterns and commodity economics and climate models. The data used needs to be:

- *relevant — providing factual social, economic, and environmental climatic information that meets requirements of users with different perspectives, interests and values;*
- *accessible — presented in a way that is easy to understand (for researchers) and readily available; and*
- *consistent and comparable — able to be integrated with other data to analyse trends in the state of natural resources.*

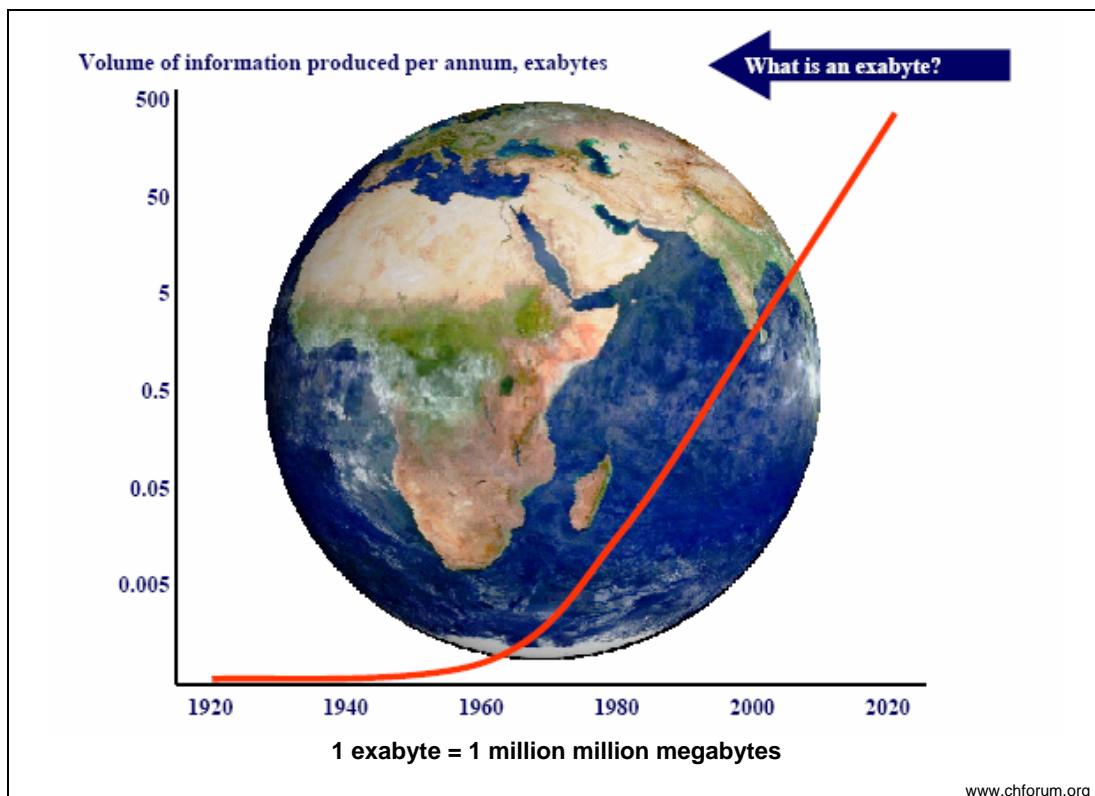
The Data Explosion

Climate, earth science, and genomic research are three examples of science areas that are generating vast volumes of data which is growing exponentially. Scientists, and science organisations, have difficulty in keeping up with this 'data deluge'¹. As a consequence, the way scientists interact with the data and with one another is undergoing a fundamental paradigm shift.

Grave concerns have been expressed about the, "...imminent flood of scientific data expected from the next generation of experiments, simulations, sensors and satellites."² Large facilities, such as the Large Hadron Collider in Switzerland, which is scheduled to begin operation in November 2007, and Australian Synchrotron (also scheduled to open in 2007) are expected to generate enormous amounts of experimental data on a daily basis. When fully functioning, the synchrotron will generate approximately 20TB per week. Numerical simulations, experiments and sensors, such as those on the Great Barrier Reef pose enormous challenges in terms of the volume of data and the storage and retrieval of such data. How to transmit these data in real time, ingest and store it effectively with current technologies is an area of research in its own right.

Figure 1

The following diagram shows one scenario projecting an exponential growth in data. For example, in 2010, according to this graph, there will be more data being generated (~50 exabytes (EB)) than has ever been generated in human history up to 2006 (~5 EB). By the year 2020, well within our own life time, the volume of data will grow to ~500 EB. Such explosion of data volume has tremendous implications.



¹ Tony Hey and Anne Trefethen, *The Deluge of Data: An E-Science Perspective*, UK e-Science Core Programme. Ch36 in *Grid Computing*, edited by Fran Berman, Geoffrey Fox, Tony Hey; in *Wiley Series in Communications Networking & Distributed Systems*, edited by David Hutchison; 29 May 2003: Digital Object Identifier: 10.1002/0470867167.ch36; [online] available from <http://www3.interscience.wiley.com/cgi-bin/summary/104535645/SUMMARY>

² *ibid.*

Data Access, Discovery and Linkage

Scientific activities and sophisticated computational technology and capacity are forcing and enabling the collection and storage of more data, often in diverse formats, such as photos, imaging files, audio records, genetic micro-arrays and visualisations. Over the last decade, universities and research groups have collaboratively or individually developed repositories and web-based subject gateways designed to provide researchers, students and specialists with access to authoritative data on specific areas of knowledge. These individual efforts have provided a substantial contribution to scholarly endeavour.

Information seekers and researchers today are more than ever faced with an overwhelming range of information and data sources of varying quality. Internet search engines are a useful immediate solution, but do not provide access to the depth of information available and to data stored behind access gateways or institutional firewalls. The ability to search across disciplines, platforms and specific data repositories is not sufficiently facilitated in Australia.

For some scientists a dramatic shift is occurring as a result of the change to next generation ICT infrastructure and research instruments generating abundant new sources of data. Large challenges are emerging in storing, managing and accessing these data. But for many other scientists, their work is based on existing sources which have not and are unlikely to change so dramatically. In both cases tools are needed to assist these scientists to access data online through improved discovery mechanisms, better access management and to provide them with an improved understanding of the role and responsibilities of researchers and institutions in making data available online. Challenges include agreements on standards for storing data, long-term preservation arrangements, protection and management of intellectual property, access to and use of data for other purposes.

Doing More with Data

Solving complex problems requires collaboration across the silos of both science and bureaucracies. A more integrated approach to data collection, management and particularly access and analysis will not only reduce expensive duplication of effort and resources, but will lead to more effective use of existing data.

In spite of the exponential rise in the range and quantity of data available, these precious resources have often been under-utilised, hidden or locked away. Yet these datasets are valuable and often unique, and it can be impractical, too difficult or too costly to replicate them. In future, data will increasingly be used for research activities beyond the scope of the research that led to the original collection of the data.

Australia has a rich collection of surveys, cohort studies, population databases and administrative records. These collections are held discretely and are under-utilised in terms of their potential to yield new information of value.

The National Data Network being established by ABS will provide infrastructure, protocols, standards, and services to support the sharing and integration of data across Australia. It will be developed to increase the availability, accessibility, and usability of information sources relevant to policy analysis and research – particularly key administrative and survey datasets held by Australian, State and Territory government agencies.

Currently, public funds are often expended on generating more data, when the answers to many problems could come from utilising existing data. For example, data collected as part of the Australian census is unique and based on a snapshot in time. The data for a particular census can be reused in a variety of circumstances and is a powerful tool when combined with previous census data, or other longitudinal and historical data.

The demand to share and re-use data presents challenges of how different data in a variety of formats can be brought together. And as technology links more and more people and data together, concerns about legal, social and ethical issues need to be addressed through legislation and guidelines.

In addition, the scale of many global collaborations and the cost of ICT infrastructure and research instruments are prime drivers for ensuring that such large, high cost science delivers maximum value through effective storage and management of the data generated.

The Mars rovers Spirit and Opportunity landed on Mars in January 2003. While mission scientists and engineers hoped to obtain at least three months out of the rovers, both continue to send back data. The Mars Exploration Rover project has received vast quantities of data such as photos, imaging, and data from soil testing and analysing all these data will take many years to complete.

Costs of maintaining the link run to millions of dollars a day. But the cost of replicating the mission is many times that. Reusing the data maximises the investment in this pioneering venture.

National Aeronautics and Space Administration Jet Propulsion Laboratory³
and Imaginova⁴

Stewardship of Data

A fundamental issue is the archiving of data and the stewardship of this process. Arguably a lot of attention is being given to archiving data from research communities such as genomics, bioinformatics, climate science and astronomy. This is mostly because they rely on small communities of interest, closely linked through global grid networks with intense use of the data in collaborative activities. The greater challenge may be from smaller isolated researchers whose data are heterogeneous, and also extensive. If they were gathered from their current locations, such as poorly organised, insecure, desktop computers, personal CDs, and other distributed storage, these datasets collectively are likely to be of similar size to those of the larger more coordinated groups.

Incentives and Disincentives

For Australian researchers, collaboration with peers both in Australia and internationally, through (global) research networks, is essential. Connecting to such networks provides a strong incentive to share data, and around these networks or communities of interest there appears to be a growing awareness of the value of data and the need to deposit and manage them.

Outside of such networks there is less incentive to preserve and manage data for the future. While there are policies – at the funding body level and at the research organisation's level – there are no strong rewards for adequate data management or penalties for poor data management. In the absence of an organised infrastructure to support the deposit of data into repositories, good governance and policy guidelines, data storage and retention will continue to be patchy and suffer from duplicated effort, lost opportunities and poor use of skill sets.

³ <http://mars.jpl.nasa.gov/missions/present/2003.html>

⁴ <http://www.space.com>

International Activities

While international science collaborations have been in evidence for a while, particularly those based on specific discipline and around costly infrastructure, the acknowledgment of the pressing need to create an international focus on developing standards and guidelines for data management is only a recent phenomenon.

Considerable work has already been undertaken to support the development of distributed learning and the development of systems to store and manage digital content in this environment. While there are flow-through effects from this work in the e-learning space at a structural level, there is substantial work still to be done to develop appropriate standards to support the curation, preservation and management of data for science.

Case Studies of Interest

Appendix A of this report contains case studies that demonstrate and elaborate on the issues raised in this introductory Chapter – particularly how sharing scientific data can contribute to solving complex problems.

Case Study 1

Avian Influenza Gene Sequence Sharing

The Global Initiative on Sharing Avian Influenza data (GISAID) represents a worldwide effort to control the spread of H5N1 avian influenza or bird flu. This case study shows us the type of issues scientists had to overcome in order to be able to share their data and the benefits of doing so.

Case study 2

Australian Ocean Data Centre Joint Facility

This joint facility is a new and collaborative initiative which aims to promote the discovery, access, long-term archiving and exploitation of knowledge about marine systems. It is part of the global system of data centres which promote the free and open exchange of marine scientific data. This case study shows us which organisations are involved and how the facility is organised and managed.

Case Study 3

PARADISEC – Pacific and Regional Archive for Digital Sources in Endangered Cultures

This case study is an example of a vulnerable dataset. PARADISEC is focussed on regional Pacific and Australian languages, knowledge of which is predominantly oral. The languages, existing research records, and the archive itself, are all vulnerable. The case study outlines briefly what is involved in digitising existing records of these languages and maintaining large datasets of complex linked objects (sound, audio, video and text).

‘When dealing with Indigenous cultural or heritage information, knowledge and/or data, more often than not information is categorised by the relevant Traditional Owners or custodians as either highly sensitive and/or restricted or available for public access. Restrictions and sensitivities are generally associated with factors such as gender, age, or right/responsibility to access particular knowledge. It is essential when dealing with Indigenous cultural or heritage information that appropriate policies are put in place to ensure the protection of information deemed restricted by Traditional Owners and/or custodians. Additional issues around intellectual property, ethics and privacy are also important to consider with Indigenous data.’⁵

⁵ Department of Environment and Heritage, pers. comm. submission to the Working Group

Chapter 2: The Emerging Research Information Environment

Chapter Overview

This Chapter explains in more detail what is involved in managing data and why access, ownership, storage, capacity to share, search and link data as well as an awareness of its existence are fundamental for collaborative research and general data re-use. The Chapter also considers privacy and intellectual property (IP) implications when sharing data as well as the skills needed by today's scientists who will work with large, often heterogeneous datasets. These data are often held in online, distributed databases and can be exploited by new tools and techniques tuned for data analysis, visualisation and mining. The Chapter suggests that the emerging information intensive environment holds both threats and opportunities.

Although the current research environment is characterised by massive quantities of data, the ability to harness these data to solve complex problems is often hampered by inadequate data management practices. Fortunately advances in information and communications technology offer some solutions to this pressure.

Making Data Available

Internationally and nationally there is a growing awareness of the importance of maximising the availability of scholarly output and scientific data, particularly that resulting from research derived from government funding. The Internet and associated digital networks, along with improved computational capacity, have created a range of opportunities and challenges – for individuals, disciplines, institutions and governments – to change the way information is stored, communicated and integrated. These changes demand that we annotate data with supplementary information (metadata) to aid in its use and re-use.

Many people in academic, research and government organisations are routinely involved with the production and use of highly complex research-oriented datasets and the production and reading of scholarly journals. The extent to which these institutions and different scientific disciplines actively manage these data for re-use purposes varies widely. Where data are actively managed, the practices and technologies used, as well as the data management services provided differ significantly. For example, a newly formed Australian marine science network is in its early stages of deploying remote micro-capture devices and aims to channel data derived from these sensors directly to an online, distributed data management system for publication. This approach contrasts with the practices and technologies used by the Australian Social Sciences Data Archive, a mature organisation with 30 years of service provision, that manages data through a centralised repository. While these approaches differ markedly, both focus on making data available in a manner best suited to serve their constituent user base.

With such a diversity of data management approaches, some of the challenges we face are how to locate sources of data that are available and how to effectively combine or analyse these disparate datasets once they are found.

The Australian Social Science Data Archive (ASSDA), located in the Research School of Social Sciences at The Australian National University, 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.

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 1050 datasets from Australian surveys and opinion polls. ASSDA also holds Australian population census data and similar 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.

*The staff at ASSDA have technical expertise in a number of areas, including digital preservation, computer programming, research methods, statistical analysis and the production of metadata.*⁶

ICT Infrastructure and Databases

Globally, technologies are changing the way research is performed. Fast, broad-band multimedia communication channels, distributed high performance computing, high volume storage area networks and innovative data management and data analysis software are driving much more collaborative research. Sensor technologies permit the collection of data over much larger spatial scales, penetrate deeper into substrates, organic tissue and the hydrosphere. Data collection rates are growing exponentially and simulations now generate vast quantities of modelled data. In Australia we need to ensure that we have a world class research information infrastructure, capable of supporting this changing research landscape.

Rapid developments in ICT infrastructure and applications in research and science – and in scientific publishing - present an opportunity to “do something bold.” But no one individual or institution has sufficiently strong incentives, or the authority, to change the current system.

With the rapidly expanding availability of primary sources of digital data, the Working Group has noticed a clear shift from the use of secondary sources, such as scientific journals, to reliance on databases of scientific data. The Microsoft Report *Towards 2020 Science*⁷ takes this concept one step further. It posits a major change in scientific method taking place where,

“[h]ypothesize, design and run experiment, analyze results” is being replaced by “hypothesize, look up answer in data base.”⁸

Databases are clearly now performing a central function in the changing research landscape. They play a key role in making efficient and effective use of data, including facilitating its repurposing and re-use, which are now essential components for achieving better research outcomes. Increasingly these databases are being aggregated or federated to form repositories, data centres or nodes in discipline specific community networks.

⁶ <http://assda.anu.edu.au/>

⁷ Microsoft Research Cambridge, *Towards 2020 Science* [online]; available from http://research.microsoft.com/towards2020science/downloads/T2020S_ReportA4.pdf; accessed 30 October 2006

⁸ Michael Lesk, <http://archiv.twoday.net/stories/337419/> 2004. [Accessed 4 December 2005] quoted in *Towards 2020 Science* (ibid) p.19

The Working Group found that, despite the inherent diversity in discipline-based data management approaches, there is also a set of common generic problems which affect organisations and disciplines in areas such as data creation, discovery and access. These problems could be addressed for mutual benefit through organised and skilled leadership. Generic issues which could be tackled systematically include:

- the development of common standards and protocols for storing, managing, and disseminating data;
- implementation of technological infrastructure capable of supporting access to information and facilities;
- regulatory environments that both enable and encourage the population of repositories and support information sharing; and
- skills development and cultural change to maximise our collective capacity to better exploit and re-use data.

The growing reliance on data as the currency of scientific endeavour, coupled with the growth in the volume of data as highlighted in Chapter 1, means that the development of a national infrastructure to better manage scientific data is vital.

Australia's early response to meet this need has been the funding of a number of programmes to put in place foundation pieces of what will be a highly complex jigsaw puzzle:

- the Systemic Infrastructure Initiative;
- National Collaborative Research Infrastructure Strategy;
- National Data Network.

National and overseas initiatives, which inform the development of Australia's fledgling infrastructure, have also been examined by the Working Group. Of significance are:

- e-Research Coordinating Committee;
- OECD Committee for Science and Technology Policy;
- International Council for Science, Committee for Scientific Planning and Review; and
- UK e-Science Programme.

These initiatives are described in more detail in Appendix C.

A significant quantity of existing Australian scientific data could potentially be re-used as valuable input to current research and to support solutions for national priorities if it were more discoverable and better able to be accessed and integrated. Achieving this requires the technology and capability to:

- link data from current and past research studies with other non-academic and government sources of information;
- provide a focal point for Australian involvement in major international research collaborations, data pooling and scientific studies;
- support individual researchers and institutions to link and use scientific datasets, and engage in multi-centre investigations and collaborations;
- build the capacity of the research community to analyse and interpret the implications of the linked information, by providing training opportunities for researchers and assisting in the development of measures and methodologies for scientific studies;
- analyse, understand and develop protocols for responding to the legal, regulatory, ethical and privacy issues attending the linking of data and sharing of data; and
- develop strategies and proposals to further invest in these capabilities as needs and opportunities arise.

Some Challenging Questions for Australia

Significant work is already underway on developing, populating and ultimately federating repositories in a number of academic institutions. Much of this work is currently funded under the Strategic Infrastructure Initiative (SII) and is being executed with a view to including the likely requirements of the Research Quality Framework. The majority of these repositories cover e-prints, reports and theses but do not cover the full scope of data for science. Outside of the academic sector initiatives such as the Australian Ocean Data Centre Joint Facility (AODCJF)⁹ are providing distributed libraries of data holdings to support policy analysis and research.

Data for Preventative Health Research

The amount of health and related data collected and generated in electronic form continues to grow rapidly. Large datasets are being produced through clinical care, pathology, diagnostic imaging, health system administration, nutritional intake and activity surveys, surveillance and monitoring, clinical trials, genomics, proteomics, and cell and structural biology.

There is enormous value to be gained in bringing together these datasets for deep analysis, and to use the findings to systematically improve our understanding of disease, disease monitoring, and health issues. For example, in bringing together genomic and lifestyle datasets we can explore and begin to understand how genetic risks are modified by environmental influences.

The CSIRO Preventative Health National Research Flagship is investing in the new ICT, mathematical and statistical approaches which will be needed to address the challenges of:

- *simultaneously understanding health data in many different formats including: structured and unstructured data, images and other multimedia, free-text reports, data streams, disease-specific registers, and genomic data and*
- *analysing these new data, for the first time simultaneously addressing the challenges posed by biodata which is multimedia, multi-format, of variable quality, large, cross-scaled, linked and longitudinal.*

The Data Linkage Unit in WA has developed a system of data linkage for population health data which both protects privacy and enables the production of integrated files to evaluate health care, seek causes of disease and investigate trends over time.

While all of these initiatives are breaking new ground and are helping to address procedural and technological challenges, there is a strong need for a more systematic approach to many of the generic data management problems which are common across research disciplines and across organisational and institutional boundaries. Systemic questions which arose as a consequence of the Working Group's investigations included:

- Are our current national approaches for managing data for science adequate?
- Is there sufficient access, storage capacity, sustainability of practices, security, preservation and capability to share data in all the areas of science, not just to address the complex societal issues already mentioned but to keep us competitively participating in the global world of science and its development?
- Are researchers and research funders aware of the important data assets which are being created? Do these parties understand that those assets must be kept so that they can be used as widely as possible not only today but for Australia's and the world's future scientists?

⁹ Australian Ocean Data Centre Joint Facility, Appendix A, Case Study 2., p.69

- Do funding agencies recognise data management as an important and legitimate part of research and adequately fund those proposals which explicitly include a data management strategy as part of their research methods?
- Are Australian scientists aware of the relevant standards for data management?
- Are research proposals that include methods to re-use existing databases for cross-cutting research and for novel analyses given appropriate weighting in grant applications compared with proposals to collect and create new datasets?
- Will researchers and data custodians want to share the data that they have so painstakingly collected, sometimes over many years?
- What are the attitudes to the ownership of data and does the push to protect data for commercialisation interfere with its availability to be used more effectively for discoveries on a much wider scale?

In general the answers to these questions are “no”. In the case of the last dot point attitudes to data sharing do negatively affect our collective ability to make best use of the data we capture. There was a high degree of agreement among all the submissions received on these points.

For example, the submission by the Australian Academy of Science argued that “unrestricted access to publicly funded research data is most satisfactorily enforced by international scientific journals with editorial policies that insist that materials and data underlying any publication are available to readers on request.”¹⁰ Indeed, such editorial policies, that support peer review processes, are part of the answer to enabling greater access to data, as the following example from the biomedical discipline illustrates:

Editorial Practices for International Scientific Journals

In the areas of biochemistry and molecular biology, papers reporting protein or DNA sequences and crystallographic structures will not be accepted for publication without proof (ie an accession number) from a recognised database in general use in the discipline.

The editorial policies in place in this regard for *Nature Genetics* are typical for biomedical research:

AVAILABILITY OF MATERIALS AND DATA

An inherent principle of publication is that others should be able to replicate and build upon the authors' published claims. Therefore, a condition of publication in Nature Genetics is that authors are required to make materials, data and associated protocols available to readers on request. Any restrictions on the availability of materials or information must be disclosed at the time of submission of the manuscript, and the methods section of the manuscript itself should include details of how materials and information may be obtained, including any restrictions that may apply. Authors may charge a reasonable fee to cover the costs of producing and distributing materials. If materials are to be distributed by a for-profit company, this should be stated.

Australian Academy of Science¹¹

¹⁰ Australian Academy of Science, *A Submission to the Prime Minister's Science, Engineering and Innovation Council Working Group on Data for Science* 11 August 2006: p. 5

¹¹ Ibid.

There are some striking examples of changing attitudes and perspectives towards the ownership and value of data. The Treasurer made the following comments in a speech celebrating the Australian Bureau of Statistics' Centenary in December 2005. In this speech he recognised that the government could derive significantly more value from data through its re-use by reducing impediments, such as charging fees for distribution and access.

In June this year I was happy to announce that ... consistent with the Government's policy of Backing Australia's Ability, many ABS publications would be available free of charge from the Internet. These publications previously cost between 20 and 40 dollars each.

Today, as a tribute to the people of Australia, and to enable them easier access to information about our country, and in celebration of the ABS Centenary, I have great pleasure in announcing that as from next Monday morning, ALL the statistical information published online by the ABS will be available free of charge. This means that electronic versions of all ABS publications, and all Census data, time series data and spreadsheets may be downloaded from the ABS web site, absolutely free.

The Treasurer, the Hon Peter Costello MP ¹²

We need to apply the Treasurer's philosophy on removing barriers to re-use to many other areas of government endeavour, particularly in the research sector. Scientists will often not share data because of a perception that restricting access to data imparts a competitive advantage for the custodian when it comes to applying for grants. Our grants system needs to combat this perception by rewarding collaborative activities which promote re-use and data sharing.

An interesting question is whether we are adequately aware of what data we already have and what data we would need to have in order to answer some of Australia's more pressing environmental and social problems. That is, do we have appropriate mechanisms operating within our museums, libraries, universities and other research institutions which allow us to ascertain gaps in information across the spectrum of scientific disciplines? Unfortunately, we do not have in place ubiquitous systems or institutional processes that easily enable us to derive such information. In Chapter 4 we illustrate this particular problem in more detail by discussing the case for a National Environmental Reporting System.

Are there precious data sets of huge importance which are in danger of being lost because of changes in personnel or technology, lack of funds for their archiving or because software or hardware have changed so much that the data are not retrievable? It seems that this is at least a possibility:

¹² The Hon Peter Costello MP, *Address to the Australian Bureau of Statistics Centenary Celebration, Canberra* delivered 8 December 2005 [online]; available from <http://www.treasurer.gov.au/tsr/content/speeches/2005/019.asp>; accessed 26 September 2006

Just 37 years after Apollo 11, it is feared the magnetic tapes that recorded the first moon walk have gone missing at NASA's Goddard Space Centre in Maryland. A desperate search has begun amid concerns the tapes will disintegrate to dust before they can be found.

It is not widely known that the Apollo 11 television broadcast from the moon was a high-quality transmission, far sharper than the blurry version relayed instantly to the world on that July day in 1969.

*CSIRO scientists began researching the role of Parkes in Apollo 11's mission in 1997, before the movie *The Dish* was made. However, when they later contacted NASA to ask about the tapes, they could not be found.*

"People may have thought 'we have tapes of the moon walk, we don't need these'," said the scientist who hopes a new, intensive hunt will locate them. If they can be found, he proposes making digitalised copies to treat the world to a very different view of history.

But the searchers may be running out of time. The only known equipment on which the original analogue tapes can be decoded is at a Goddard centre set to close in October, raising fears that even if they are found before they deteriorate, copying them may be impossible.

"We want the public to see it the way the moon walk was meant to be seen. There will only ever be one first moon walk."

*The Sydney Morning Herald*¹³

Postscript: These tapes were recently found.

In many instances even if we do know where data are, if they are not being actively managed, there is a high probability that they are in danger of being lost to us forever because of changes in the chain of custodianship or their curation technology. Many submissions indicated that both poor foresight and a lack of funds for archiving render important datasets unusable over time. Rapid changes in software or hardware mandate that electronic data are managed continuously to ensure that it is readable using current technology. It is not feasible to maintain a myriad of obsolete systems just to ensure the longevity and usefulness of data. More sophisticated data management approaches are required, which necessitates investment in dedicated repositories and data centres staffed by skilled personnel. The cost of re-collecting data is vastly more expensive than the investment required to adequately secure data once collected.

The Australian Partnership for Sustainable Repositories (APSR) is solving some of these issues through its Automated Obsolescence Notification System (AONS). The APSR system aims to help librarians, archivists, and repository managers in the increasingly onerous task of coping with the diverse range of file formats in use – not to speak of the new ones arriving all the time.

AONS is a software solution that works in conjunction with leading open-source repository platforms, such as Dspace and Fedora. AONS is essentially an obsolescence detector – an autonomous process that periodically scans all digital objects collected in a repository looking at the file formats that have been used. Each file format found is checked to see if it is obsolete, or at risk of becoming so. If the repository is found to contain such formats a notification report is sent to the repository manager responsible.

This is clearly a first step in the long term challenge of preserving important datasets and automating the process of continually migrating out-of-date data formats to current technologies.

¹³ Richard Macey, *One giant blunder for mankind: how NASA lost moon pictures* The Sydney Morning Herald, 5 August, 2006 [on-line] available from <http://www.smh.com.au/news/national/one-giant-blunder-for-mankind-how-nasa-lost-moon-pictures/2006/08/04/1154198328978.html>; accessed 6 September 2006

What Role for Industry?

What leadership role should industry be playing? Businesses have data needs, are keen to capture new knowledge and discoveries, are designing and developing new software, hardware and ways of sharing and analysing data that are potentially very useful to researchers and others needing and using data. Do we have appropriate strategies across the nation that enable industry to contribute and share in the data and knowledge revolution? What role could industry and public/private partnerships play in funding repositories or enabling discovery and preservation of data, likely to be an extremely expensive exercise when all areas are considered?

Part of the answer is to ensure that information about research is as widely available as possible, and to develop a consistent, national approach to the management of and access to data. This will have flow-on benefits to industry, as opportunities for commercialisation and collaboration become known.

There are some industry associations such as the Australian Spatial Information Business Association (ASIBA) that have been formed to exploit spatial data. The ASIBA, a relatively new industry peak body, is a good example of how a co-ordinated, but diverse industry sector can effectively mobilise to partner with the public sector on data issues. ASIBA was a major contributor in formulating an action agenda to grow the spatial information industry and regularly works with the government on the development of policy frameworks aimed at maximising the use, distribution and creation of publicly funded data products and services. The Spatial CRC is also another mechanism for encouraging industry, particularly software vendors, to become involved in research and development that can lead to better management and exploitation of public data assets. The spatial industry sector is particularly well organised and active in public/private partnering. Unfortunately this is not always replicated in other sectors.

A number of recommendations made by the Digital Content Industry Action Agenda, in relation to statistics, standards and data management, were endorsed by the Government.

The recommendations include:

- assessment of existing industry frameworks to improve measurement of the digital content industry;
- exploration of opportunities to generate research data to reduce the current high-risk profile for investment in digital content activity;
- identification of opportunities for regular and follow-up industry surveys to identify benchmarks and change;
- identification of strategically important industry standards in key areas and encouragement of wide dissemination; and
- support of industry involvement in the formulation of next-generation standards. ¹⁴

The Impact of Privacy on Research

Some concern was expressed to the Working Group about the impact of privacy legislation on research, particularly in the area of health and welfare, although privacy issues also arise in other disciplinary areas, for example with commercial fishing data in the marine sector. This has led the Working Group to ask whether the much publicised concerns about privacy are appropriately balanced against the public good which arises from using individual data, particularly in relation to enabling the best government services to be developed. Do privacy concerns unreasonably interfere with research of national importance? Is the variation in privacy legislation across Commonwealth, States and Territories and federally creating unnecessary barriers to sharing of important data?

¹⁴ Department of Communications, Information Technology and the Arts, pers. comm. submission to the Working Group

Privacy regulations apply to the collection, storage, use and security of personal information. Concerns around the use of private information in research usually relate to fear of unauthorised disclosure to others; misuse of information, especially in a discriminatory way, that may affect health or life insurance or employment; or use of information for commercial transactions. The challenge for those using any personal information is to ensure that these concerns are acknowledged and that actions are taken to minimise the risk of disclosing private information, or it being misused. In many cases data can be de-identified, aggregated or filtered to remove components of the data which then protects an individual's or a company's identity.

In evaluating health services, finding the causes of cancers or other diseases, ascertaining the harmful effects of drugs and a range of other "public good" health research, the best science is done by using individual patient records and linking them with other information to ensure that complete population health data are used. This can only be done without consent as it is impossible to contact the large numbers required for accurate statistical analysis. Current NHMRC guidelines and Commonwealth privacy legislation in Australia allows use of individual data for such research within strict guidelines and conditions.

With new methods of record linkage, a "win-win" approach can be used in which the identifying information is linked independently of the researchers and without the sensitive information attached; the researchers are then given a de-identified dataset with the important sensitive clinical and other information to enable them to do their research.

The problem which has occurred lately, both here and overseas, is that ethics committees are interpreting guidelines in an increasingly conservative way¹⁵. This is starting to limit many very important projects which have the capacity to improve health services or find the causes of diseases. The potential social benefits of research seem to be discounted in favour of privacy concerns – the Working Group do not believe that this is serving the best interests of the community.

Other Regulatory Factors

Also of concern is the complexity and uncertainty of copyright legislation. Many of the large databases that are available online in Australia have been compiled in overseas jurisdictions. Generally these databases are subject to the laws that apply in those jurisdictions, and the treatment of these databases varies.

Owners of private databases compiled in Australia may seek to protect their database through a combination of means: copyright law, technological protection measures, or contract law. Where the database is compiled as a result of public funding, access regimes should be clear.

Significant momentum is being generated, particularly in the UK, for researchers to make their published papers freely accessible by placing them in the public domain, such as through an institutional repository or other public database, or in open access journals.

Similar principles should be considered for scientific databases comprising observational and experimental data which are generated as a result of public funding. As discussed in the biomedical example on page 31, there is a growing trend towards submitting journal articles that are openly linked to scientific databases to ensure transparency in the conclusions reached and to enable critical peer review of the research.

¹⁵ Academy of Medical Sciences (UK) *Report on Personal Data for Public Good: Using health information in medical research*. January 2006

Data Analysis and Data Management Skills

Submissions received by the Working Group indicate that we should be concerned about the low number of skilled scientists and data managers available in Australia who are capable of both developing the emerging information infrastructure and working within it. What can be done to rectify this situation in the short- and long-term? We need to be confident that Australia has the governance structures and institutional arrangements in place to ensure that decisions about building the emerging information infrastructure can be made and then executed. A related issue that was raised was whether Australia currently has the right educational framework to develop a skilled workforce both aware of and competent in, data management. This educational framework includes the initial training of scientists as well as their ongoing skills maintenance and updating and the support for multi-disciplinary teams.

The Australian Partnership for Sustainable Repositories, a project funded under SII, recently conducted a survey, examining sustainability issues for Australian research data. The report of this survey was published in October 2006.¹⁶ Among other findings which support many of the Working Group's observations, the study found that researchers and research groups are generally very insular in their management and use of data, a situation which 'does not support the sustainability and sharing of research data.'¹⁷ The report also laments the lack of any training in data or repository management in most existing tertiary courses covering scientific disciplines. The Working Group itself has noted that there are few undergraduate courses that appropriately prepare scientists to perform the types of quantitative analyses required to mine and manipulate large volume datasets and find patterns in heterogeneous, multi-disciplinary databases. These are skills which will be crucial to a successful career in science.

Importance of International Research Collaborations

Many problems and scientific discoveries are global in scope and require countries with relevant knowledge and scientific know-how to work together for mutual benefit. Sharing our capacity to diagnose, monitor, treat and, most importantly, to prevent or ameliorate global scourges such as HIV/AIDS and bird flu, and challenges such as climate change, water shortages and terrorism, is becoming the norm in science.

Most international problems ignore national boundaries so collaborating and sharing data are essential. The oceanographic and atmospheric communities, for example, have recognised this for many years and have developed a very collegiate approach to conducting research. As a result data sharing and data sharing networks are intrinsic to the conduct of most science within these disciplines, enhancing both the breadth and depth of the questions tackled. In contrast, other research communities have only recently recognised the benefits of cross-border, collaborative activity and therefore have significantly less mature systems to support such endeavours. These communities have traditionally been characterised by science conducted by individuals and small groups, relatively independent of each other. This trend is, however, changing.

In its discussion paper, released in June 2005, the e-Research Coordinating Committee (eRCC) noted considerable Australian involvement in international projects based on e-research and promoted the capacity of e-research to cost-effectively:

- increase opportunities for international involvement by Australian scientists, including leveraging their strengths;
- enable Australian science to benefit from international collective investments in e-research infrastructure; and

¹⁶ Markus Buchorn and Paul McNamara, *Sustainability Issues for Australian Research Data: the Report of the Australian e-Research Sustainability Survey Project*, APSR Publications, October 2006 [online] available from <http://hdl.handle.net/1885/44304>; accessed 4 November 2006

¹⁷ *ibid.*, p.48

- promote Australia's excellence in ICT, including middleware and advanced networks development.

The eRCC found that progressing the e-research agenda is as much about people as it is about technology, because e-research is challenging existing research practices and cultures. In its interim report, published in September 2005, the eRCC identified the following barriers to national and international collaboration:

- a lack of awareness by Australian researchers of the collaborative potential of e-research;
- cultural barriers to recognition and reward of e-research methodologies and the need for cultural change and skills development;
- data access and security issues;
- a need for interoperability and standards development, particularly for data management; and
- a lack of support for researchers (in terms of developing and using e-research technologies and middleware).

The eRCC identified the need for a strategic framework for e-research to increase Australia's involvement in both national and international collaborative activities. Many of the key issues identified relate to the need for improved support for researchers; awareness raising, training and cultural change to foster participation; along with data management, accessible databases and authentication and authorisation strategies.

The international organisation established to develop the Square Kilometre Array radio telescope illustrates the potential of large-scale international collaborations and is described below:

In September 1993 the International Union of Radio Science (URSI) began a worldwide effort to plan for a next generation radio observatory.

The instrument – the Square Kilometre Array (SKA) – is projected to cost US\$1 billion, beyond the capacity of a single country's astronomy community to fund.

The particular requirements of a SKA site, including radio quietness, physical characteristics of the site, and tropospheric, atmospheric and ionospheric conditions, severely limit possible locations for a SKA. These requirements, combined with the prohibitive cost of the instrument, mean that a single SKA will be built to service the global radio astronomy community.

On 28 September 2006, the International SKA Steering Committee announced that Australia and Southern Africa have been short-listed as the countries to host the SKA.

The Square Kilometre Array will:

- *have a collecting area of almost one million square metres, giving it 50 times the sensitivity of today's best radio interferometer;*
- *be the first aperture synthesis telescope with multiple independent fields of view (up to 100 at one time); and*
- *integrate computing hardware and software on a massive scale, in a way that best captures the benefits of these exponentially developing technologies.*

A broad scientific community has mobilised to cooperate in achieving this common goal.¹⁸

¹⁸ <http://www.skatelescope.org/>

Case Studies of Interest

Appendix A of this report contains case studies that demonstrate and elaborate on the issues raised in this Chapter – particularly what is involved in managing data and why access, ownership, storage, capacity to share, search and link data as well as an awareness of the data's existence are so important.

Case Study 4

Australian Digital Thesis Project

This project is an Australia-New Zealand version of the worldwide movement to release research theses on the Internet for public access. The aim is to establish a distributed database of digital versions of theses available worldwide on the web. The case study shows the range of institutions involved, what each partner is responsible for doing and the characteristics they share.

Case Study 5

Species 2000 and ITIS Catalogue of Life

The Species 2000 and ITIS Catalogue of Life will become a comprehensive catalogue of all known species of organisms on earth by the year 2011. The case study outlines the difficulties this project faces given the complexity of compiling an index of this nature, and the disparate sources of information. It describes the range of data base projects which are collaborating to produce this catalogue and the organisations involved.

Case Study 6

Biometric Technologies

Biometrics is the term applied to technologies used for recognising an individual based on their physical and behavioural characteristics, such as face recognition, iris recognition, speech recognition and hand geometry. The case study describes the way this information is stored and discusses the importance of consistent metadata, protocols, standards and systems.

Case Study 7

ICTVdB, Universal Virus Database built for the International Committee on Taxonomy of Viruses

This database grew out of a think tank that examined the need for an all inclusive virus database. The case study describes the lessons learnt and achievements of building a database from data that is recorded, evaluated and published by individual researchers or teams after processes of peer review. It also highlights the fact that individuals and teams differ greatly in their creativity, styles, standards and research methods.

Chapter 3: The Way Forward

Chapter Overview

This Chapter presents the recommendations of the Working Group. These relate broadly to three areas:

- *the whole of the research system – a national strategic framework, a national network of digital repositories, data access and sharing protocols and the need to ensure that privacy and IP regulations do not impede data sharing;*
- *cultural and institutional change – which covers how to encourage better data management practices; and*
- *how to develop the skills required for researchers and others to be able to work within the emerging information infrastructure and the new data environments.*

The overwhelming majority of institutions and individuals who made submissions to the Working Group recognised that national action is now necessary to address scientific data related issues. A number of loosely connected Australian initiatives have begun tackling some of the identified problems. Some of these projects are developing various technological components that underpin a research information infrastructure: using grid technologies, promoting methods to preserve data for re-use, and exploring how to link government and academic data holdings in an open access environment.

The Working Group acknowledges that developing a national strategic framework, a national network of digital repositories and increasing training is a large undertaking that will require serious investment as well as cultural change. It may well require a detailed feasibility study (or studies) as a first step.

The biggest challenge will be to channel the knowledge and experience of the range of stakeholders dealing with the problem of data for science into a coordinated, productive whole, and to understand the scope and nature of the inherent issues and address these in a coordinated way.

To this end, the Working Group has made recommendations for action to address the challenges for better usage, linking and management of data in Australia.

The recommendations are aimed at ensuring Australia's science data assets are preserved, accessible, discoverable, re-usable and professionally managed. We propose approaches to systemic problems which must be tackled at the national level.

These recommendations are intended to apply to governments of all jurisdictions, the higher education sector, publicly owned research funding agencies and public research organisations. The goal is to encourage best practice in data management and to promote data sharing, which can only be achieved if research institutions embrace cultural change.

The Working Group also proposes that a specific focus on changing individual behaviour is a key component of the overall approach. The education and training of researchers, research support, data scientists and others should ensure that they have the basic skills required to manage and manipulate electronic datasets and that they understand what is required to curate data for long-term re-use.

Unfortunately, skill development alone will not solve the underlying problem. The Working Group acknowledges the need for fundamental cultural shifts, which are also highlighted below.

A. A National Strategic Framework for Scientific Data

The Working Group consulted with a wide range of stakeholders as part of its deliberations. (see Appendix B for details). The central focusing question posed in the consultations was: Should Australia have a national approach for managing the lifecycle of data for science?

The overwhelming majority of respondents agreed with that proposition.

As with similar reports conducted internationally, including those led by ICSU, the OECD, and the National Science Board (NSB) in the United States, the Working Group has concluded that the absence of a long-term strategic framework for scientific data management is a barrier to the best use of Australia's data assets. Without this national context and agreement as to common data management policies between and across jurisdictions and institutions, the current disparate approach to data will persist.

This was reflected in specific comments from a number of stakeholders, along the following lines:

'The large investment that government agencies and other players have made in health data has been based on the same belief as is expressed in the Working Party's documents namely:

Better national management of the data lifecycle leads to better, more efficient research and hence to better application of research to improve social and economic outcomes for Australians.'

Dr P Allbon, Director, Australian Institute of Health and Welfare¹⁹

Given the range and number of stakeholders engaged in aspects of this problem, coordination and leadership at the national level will be critical to ensure the engagement of a sufficiently wide range of interests.

The organisations consulted by the Working Group represented diverse research disciplines and research funding organisations, including from the humanities and social sciences as well as domains more usually thought of as "science". All of these areas indicated that they are affected by data management issues. There was general agreement that a national approach would be desirable.

The Working Group did not consider that it was appropriate to nominate any single institution or agency to take this important work forward. Instead, a high-level expert committee, informed by this report and others, could take a sufficiently cross-sectoral view of the problem and engage with relevant experts in different domains.

Recommendation 1:

That Australia's government, science, research and business communities establish a nationally supported long-term strategic framework for scientific data management, including guiding principles, policies, best practices and infrastructure.

Recommendation 2:

That a high-level expert committee be established to provide the leadership role in progressing the formation of the long-term strategic framework for scientific data management.

¹⁹ Dr Penny Allbon, Australian Institute of Health and Welfare, submission to the Working Group, 6 September 2006

B. The National Network of Digital Repositories

There has already been investment in technology through SII to enable the federation of data repositories in the higher education and research sector.

For example, the MAMS²⁰ project at Macquarie University and the e-Security collaboration between Macquarie and the University of Queensland²¹, which includes the involvement of a range of government and research sector interests, have been building the technological infrastructure to allow access to federated datasets based on a trust fabric of agreed authentication processes.

The technology in these fields is maturing to the point where it will be possible to move research projects into a deployment phase to capitalise on this investment in information infrastructure research.

There has also been considerable activity in the government sector, including, as mentioned earlier, the establishment of the National Data Network (NDN), led by the ABS. The NDN model for connected statistical data repositories is an important example of enabling linked-up use of data from a variety of sources by providing institutions with generic repository software tools. This is a great start towards building a federated system for this type of data but we must also cater for the development of repositories and repository networks where institutions can leverage their existing investments in repository tools and local ICT infrastructure by using open technology standards.

Achieving interoperability of systems to underpin better integration of data must be based on agreed and accessible standards. These standards are required at many levels within the system. The development and widespread distribution of standards is a sometimes costly exercise, but without them there is no basis on which different systems can work effectively with each other. The alternative to a standards-based approach is to enforce a single set of software tools, communication protocols and data definitions across the board. This is neither practical nor necessary. Interoperability is achievable with a constructive approach to standards-based developments. Standards, once developed and agreed, should be available free of charge to ensure broad adoption and use.

Apart from investment in software, there has also been investment in physical infrastructure in the higher education sector particularly through the Australian Partnership for Advanced Computing (APAC), and the Australian Research and Education Network (AARNet3). These physical assets could be better utilised in a more strategic approach to information infrastructure development.

The National Broadband Strategy Implementation Group (NBSIG), a Standing Committee of the Online and Communications Council (OCC), is responsible (through the OCC) for assisting Australian governments to develop strategies and programs that contribute to achieving the vision and the objectives outlined in the National Broadband Strategy (NBS). The NBSIG formed a number of Working Groups to facilitate and focus its activities, with one such working group being the Digital Content Working Group (DCWG).

The DCWG has identified a number of significant issues and opportunities regarding public domain digital content that require attention and potential action by Australian governments. These included:

- Networking, for example, the desirability of establishing a national broadband network that links up schools, colleges, universities, libraries, museums and other public institutions to content repositories and other shared services.

²⁰ MAMS is the Meta Access Management System. It was funded by DEST's Systemic Infrastructure Initiative (part of BAA 1) in 2003. This project allows for the integration of multiple solutions to managing authentication, authorisation and identities, together with common services for digital rights, search services and metadata management. More info [online] at <https://mams.melcoe.mq.edu.au/zope/mams>

²¹ <http://www.esecurity.edu.au/>

- Making content available, for example, through an Online Access Program for Australia's cultural collections
- Supporting demand and public engagement – for example, effective search engines.²²

The Government's Broadband Blueprint, which is currently in draft form, but is soon expected to supersede the NBS, acknowledges that the Australian Government will continue to give priority to ensuring that the Australian research community has access to the bandwidth necessary to participate fully in, and, as appropriate, lead international collaborative research.

The Australian Government's National Collaborative Research Infrastructure Strategy (NCRIS) has recognised that all areas of modern research are becoming dependent on technological platforms that enhance the research community's ability to generate, collect, share, analyse, store and retrieve information. Through its capability 5.16, Platforms for Collaboration, NCRIS is seeking to address this dependency by investing in a range of information infrastructure initiatives. Five key areas have been identified:

- data storage, management, access, discovery and curation to improve interaction and collaboration;
- grid-enabled technologies and infrastructure to enable seamless access to the facilities and services required by various research fields;
- support skills to assist researchers in developing and using this infrastructure effectively;
- high performance computing to allow modelling, analysis, simulation and visualisation; and
- high capacity and high quality network access to permit interaction with, and sharing of, diverse data and computing resources.

As discussed in earlier Chapters, there are currently a large number of disparate data repositories of varying standards and differing levels of accessibility. Over the next year or two, most universities are expected to have a digital data repository of some kind. It should be emphasised that most of these repositories will primarily focus on managing publications and related material, far fewer will attempt to manage research data. There are exceptions such as the BlueNet project (funded under SII), which is using the virtual hosting environment provide by the AODCJF to manage marine data sourced from academic institutions and discipline-specific data repositories such as ASSDA (discussed in the previous Chapter). In addition most government departments at Commonwealth, State and Territory level hold data in a variety of ways and some of the more advanced management mechanisms could be classified as repositories.

In developing a strategic, national approach to the management of data for science, it will be important to "build on existing data and information structures and services where it is advantageous to do so"²³ to achieve engagement from all elements of the research sector. Australia needs to agree upon and articulate a set of principles that characterise the type of repository networks we are attempting to establish as part of a national system. While this requires further debate, the Working Group considers that a national system should be established that:

- captures data that is of high relevance to specialist user groups, but which also has broader significance through its re-use in other communities of interest;
- provides hosting services for content that would not otherwise find its way into a repository;
- provides data access services, which may require user authentication for transactions;

²² Department of Communications, Information Technology and the Arts, pers. comm. Submission to the Working Group

²³ International Council for Science (ICSU), *ICSU Report of the CSPP Assessment Panel on Scientific Data and Information* 2004: p.7 [online] available from http://www.icsu.org/1_icsuinscience/DATA_Paa_1.html; accessed 23 August 2006

- supports an intuitive, single point of access discovery service for the entire network; one that does not require the user to understand the nature of the repository nodes in the network, nor have an intimate knowledge of the way in which the individual repositories classify their data;
- operates within legal, intellectual property and copyright bounds;
- uses, adopts and develops common standards that promote repository interoperability at both the technical and content level;
- provides tools that facilitate the creation of dataset documentation (metadata) and the deposition of both metadata and data into the repository;
- captures and provides metrics on data usage;
- provides permanent archival facilities for long-term data preservation; and
- conducts outreach, education and advisory services for data consumers and, importantly, potential data contributors.

The ICSU report notes that much historical data and traditional knowledge is in danger of being lost to science because this information is not available in digital formats.

The loss can occur because the data is paper based and paper decays or because the electronic format or software that is used to store them becomes corrupt or superseded.

The problem of loss of knowledge is major in the case of traditional knowledge – for instance relating to plants and foods which is so important for modern medicine and agricultural science. Because so much of this knowledge is passed down orally it is in danger of being lost as communities disperse and change and as languages die out.

The ICSU report states *‘Digitisation, data rescue, transcribing, and improved management of traditional or historical data are necessary to preserve these types of data for current and future scientific research. However the process of data recovery is expensive and often labour intensive and it requires trained personnel.’*²⁴

The report notes that there is a need for new and inexpensive methods of data digitisation and rescue that that this is an areas where the public and private sectors could work together to develop them.

To satisfy any goals that are set we must first fully understand the current information infrastructure landscape and take stock of the repositories that already exist. This will permit the development of future robust implementation models cognisant of what can be achieved with what we already have available to us.

Recommendation 3:

That the necessary policy and programmes be implemented with a view to establishing a sustainable publicly funded national network of federated digital repositories.

²⁴ International Council for Science. 2004. *ICSU Report of the CSPR Assessment Panel on Scientific Data and Information*. p.18.

Recommendation 4:

That the expert committee consider the development of a strategic roadmap for the implementation and evolution of the national network of federated digital repositories.

C. Data Management, Access, Sharing and Collaboration – Changing the Culture

‘...our current data management practices do not support easy access to data. As a result, the explosive growth in data means that we will grow a mountain of lost or unusable data unless we improve our data management practices and improve them quickly.’

Dr R Francis, NCRIS Platforms for Collaboration Facilitator²⁵

Scientific knowledge and data are of enormous importance in a global Information Society:

- *To foster innovation and promote economic development*
- *For efficient and transparent decision-making, particularly at government level*
- *For education and training*

Scientific data and information should be as widely available and affordable as possible: the more people that are able to share them, the greater the positive effects and returns to society. Scientific knowledge is a “public good”.

The development of new ICTs opens up unprecedented opportunities to ensure universal and equitable access to scientific data and information to enhance the global knowledge pool. However, excessive privatization and commercialization of scientific data and information is a serious threat to the realization of these opportunities for the benefit of society as a whole.

- International Council for Science (ICSU)²⁶

As a general proposition, publicly funded research should be publicly available. The Australian Government acknowledged this when it instigated the development of an Accessibility Framework for publicly funded research as part of *Backing Australia’s Ability*, to ensure that information about research and how to access it is available to researchers and the wider community.

The protocols needed to support data sharing must cover several key aspects. There is often an assumption that the barriers to sharing or accessing data are technological, and in many instances they are. More and more, though, the technological impediments to sharing or linking data are able to be overcome, as proposed in Recommendation 5. The protocols that are becoming more important and less tractable as the technology advances are around legal and policy frameworks for sharing data.

²⁵ Dr Rhys Francis, NCRIS Platforms for Collaboration Facilitator, submission to the Working Group, July 2006

²⁶ International Council for Science. 2004. *ICSU Report of the CSPR Assessment Panel on Scientific Data and Information*. p.40

As more organisations seek to use each other's data, agreeing on protocols for data access that can work between jurisdictions, sectors and discipline areas presents a complex problem. The different privacy and ethics regimes in the Commonwealth, States and Territories, varying policies relating to ethics at an institutional level and detailed considerations as to the possible identifiability of individuals within anonymised data all contribute to the difficulty in accessing data.

Projects such as the Molecular Medicine Informatics Model²⁷, based at the University of Melbourne and the WA Data Linkage System also for health data, have developed innovative solutions for the re-use of data which may have wide-scale application in health and similar research. However, to overcome the differences in laws, policies and approaches that face researchers every day demands a better national approach further addressed under Section D below.

Making data accessible

In addition, there are benefits for researchers and their institutions in making research data and research outputs publicly accessible. The impact of research is not felt until it is disseminated. Some universities, including the Queensland University of Technology for example,²⁸ are encouraging their researchers to deposit their research in the university's repository as a means of improving citation rates. Publication of research outputs in this way will also lead to greater discoverability of the datasets underpinning such research, opening up opportunities to share and link sets of data.

It is costly to collect data. If research has been funded to collect and organise a set of research data, it is more costly to recreate the data than to re-use it for different purposes. It's important to note that sometimes specific-purpose data does not exist, but relevant data, collected for unrelated purposes, can provide the required information. (For example, researchers looking for evidence as to the shrinking of the polar ice caps have used whaling records to examine data that is much older than the measurements of the ice caps themselves. Before the banning of commercial whaling in 1987, ships often moored close to the Antarctic pack ice while taking their catch, allowing researchers to retrace the extent of the pack ice by examining the positions of whaling ships).

Work by the National Health and Medical Research Council (NHMRC) to develop the *Australian Code for the Responsible Conduct of Research*²⁹, has included, among other issues, publication and dissemination of research findings and the management of research data and records. The Australian Research Council (ARC) plans to include a form of words encouraging wide dissemination of research findings in future funding agreements.

It is important, when considering the issues around wider access to data, to distinguish between knowing that a set of data exists and being able to access that data easily or at no cost. It is appropriate that a large number of interested parties and potential collaborators be aware of what research is occurring and what types of data are being gathered. Much fewer people should have access to be able to use the data, and this access should be governed by appropriate policies, processes and technologies to facilitate legitimate sharing of data, but prevent casual access. These issues are being considered in DEST's development of the Accessibility Framework for publicly funded research and should be incorporated into any national strategy for data for science.

²⁷ <http://mmim.ssg.org.au>

²⁸ Queensland University of Technology (QUT) *Measuring the research impact of your publications: citation indexes and alternatives, Strategies to increase citations to your publications* [online] available from http://www.library.qut.edu.au/subjectpath/citation_indexes.jsp#strategies; accessed 24 October 2006

²⁹ ARC, NHMRC, AVCC *Australian Code for the Responsible Conduct of Research – Second consultation Draft, February 2006*, Revision of the Joint NHMRC/AVCC Statement and Guidelines on Research Practice [online] available from http://www.nhmrc.gov.au/publications/_files/acrcr.pdf; Information about the Review process is available [online] at <http://www.nhmrc.gov.au/funding/policy/code.htm>

Much of the work in the 'open access' movement has focused on research outputs rather than research data.³⁰ However, while the issues are closely connected, it is the sharing of raw data which is the major emphasis here, rather than making available the reports, papers and theses which are the outputs of data. The latter are quite different and usually more readily available and accessible, and increasingly via electronic means. Whilst all data should be made available to researchers, scientists and policy makers it is the raw data that we want to make a special case for in this report. This includes data from government agencies or data collected from one source or from one study or data collection being made available for other purposes, re-use, secondary analyses etc. Hence health data collected from hospitals can be used to investigate the causes of cancers when linked to mobile phone use or occupational exposures; data collected by the Bureau of Meteorology can be used to investigate the effects of climate change over time by linking it to land degradation and so on. This can only happen if the raw data sets are made available and linked together to enable these complex analyses - this is what the ABS National Data Network is all about.

Some disciplines, including astronomy, have well-developed and widely accepted protocols covering the exclusive access to research data for defined periods. In other domains, data are never released; only the research findings – publications, conference papers etc. – are available for wider information. However, access to the building blocks of research and knowledge is as important as access to the refined knowledge products that research can produce, as it is likely to lead to more discoveries and useful outcomes.

A new approach to the management and integration of data for science will require a culture that recognises science data as a key organisational, national and global asset and that rewards the sharing and re-use of such data across disciplines and organisations. There are already a number of excellent examples within certain research communities where this has been achieved and a multi-disciplinary and global approach has been applied, in particular in the marine, atmospheric and astronomy science communities.

Such a culture, which implies greater openness and transparency among researchers, will be a challenge to many who may have grave concerns about any notions of making "their" data available to others. It is therefore all the more important to not only increase these scientists' awareness of the additional data and collaboration opportunities that may be available to them through data sharing and re-use, but also to reassure them that such initiatives will be governed by a strict adherence to intellectual property and copyright policies.

A related element of the culture that is needed to support the national strategic framework for scientific data is a commitment to best practice and continuous improvement in all data management activities. This will require both an organisational focus on such matters through adequate governance and auditing mechanisms and an awareness campaign among researchers that will highlight not only the vast potential of the collaborative use of data repositories but will also lead to an appreciation of and commitment to the principles and guidelines underpinning such activities.

Integrating good data management practices into existing research processes will, in many cases, constitute a significant change and will only be achieved over time and as the data management maturity of research organisations and individual researchers increases. However, it has been demonstrated in a variety of contexts that researchers and institutions are able to respond to changes in expectations, funding models and other elements of the overall policy framework.

³⁰ <http://www.eprints.org/openaccess/>

Financial support for data management

Financial support for data and information management should become a routine component in all research budgets and the evaluation criteria for assessing research funding proposals should include an evaluation of data management. The Working Group feels strongly that funding agencies should insist upon data management plans for all grants, tenders and other funded activities which involve data capture or data generation. These plans should encompass proposals for long-term data archival and explain how the data will be made available for secondary use. There is also a need for granting bodies to appropriately value proposals which include research that will utilise existing data (as discussed in the previous Chapter).

Recommendation 5:

That standards and standards-based technologies be adopted and that their use be widely promoted to ensure interoperability between data, metadata, and data management systems, providing authentic users of the data with appropriate processes and safeguards.

Recommendation 6:

That the principle of open equitable access to publicly-funded scientific data be adopted wherever possible and that this principle be taken into consideration in the development of data for science policy and programmes.

As part of this strategy, and to enable current and future data and information resources to be shared, mechanisms to enable the discovery of, and access to, data and information resources must be encouraged.

Recommendation 7:

That funding agencies offer incentives to encourage researchers and institutions to:

- develop data management plans for each research grant application involving data collection and generation, and that standards be made freely available and widely disseminated so as to encourage best practice in data management;***
- introduce policies and practices to encourage collaboration and sharing of data across Australia's scientific research institutions and across agencies; and***
- analyse and re-use existing data.***

D. Ensuring there are no Regulatory Impediments

Ethics and Privacy

In the earlier section on discussing issues around data access and integration, there was passing mention of the issues of ethics and privacy. These questions present themselves most commonly in the areas of population, health and welfare research, where identifying individuals would constitute a breach of their privacy and conducting some types of research may be considered unethical. As part of developing a national strategic framework for data for science,

the high level committee should determine the impact of privacy, and ethical concerns and the impact of these critical issues on specific data integration activities. If the sample sizes in a population health study using current disease data are too small, even in de-identified datasets, anonymous subjects may be able to be identified. Researchers and institutions need both to comply with privacy regulations and be able to negotiate their way through the complex rules that ensure sufficient safeguards are in place, to ensure such identification and other unacceptable practices do not occur.

Two recent documents from the UK, *Better use of personal information: opportunities and risks*³¹ and *Report on Personal Data for Public Good: Using health information in medical research*³², make strong recommendations to encourage the use of personal data sets as a major way of improving public services (such as health, education, employment and family services).

Some of the Recommendations from the UK Council for Science and Technology are:

We consider that there are major benefits to be delivered from developing better linkages between, and wider access to, personal datasets provided the risks are carefully assessed and managed. In order to realise these benefits there is a need to:

- *engage in dialogue with the public and stakeholders on the full range of benefits and risks, in particular to individual citizens as well as to society and to government;*
- *carry out risk analyses with a balanced approach to the benefit:risk equation to strengthen the evidence base for policy formulation and enable improved service delivery;*
- *have a focal point within government to think through, plan and coordinate joined-up working across different datasets.*

*Governments should put in place mechanisms to ensure that the linkage and use of personal data sets is achieved in a much more coordinated, coherent, and transparent way across the public sector. We consider that this should be done on the basis of three linked principles which relate to i) that personal data must be anonymised where-ever possible; ii) access to data should be facilitated where that access is for research or statistical purposes and iii) appropriate safeguards and transparent processes should be in place before personal data can be accessed and used.*³³

Recommendation 8:

That funding agencies such as the NHMRC and ARC ensure that best practices and policies are developed and followed that allow bona-fide researchers to access individual population data, including the integration and linking of data from multiple sources, whilst protecting privacy, and ensuring that ethics committees fully understand these policies and their rationale.

Intellectual Property Rights

Issues surrounding the impact of intellectual property (IP) rights on access to and use of data have been examined at the national and international level. In the digital age it can be IP rights that impede access to data once it is known and located.

³¹ Council for Science and Technology (UK) *Better use of personal information: opportunities and risks* (2005)

³² Academy of Medical Sciences (UK) *Report on Personal Data for Public Good: Using health information in medical research*. (January 2006)

³³ Council for Science and Technology (UK) *Better use of personal information: opportunities and risks* (2005)

In particular, the OECD Committee for Scientific and Technological Policy (CSTP) Ministers meeting in 2004 recognised that fostering broader, open access to and wide use of research data will enhance the quality and productivity of science systems worldwide. They adopted a *Declaration on Access to Research Data from Public Funding*³⁴, asking the OECD to take further steps towards proposing *Principles and Guidelines on Access to Research Data from Public Funding*, taking into account possible restrictions related to security, property rights and privacy.

The Australian Academy of Science cited this work in its written submission, suggesting that the Working Group may wish to examine the extent to which Australian facilities are complying with the OECD Declaration.

The Ministers (in the OECD declaration) also noted that coordinated efforts at national and international levels are needed to broaden access to data from publicly funded research and contribute to the advancement of scientific research and innovation.³⁵

The International Council for Science (ICSU) report argued that “recent trends towards the appropriation of data ... pose serious obstacles to full and open access to data for scientific purposes.”³⁶ ICSU identified the problem of researchers holding intellectual property rights in **data**, rather than in the **discoveries** made from the data, as an impediment to innovation and the sharing of knowledge.

As the Defence Science and Technology Organisation (DSTO) point out in their response to a draft of this report:

‘...the Australian research community in recent years has been under significant pressure to commercialise. In order to prove potential commercial opportunities scientists need real data. Therefore if scientists share their data they are making the task of replicating their research easier for their fellow research competitors which explains today’s non-sharing culture. Thus continued pressure to commercialise may prove counter productive to the proposals made in this report.’³⁷

Paradoxically, if data were more freely available its inherent value would increase, for Australia as a nation and for the rest of the world, as its re-use would lead to more scientific discoveries.

The Bureau of Meteorology points out that:

‘... An important experience of the meteorological community is that foregoing proprietary rights to data and making them freely available actually benefits the individual as well as the community at large – it attracts scientists to your area and accelerates progress over and above that which you could achieve without data sharing’

³⁴ The governments (including the European Community) of Australia, Austria, Belgium, Canada, China, the Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Luxembourg, Mexico, the Netherlands, New Zealand, Norway, Poland, Portugal, the Russian Federation, the Slovak Republic, the Republic of South Africa, Spain, Sweden, Switzerland, Turkey, the United Kingdom, and the United States, *Declaration on Access to Research Data from Public Funding*, 30 January 2004 in Paris [online] available from <http://www.codataweb.org/UNESCOmtg/dryden-declaration.pdf>; document provided in the Academy of Science submission to the Working Group. Full text available at Annex 1

³⁵ Organisation for Economic Co-Operation and Development (OECD) *Principles and Guidelines on Access to Research Data from Public Funding*, Science, Technology and Innovation for the 21st Century. Meeting of the OECD Committee for Scientific and Technological Policy at Ministerial Level, 30 January 2004 – Final Communiqué [online] available from:

http://www.oecd.org/document/0,2340,en_2649_34487_25998799_1_1_1_1,00.html; accessed 4 November 2006

³⁶ International Council for Science (ICSU), *ICSU Report of the CSPR Assessment Panel on Scientific Data and Information 2004*: p.10 [online] available from http://www.icsu.org/1_icsuinscience/DATA_Paa_1.html; accessed 23 August 2006

³⁷ Defence Science and Technology Organisation, pers. comm. submission to the Working Group

The International Council for Science in their report *Scientific Data and Information* note that:

*'Scientific data and information are increasingly considered both as input to research, decisions, policy, and management, but also, like digital data in entertainment and commerce, property in and of themselves. Science has well been served by a system of minimal restraints (e.g., those based on privacy considerations) on access to and use of data, such as genetic information, and the protection of databases under sui generis regimes, as well as limitations to the fair use of digitized data (e.g. anti-circumvention measures) pose serious obstacles to full and open access to data for scientific purposes.'*³⁸

The ICSU's recommendation number 39 states:

*Governments, and other bodies concerned with international and national policy development, should ensure that IPR (Intellectual Property Rights) legislation recognizes the value of ensuring full and open access to data for scientific research and education purposes.'*³⁹

Our Working Group discussed how the freely available data from the human genome has enabled the enormous progress on how genes work in a range of human developmental and disease processes. What would progress have been if the genome had been privately owned and controlled by IPR? The costs of accessing the information would have dramatically decreased our ability to participate in this important research and much less research would have been done internationally.

The OECD CSTP has made progress on the principles and guidelines mentioned above. The Working Group considers that more work remains to be done to understand the implications of these guidelines for Australian data for science.

There may also be a need to consider a wider range of regulatory factors that impact on the conduct of research and the treatment of research data including archives and records legislation, national security, confidentiality, conflict of interest and research misconduct.

Recommendation 9:

That in the context of developing the strategic framework for scientific data management, Australia's intellectual property approaches be checked to ensure they do not impede the sharing of data.

In particular, it should take into account the OECD Committee for Scientific and Technological Policy guidelines on access to research data and the International Council for Science statements about the benefits of sharing data.

E. Skills for Data Management

The e-Research Coordinating Committee examined in detail the skills issues impacting on greater online data sharing and collaboration by researchers. The Committee found that adoption of e-research methodologies in Australia is constrained by a shortage of skills in e-research methodologies, including information management and curation skills.

³⁸ International Council for Science. 2004. *ICSU Report of the CSPR Assessment Panel on Scientific Data and Information*. p.10

³⁹ *ibid.*, p.10-11

The e-research Committee's Interim Report⁴⁰ found that "e-research skills include cross-disciplinary skills that bridge and cover both the relevant research area and ICT," but that "the limited number of opportunities and the finite period of time associated with a research grant work against the development of career systems and sufficient rewards for ICT professionals working in the support of e-Research."

They also reported that ICT professionals working on e-research projects can be relatively isolated from their peers and, in the absence of wider adoption of e-research, may have limited subsequent employment opportunities elsewhere. Overseas experience highlighted a need to move away from the dominance of a traditional disciplinary focus to encourage multi-disciplinary collaboration and education.

The need for cultural change to more fully integrate a collaborative, multi-disciplinary, digital way of working was also reflected in a widespread lack of awareness by researchers of the potential of new technologies to improve their research practices, particularly through data mining and sharing. For example, the Committee noted the experience of the Grangenet projects, which highlighted the need for extensive awareness and communication activities over two years to make researchers aware of the facility, and to support them in exploring the collaborative potential of new technologies.

Providing training to all researchers was seen as a critical part of addressing the need for cultural change, and the e-Research Committee identified three groups of skills that need to be developed as a basis for advancing e-research in Australia. Researchers need:

- to acquire basic level e-research skills and need easy and structured ways of acquiring those skills;
- ongoing (day-to-day) support from ICT professionals to assist with more complex ICT problems; and
- high level ICT and information management professional support to make efficient and effective use of available data resources.

Data management is recognised as a key element of the skill set needed by all researchers. For example biologists researching genetic influences need mathematical skills themselves or to work closely with those who have them. While some e-research skills can be incorporated into higher education and training, others will be developed 'on the job', through exposure to working with experienced professional support staff. The most important issue here is to ensure that researchers and others who use data in any scientific discipline are better supported in terms of modern data management practices, either by being up-skilled themselves or by working with specialists or within specialist groups who can provide the skills, use the technology and ensure access to the latest supportive infrastructure.

The skilling of researchers and their subsequent adoption of e-research methodologies (and the flow-on of positive outcomes to society) are dependent on the availability of complementary professional ICT and data management skills. The Committee found a strong need for both skills development [and professional recognition] for ICT specialists and information managers.

The UK e-Science Programme has shown that skills development needs to be undertaken early on in an integral way, to produce multidisciplinary teams, capable of working effectively in the virtual, collaborative e-research environment. Such overseas experience also demonstrates a need to increase the number of advanced ICT graduates emerging from the higher education system, which means we need to excite more school leavers into these vital areas.

⁴⁰ Department of Education, Science and Training (DEST) and the Department of Communications, Information Technology and the Arts (DCITA), *The e-Research Coordinating Committee's Interim Report* September 2005: Chapter 4, p14-17 [online] available from <http://www.dest.gov.au/NR/rdonlyres/B6F765A7-DD2C-432B-9064-2F9CD4E17E66/10518/InterimReport2.doc>; accessed 1 May 2006

The Working Group notes the skills gap in the area of qualified multi-disciplinary scientists and engineers, and in particular notes the need for suitably skilled Data Scientists able to provide the crucial linkages between data management skills and domain expertise.

The *Report of the Australian e-Research Sustainability Survey Project on Sustainability Issues for Australian Research Data* (October 2006)⁴¹ found that 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.”⁴² The report also noted that it would “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.”⁴³

There are a number of approaches that could inform a strategy to address the lack of necessary skills in the research sector. Work should be done to chart the path from minimum to optimum maturity in key areas of science-related data management from an organisational and individual's perspective. This information should be used to inform a high level science data management curriculum.

An analysis of training needs could identify the focus areas and scope of the data management curriculum for specific target groups with the aim of adequately addressing the different requirements of scientists, data managers, ICT support staff and other key parties involved in activities related to the development and/or use of science data systems.

Together, these assessments of data management maturity and training needs could be used to inform the development of relevant curricula for graduate and post graduate educational content in science, ICT and information management disciplines.

Recommendation 10:

That data management expertise becomes a core skill for researchers, including graduate and postgraduate science students across all disciplines, and that they receive data management training as part of their education.

Recommendation 11:

That the Australian Government give early consideration to the findings of the e-Research Coordinating Committee regarding changing research behaviour, practices and skills.

Comment on a National Centre for Data for Science

The Data for Science Working Group discussed at length the idea of a new National Centre for Data for Science. There was considerable support within the Group for a Centre; it was felt that such an initiative would be of benefit and may be a useful mechanism for progressing many of the above recommendations.

⁴¹ Markus Buchorn and Paul McNamara, *Sustainability Issues for Australian Research Data: the Report of the Australian e-Research Sustainability Survey Project*, APSR Publications, October 2006 [online] available from <http://hdl.handle.net/1885/44304>; accessed 4 November 2006

⁴² *Ibid.*, p.8

⁴³ *Ibid.*

The Working Group considers that there is a range of functions that a Centre could assist with, including:

- facilitating and promoting the changes reflected in the recommendations;
- working with those in specialist scientific disciplines to discover datasets;
- establishing vital repositories; and
- working collaboratively with the research, government and business communities to support the proposed new approaches to data for science.

The Working Group stopped short of recommending the establishment of a Centre. The Working Group concluded that the high-level expert committee (recommendation 2) should decide whether such a centre was desirable and, if so, where it may be hosted, and what its role and governance mechanisms should be.

Chapter 4: Possible National Initiatives

Chapter overview

The aim of this Chapter is to present two important initiatives that would benefit greatly from the recommendations in this report being implemented. The national drug research and policy initiative highlights how novel data linkages could lead to lives being saved by the early detection and avoidance of adverse drug effects. The national environmental reporting initiative demonstrates how better information about the environment, and especially water, can lead to an improved understanding of better management of scarce resources and can assist decision-making about where to spend money to the best advantage.

Opportunities for Change

As outlined in Chapter 1, optimal sharing and re-use of scientific data can help provide answers to the many complex problems and challenges faced by Australia and the world. In making its recommendations, the Working Group considers it important to illustrate what is possible if scientific data were used, linked and managed in a more coherent way. To this end, the Working Group has identified two national initiatives which present exciting opportunities to transform the way health and environmental data, respectively, are coordinated and managed on a national level. These initiatives are:

- the National Drug Research and Policy Initiative; and
- the National Environmental Reporting Initiative.

The cases put forward for these initiatives highlight the opportunities Australia has missed in pharmaceutical and environmental reporting, respectively, because of the current lack of national coordination and problems caused by data not being available, inconsistent data or lost opportunities for data linkage. These initiatives also present a way forward to address such data management issues.

Note:

Pharmacovigilance is defined as the detection, assessment, understanding and prevention of adverse effects, particularly long-term and short-term side effects, of medicines. It is gaining in urgency and importance for doctors, researchers, pharmaceutical companies and health department committees which advise on drugs, as the number of stories in the media of unexpected drug effects resulting in recalls and increases in expensive litigation. (Source: The Importance of Pharmacovigilance, WHO 2002, accessed at http://www.google.com.au/search?hl=en&lr=&defl=en&q=define:Pharmacovigilance&sa=X&oi=glossary_definition&ct=title)

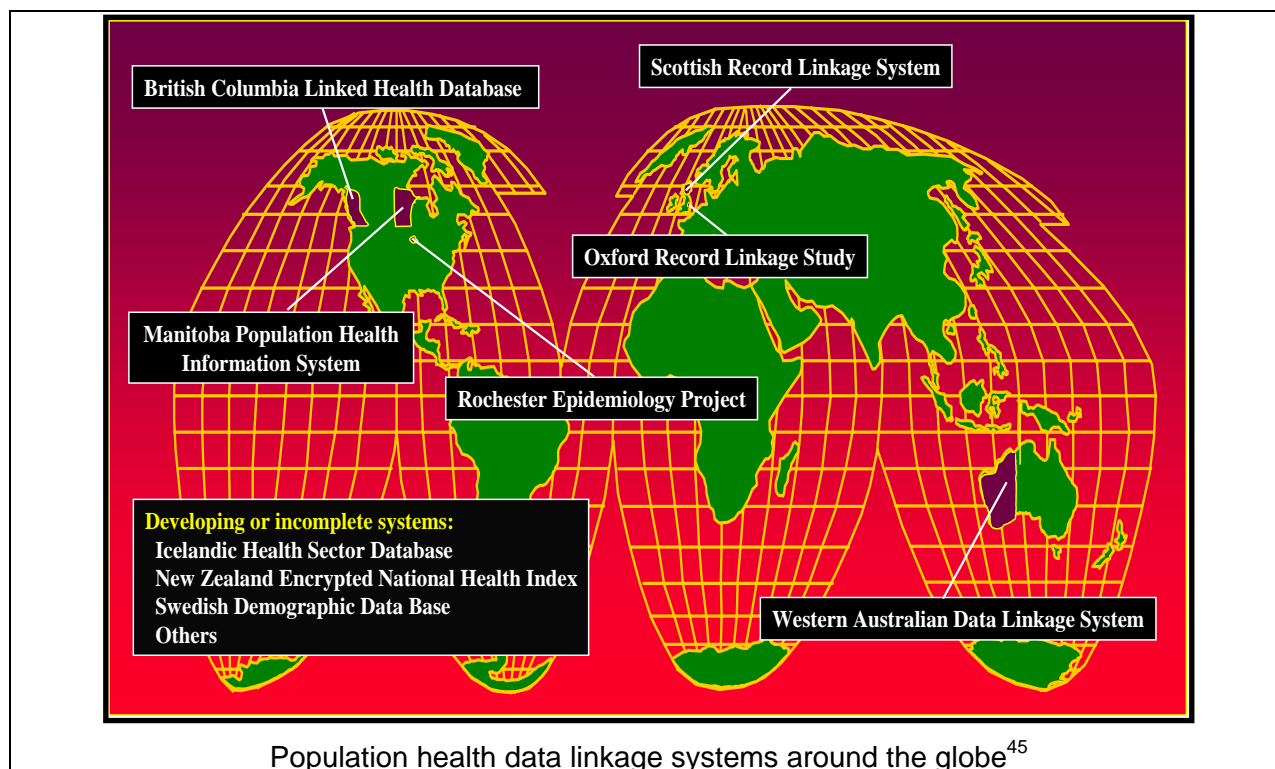
The Case for a National Drug Research and Policy Initiative (Pharmaco-epidemiology/ Pharmoco-vigilance)

Introduction

This initiative could potentially save thousands of lives and billions of dollars through avoiding preventable complications, litigation, unnecessary hospitalisations, deaths and disabilities. It would also ensure Australia has the best data in the world for a vast range of pharmaco-epidemiological research, the basis for pharmaco-genomic research and new drug development.

The availability of data bases which record all health outcomes for a total population of people is rare and exists in few centres internationally. In the Western Australian Data Linkage System (WADLS)⁴⁴ data has been available, for many years, on all deaths, hospitalisations, cancers, birth defects and other selected disabilities, cardiovascular diseases (such as heart attacks and strokes) and all emergency department visits. With the recent capacity to link these data to all prescription data via the Pharmaceutical Benefits Scheme (PBS), an exciting and innovative opportunity exists to create a major initiative for pharmaco-epidemiology to serve a national and possibly international agenda for ensuring the best and safest use of drugs.

Figure 2:



Why do we need this initiative?

The tragic, unforeseen effects of drugs, such as thalidomide⁴⁶ and DES⁴⁷ in the past and the most recent incidents caused by Vioxx⁴⁸, clearly illustrate the need for more accurate information on the safety and effectiveness of pharmaceuticals after they have been marketed. If the WADLS had had an active and working link to PBS data in the late 1990s with automatic review of major drugs in place, thousands of deaths and heart attacks caused by Vioxx could have been prevented.

There are other reasons why we need to access data on drugs and their effects in the “real world” of population-linked databases. Randomised controlled trials to investigate both efficacy

⁴⁴ <http://www.publichealth.uwa.edu.au/welcome/research/dlu/linkage/system>

⁴⁵ Brook, E personal communication with Professor Fiona Stanley 2006

⁴⁶ Lenz W. *Thalidomide and congenital abnormalities*. Lancet 1962;1:45

⁴⁷ Bibbo M, Haenszel WM, Wied GL, et al. *A twenty-five-year follow-up study of women exposed to diethylstilbestrol during pregnancy*. N Engl J Med 1978;298(14):763-767;

Giusti RM, Iwamoto K, Hatch EE. *Diethylstilbestrol revisited: a review of the long-term health effects*. Annals of Internal Medicine 1995;122(10):778-88;

Bishun NP, Smith NS, Williams DC, Raven RW. *Carcinogenic and possible mutagenic effects of stilboestrol in offspring exposed in utero*. Journal of Surgical Oncology 1977;9(3):293-300

⁴⁸ McGettigan P, Henry D. *Cardiovascular risk and inhibition of cyclooxygenase: a systematic review of the observational studies of selective and nonselective inhibitors of cyclooxygenase 2*. JAMA 2006;296(13):1633-1644

and safety in humans before drugs are marketed can never uncover all untoward effects. Once drugs are approved, there are few regulations which prevent doctors prescribing drugs for those groups of people not included in the trials – particularly pregnant women and, increasingly, children.

For example, psychotropic (mood-altering) drugs are now being prescribed increasingly for children, on whom their effects have never been tested.⁴⁹ They may not be effective for children's illnesses and they may cause considerable harm, either immediately, or interfere with vital developmental processes and result in later problems. It is likely that this is happening, but we are currently not monitoring this nor detecting the harm that may be done.

There is also a worldwide trend of doctors prescribing drugs that have been trialled and marketed successfully for one disease or symptom for another disease or symptom for which there is no evidence of efficacy and as a result may indeed cause harm. Some of these so-called "off-label" drugs have caused thousands of deaths in people who should never have been given them in the first place.⁵⁰ And with the ageing of the population, a related matter is the concern about "poisoning" by polypharmacy and over-prescription to the aged and its effects, namely more disease, more accidents and falls with further costs to the health-care system. In Western Australia nearly 5% of hospital admissions (80 000 per year) are due to prescription drug related problems and cost the health system about \$360 million. This is likely to be an under-estimate as this data is not yet based on total population information and thus exclude out-of-hospital episodes.

What would this drug initiative produce?

There is an urgent and exciting opportunity for Australia to develop a world leadership capability in pharmaco-epidemiology and "real world" safety and effectiveness of drugs by building on our unique datasets in WA, expanding them nationally and making them, with strict privacy regulations, linkable to PBS. With appropriate routine and special analyses, all new and existing drugs could be evaluated to enable us to properly meet nationally stated aims as required from the various national responsible organisations (*DOHA National Medicines Policy, 2000*; Pharmaceutical Health and Rational Use of Medicines, DOHA 2004; Therapeutic Goods Administration; Adverse Drug reactions committee; Australian Drug Evaluation Committee etc).

The evaluation of the safety and efficacy of drugs once marketed and sold, is called "post-marketing drug surveillance". Currently it depends on alert clinicians and patients reporting occurrences, and is sporadic and inaccurate. This means that our national committees currently conduct decision making on drug availability, usage and adverse effects with woefully inadequate data.

When fully implemented this drug initiative will help us find out more about:

- a product's safety;
- the therapeutic effects of drugs – how effective they are and if they are effective in all population groups;
- drug interactions and their safety/adverse effects;
- the therapeutic use of a drug – whether the drug is being prescribed appropriately and whether it is working;
- drug utilisation/uptake – who is getting it and where (e.g. rural vs urban);
- new drugs replacing old ones – why, how and what are the costs and consequences;
- trends and patterns in prescribing drugs;

⁴⁹ Bramble D. Annotation: *The use of psychotropic medications in children: a British view. J Child Psychol Psychiatry* 2003;44(2):169-179

⁵⁰ New Scientist, *Dicing with Death*, 29 July 2006

- whether we are getting value for money from our publicly funded drug expenditure.

What are the privacy issues?

There are considerable concerns about the use of a person's health data and the threat that the availability and misuse of such data may harm that person. With the establishment in 1994 in Western Australia of the Data Linkage Unit (DLU)⁵¹, the linkage of such data is conducted in a way that protects privacy and ensures confidentiality⁵². Australia does not yet have unique numerical identifiers as in Scandinavia; in the DLU, the individual identifying information is linked across records without any of the sensitive health information. The researcher or policy analyst, after ethics committee approval and following national guidelines⁵³ then obtains a linked file of the sensitive information (e.g. drug exposure and cancer occurrence) without any individual identifying information. The risk of any privacy breach is negligible and has not eventuated in 30 years of similar linkages in WA⁵⁴.

A recent UK report on privacy and use of medical records for research⁵⁵ suggested that privacy concerns had recently over-ridden those of public-good research and resulted in environments which prevented or discouraged such research. Australia needs to ensure that we continue to maintain our unique "niche" position in health record linkage to enable this important work to continue and to ensure that privacy is protected – a "win-win" for all who need safe and effective health care.

International activities – does Australia really have a niche?

Most international activity in pharmaco-epidemiology is within the USA, Canada and the UK with all jurisdictions concerned about drug safety, reducing the cost of drugs and ensuring efficacy and appropriate prescribing. Few countries have total population health outcome data linked to all prescriptions, as proposed here. The Federal Drug Administration (FDA) in the US has put out collaborative research agreements (CRAs) and centres for education and research on therapeutics (CERTs) with special requests for rapid linkage to large datasets which are geographically and demographically diverse and have longitudinal databases to enable outcomes and adverse drug effects to be measured. Health Maintenance Organisations and the Veterans Health service were the only groups for which some data were available, but no jurisdiction in the United States has complete population data on outcomes and complications linkable to drug exposure data. In the UK, there is a data linkage project in Tayside in Scotland with linkage of prescriptions to hospitalisations but the field is too small to enable precise estimates (only 400 000 people). A feasibility study is being planned to link data for all of Scotland (5.1 million people). The UK's GP Research database on 9 million patients forms the UK's pharmaco-vigilance programme – but is based only on GP consultations, with no linkage to hospitalisations for more serious illnesses and deaths.

Would the Pharmaceutical Industry welcome this initiative and participate?

Two of the largest pharmaceutical companies worldwide have been consulted about this initiative and have both indicated that they would be extremely pleased to have such a centre established anywhere and would participate fully to ensure drug safety and monitoring of

⁵¹ Holman CDJ, Bass AJ, Rouse IL, et al. Population-based linkage of health records in Western Australia: development of a health services research linked database. *Aust N Z J Public Health* 1999;23:453-459; Brook EL, Rosman DL, Holman CDJ, Trutwein B. *Summary report: research outputs project, WA data linkage unit (1995-2003)*. WA data linkage unit, Department of Health 2005

⁵² Trutwein B, Holman CD, Rosman DL. Health data linkage conserves privacy in a research-rich environment. *Ann Epidemiol* 2006;16(4):279-280

⁵³ National Health and Medical Research Council (NHMRC) Section 95A privacy act

⁵⁴ Brook, E personal communication with Professor Fiona Stanley 2006; Brook EL, Rosman DL, Holman CDJ, Trutwein B. *Summary report: research outputs project, WA data linkage unit (1995-2003)*. WA data linkage unit, Department of Health 2005

⁵⁵ Academy of Medical Sciences (UK) *Report on Personal Data for Public Good: Using health information in medical research*. January 2006

therapeutic efficacy. It would also help “big Pharma” to know about the extent of uptake of drugs, patterns and trends in prescribing, and the effectiveness of drugs in groups for whom they have little data, such as children.

Are there any examples of this kind of initiative on a smaller scale already happening?

It has already been demonstrated that Australian researchers have the capability to develop the required underlying statistical and mathematical methods for pharmaco-vigilance based on administrative health datasets. The proposal has therefore been shown to be feasible. The WADLU has already several projects which are linking PBS data to outcomes in Western Australia (pharmaceutical prescriptions in aged care; drugs in pregnancy and birth defects). The CSIRO has linked PBS data to the Queensland hospital morbidity data to investigate patterns of utilisation, adverse events and other health outcomes.

Vision

If a national framework for scientific data management as suggested in this report is adopted, for this area of health this would support the development of an effective national drug research and policy initiative.

A committee with representatives of the relevant groups within DOHA and NHMRC, CSIRO, the West Australian Data Linkage Unit and Data Linkage Australia (a new state Centre of Excellence), the States and Territories, pharmaceutical, clinical and epidemiological researchers, privacy experts, consumers and industry could be formed which would then:

- develop a business plan and funding strategy for this initiative;
- evaluate the capacity (for example, size of population, adequacy of data including not only the PBS data but information on other explanatory factors, and how representative of other jurisdictions) and feasibility of using the current WA Data Linkage System;
- evaluate the availability and completeness of data on health outcomes in national and state databases and the Australian Institute of Health and Welfare (AIHW), to build capacity towards a national centre; and
- identify the first projects, of highest national priority, to be undertaken.

The Case for a National Environmental Reporting Initiative

The State of the Environment 2006 report indicated that in 2002–03 the annual spend was over \$12 billion on the environment. That figure has now probably doubled to \$24 billion per annum and is increasing. This indicates the significance of environmental sustainability for Australia's well being.

The State of the Environment (SOE) reporting system offers the use of indicators of environmental performance ranging from changes in vegetation condition and extent, river flows and water quality, the condition of our air, land, soils, cultural and natural heritage and biodiversity in order to inform and target expenditure.

However, reasonably comprehensive data is only available for 35.7% of these indicators.

Environmental Reporting

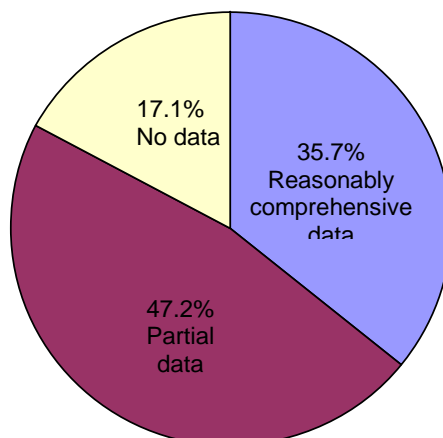
Australia has a leading position in the world in its recognition of the importance of the environment, and the need to measure environmental performance and to invest suitably in a sustainable environment.

We have been world leaders in, for example, the early measurement of climate change with the pioneering work of CSIRO at Cape Grim. At the other end of the scale, we are also world-leading in taking a comprehensive survey of the State of the Environment on a regular basis. Indeed SOE reporting is legislated and involves considerable cooperation between various levels of Government and covers the whole continent.

There is however a major problem: the SOE report cannot give a comprehensive national picture of the State of Australia's Environment because of the lack of accurate, nationally consistent environmental data.

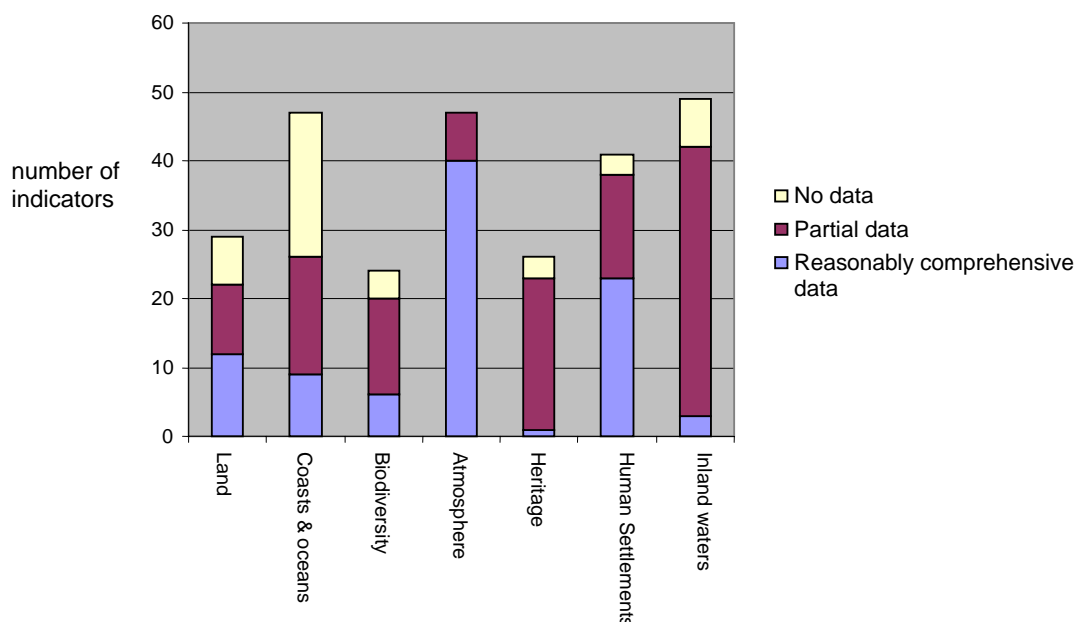
The SOE 2006 report is based on 263 environmental indicators under eight themes: Land; Coasts and Oceans; Biodiversity; Atmosphere; Natural and Cultural Heritage; Human Settlements; Inland Waters; and Antarctica and Other External Territories. The indicators were selected as being potentially capable of being populated by the SOE committee following an indicator audit in 2004 when several hundred other indicators were rejected as impracticable at the time. The overall picture (Figure 3) shows we still have a long way to go.

Figure 3. National data coverage across 263 indicators (SOE 2006):



As shown in figure 3, above, there exist useful national data for only 35.7% of the chosen indicators. We have some data for 47% of indicators and no data for the remaining 17%. Of the eight themes, five (Land, Biodiversity, Coasts and Oceans, Inland Waters, and Natural and Cultural Heritage), lack data needed on more than half the indicators to make a comprehensive national assessment (see Figure 4). This is clearly an issue.

Figure 4. Indicator coverage SOE 2006 in seven of the eight themes :



The Solution – a National Environmental Reporting Initiative

In order for Australia to have the capacity to accurately monitor and assess the condition of the environment on an ongoing basis, not just to make an informed approximation every five years through the State of the Environment report, the SOE Committee needs to be able to provide data interpretation and commentary, using comprehensive, accessible, up-to-date, relevant national data. Indeed, the SOE report suggests that environmental data should be continuously updated and made publicly available on the Internet.⁵⁶

With such capacity, we could more clearly illustrate returns on investment in the environment and changes to environmental governance, we could also achieve more efficient targeted use of government funds and greater leverage of private sector inputs, including capital, information and knowledge, to better integrate production systems with natural resource management.

To achieve this, Australia needs a properly funded, comprehensive and enduring system for all nationally significant research and management of relevant data – not least, and not only environmental data. Science can not inform innovation and management without relevant data.

⁵⁶ <http://www.deh.gov.au/soe>

A Further Example: Australian Water Resources

Australian Water Resources 2005 is the National Water Commission's baseline assessment of water resources for 2004–05 (the first year of the National Water Initiative). The report synthesises information at the national scale to increase the understanding of Australia's water resources, and identify knowledge gaps that reduce Australia's ability to manage these resources effectively and sustainably.

On a scale of A (complete coverage) to E (inadequate, no coverage) the average ranking is between C (Adequate – reasonable coverage, some gaps or limitations) and D (Poor – patchy coverage, significant gaps or limitations)⁵⁷

The Australian Water Resources assessment identifies a number of gaps in the consistency, extent and availability of information to provide for robust resource management, these gaps include the:⁵⁸

- inability of some states and territories to provide information in a readily-accessible manner;
- lack of systems to support information integration, resulting in problems with data collation;
- lack of consistent data standards, terminologies, attribute descriptions and, in addition, highly variable methods for determining them;
- information about actual water use and levels of resource development is highly variable across Australia, making accurate reporting and assessment difficult and error prone; and
- lack of integration and coordination within the jurisdictions that address different issues (such as water availability and river health).

Through the National Water Commission (NWC), the Australian Government is investing \$2 Billion in reforms to water planning, regulation and trading in an environmentally sustainable framework. If the overall water reform process is to be transparent, credible and evidence based, having a robust scientific foundation is critical. Accordingly, better information on our water resources is seen as having seminal importance.

The way forward – How to Achieve a National Environmental Reporting Initiative

Australia's environmental capital is as significant as its other assets, yet, as indicated above, the data needed to effectively manage the environment is lacking.

Australia must build its capability to live with its environment and respond appropriately to changes in that environment. An adaptive approach to appropriate investments in the environment requires good decision making information. Cooperation across all levels of governance is critical to obtain such information. This cannot occur without consistent, time-series measures of appropriate environmental indicators. Ready availability of knowledge that is accumulated through national expenditure on environmental research is also crucial.

One of the root causes of the current challenge is that data is dispersed through many different levels of government, departments, authorities and companies. There are many databases and the access to these databases is variable at best. The longevity of the data is likewise questionable in many cases.

⁵⁷ http://www.water.gov.au/Keymessages/StatusOfInformationIn2004_05/index.aspx?Menu=Level1_1_2

⁵⁸ http://www.water.gov.au/Keymessages/InformationAndSystemGapsLimitations/index.aspx?Menu=Level1_1_9

The view of the Working Group with regard to the significance of scientific data management is aligned with the view of the SOE 2006 Committee and other sources of advice to the Australian Government. Australia needs a properly funded and enduring system for nationally significant research and management data. Such a system is required to advance Australian science and draw scientifically valid inferences from temporal datasets.

A truly national system would require cooperation between the Department of Environment and Heritage; the Department of Agriculture, Fisheries and Forestry; the Australian Bureau of Statistics; the National Land and Water Audit and many other Commonwealth, State and Territory instrumentalities.

SOE 2006 has built a data reporting system and linked it to other systems. This has created a basis for Australia to adopt an enduring environmental reporting system that has the potential to track changes in environmental pressures, conditions and responses. However, this is only a beginning and does not, for examples, include a system for data aggregation.

CSIRO's Water Resource Observation Network (WRON) has begun to address, for water, issues that are being faced across most areas of environmental data reporting.

Case Study: CSIRO's Water Resources Observation Network

WRON commenced in 2006 with an initial \$9 million investment to establish a technology platform for a water information system. Areas of research include work in standards development for interoperability, next generation web technologies, model integration, sensor networks, reporting, analysis and forecasting tools for use at multiple scales across the continent. Within the WRON framework, these reporting tools will integrate actual and modelled water resource information from the past, present and predicted future.

Figure 5



Such a framework would clearly be relevant to other areas of natural resource management and environment reporting more broadly, but would require a high level of coordination to ensure interoperability across domains is not compromised.

Other initiatives are under way involving the National Land and Water Resources Audit (DAFF and DEH), the Commonwealth Environmental Research Facilities Programme (DEH) and the National Collaborative Research Infrastructure Strategy. Each has a different purpose but ultimately all are a form of data generation and are data dependent. In addition, the reviews of National Research Priority (NRP) reports by the reporting agencies to the NRP Standing Committee make it clear that much work is done and collaboration is improving. However, no data protocols exist outside localised agreements or where a specific project is involved. These are often international in scope, for example climate change research and policy. It would not be unreasonable to suggest a data management NRP in due course.

The SOE 2006 Committee has brought the environmental data problem to the attention of the Natural Resources Management (NRM) and the Environmental Protection and Heritage Standing Committees, the relevant ministers and departmental secretaries. Currently a working party of the two standing committees is investigating a national environmental reporting system. A strong direction from the Australian Government would ensure the timeliness and success of this activity.

Vision – Creating a National System

There are many routes to a national environmental system as a component of national science data management. All will require a mandate and a cooperative approach because of Australia's constitutional arrangements.

The objective of an environmental reporting system is to identify and assess any change in the environment. In addition it should assist in establishing if these changes are a result of human interventions or natural causes. This will enhance environmental decision making based on sets of time-series data that are both temporally and spatially consistent.

The characteristics of a national environmental reporting system are:

- the identification of nationally important datasets;
- the commitment of data custodians to the continuous collection of data in a nationally consistent way;
- appropriate data aggregation and management protocols through a single point;
- agreed principles for data sharing ;
- a span across all national SoE themes;
- data custodianship remaining with data providers ; and
- the ability to report on different scales.

The Cost of Inaction and the Benefits of Action

If unattended the problems associated with the deficit in environmental information will increase. Regional groups will start to acquire data without uniform standards. Sensor technology proliferation will create a data explosion without a knowledge framework. Data aggregation from multiple sources, including NRM and statutory planning geographic information (GIS) systems, will expand. This would lead to adverse outcomes.

If resolved the clear benefit is an improved understanding of the problems, improved prioritisation and ultimately improved decision making. This would allow Australia to measure, for example, the anticipated improvements in modified and natural systems from existing and foreshadowed investments. It would allow greater leverage of private sector inputs, including capital, information and knowledge, to better integrate production systems with natural resource management.

Specifically this involves the:

- linking and quality assurance of a knowledge base.
- development of data interpretation systems so data becomes useful knowledge.
- linking of data to multiple applications including environmental valuation and environmental service markets.
- development of predictive modeling, with appropriate validation, in key areas of environmental concern; and
- linking of environmental, human and social data leading to better understanding of exchanges between.

The need to live in and pass on a sustainable environment is widely recognised. Australia invests almost \$25 billion per annum to achieve this goal. This initiative presents the opportunity to move from datasets that are largely unconnected, are unsustainable or lack durability, and cover only one third of the needs to a much stronger position.

Appendix A: Case Studies

Case Study 1: Avian Influenza Gene Sequence Sharing

The Background

Since 2003 there have been widespread outbreaks of a particularly virulent form of avian influenza known as H5N1, or the “bird flu.” This has resulted in the death or culling of hundreds of millions of poultry and has also caused more than 150 human deaths. The extent of these outbreaks and the potential for this virus to mutate into a potentially pandemic strain has resulted in an unprecedented worldwide effort to control the virus and to prepare in case significant levels of infections occur in the human population causing large outbreaks. These preparations include developing a variety of vaccines against H5N1 and stockpiling influenza antiviral drugs such as Tamiflu and Relenza.

In order to track the changes in H5N1, scientists have relied heavily on molecular techniques such as sequencing of the genes of many strains of the virus which have been isolated in different locations and covering the entire time span of the outbreaks. This type of analysis has been done for many years with both human seasonal influenza viruses and for other avian influenza viruses. These sequences are usually placed on public databases such as GenBank or the virus-specific Influenza Sequence Database (based at the Los Alamos National Laboratory). A recent example of the power of these techniques has been the sequencing and subsequent “resurrection” of the 1918 Spanish Influenza virus by Drs Taubenberger, Tumpey and colleagues.

The Problem

While scientists have freely lodged seasonal influenza virus sequences onto publicly available databases, until recently there have been several issues surrounding the lodging of H5N1 sequences. These included the desire for scientists to publish their results before divulging the sequences, restrictive agreements between countries and scientists regarding publishing of results that include sequences from H5N1 viruses and failure of some countries to supply viruses to laboratories capable of performing sequencing because of sensitivities and economic complications which follow disclosure of H5N1 outbreaks. This has been highlighted by the World Health Organisation (WHO) and others such as the Italian scientist, Ilaria Capua (see attached story from Nicolas Zamiska from *The Wall Street Journal*, 13 March 2006).

The Solution

A new approach described in a letter to *Nature*⁵⁹ signed by some 70 scientists and health officials from various countries, proposing the Global Initiative on Sharing Avian Influenza Data (GSAID) is designed to address this issue. Those participating in the consortium have agreed to share sequence data in public databases in Japan, Europe and the United States, analyse results jointly and publish collaboratively. This has removed many obstacles to the sharing of information and should help in the global push to understand and prepare for any eventuality associated with these deadly viruses. A similar approach was taken with the mapping of the DNA sequence variation for the human genome. This consensus has already had a profound effect and has led to the release of hundreds of H5N1 sequences from the Center for Disease Control (CDC) (USA) and other laboratories around the world and led to the release of new H5N1 isolates to laboratories for sequencing.

⁵⁹ Bogner P et al *Nature* 2006:442; 981

Case Study 2: Australian Ocean Data Centre Joint Facility

The Australian Ocean Data Centre Joint Facility (AODCJF) is a new, collaborative initiative conceived in 2004 from recommendations made in a review coordinated by the Royal Australian Navy, which examined the future structure and function of Australia's existing National Oceanographic Data Centre. At the time of the review this Centre was located within the Department of Defence.

The purpose of the AODCJF is to promote the discovery, access, long-term archival and exploitation of knowledge about marine systems. It is one of 65 designated National Centres established under the UNESCO Intergovernmental Oceanographic Commission's (IOC) International Oceanographic Data and Information Exchange (IODE) network. This global system of data centres was established to promote the free and open exchange of marine scientific data. The main differences between the AODCJF and the previous National Data Centre are that the AODCJF will: (a) operate as a virtual, distributed data centre, leveraging new technologies available to support distributed computing and (b) be owned and operated by a number of domain experts, who collectively have significantly more capability and capacity than was available to the previous RAN-based centre.

In the first instance, the AODCJF comprises the Australian government bodies involved in marine research and policy implementation: the Australian Antarctic, the Australian Institute of Marine Science, the Bureau of Meteorology, CSIRO Marine and Atmospheric Research, Geoscience Australia, the Department of Environment and Heritage (Marine Division) and the Royal Australian Navy. The AODCJF is now active as an entity and has been formally established by signing of a Collaborative Head Agreement by each of these parties. The Head Agreement used by the parties was obtained from the Australian Government Information Management Office (AGIMO) developed under its National Service Improvement Framework¹. The Framework provides the basis to enable different agencies to effectively collaborate across technical, business, legal, financial and governance areas.

The operation of the AODCJF is coordinated by a technical committee, comprising the Managers of the respective agency data centres. A governing board oversees the strategic direction and governance of the AODCJF. The board has an independent, paid Chair, a senior executive representative from each partner agency and two independent members (from the university and industry sectors).

The vision of the AODCJF is that the network will expand to ultimately encompass all of Australia's marine data, including marine data held or generated by universities and other institutions outside of Australian government agencies. This expanded vision is currently being pursued through BlueNet, a Systemic Information Infrastructure backed project.

The current AODCJF data curation and data access network utilises the existing hosting facilities of the AODCJF partners and is connected via an infrastructure based upon open, geo-spatial web service standards and a services catalogue.

Links

1. National Service Improvement Framework <http://www.agimo.gov.au/services/services>

Case Study 3: PARADISEC – Pacific and Regional Archive for Digital Sources in Endangered Cultures

Vulnerable languages of our region

30% of the world's languages (3,000) are in the East Asia–Pacific Islands region. Papua-New Guinea alone has 800 languages. Each of these societies has a rich tradition of botanical, zoological, ecological, ethnographic and historical knowledge, which often has important scientific and economic implications. The only record of this knowledge is oral, but with globalisation this knowledge is rapidly vanishing. By 2088, for instance it is predicted that Australia will have lost 90% of its Indigenous languages.

Australian leadership

As one of a handful of countries in the region with the scientific infrastructure to respond adequately to this crisis, Australia has shown leadership in our region, both in documenting cultural knowledge and in developing procedures for digital storage of the resulting sound, video, pictorial and text documentation. Since the 1950s Australian researchers have been active in the region making unique and unrepeatable audiovisual recordings of these endangered languages and cultures.

Vulnerability of the research record

With the obsolescence of all analogue recording media of the twentieth century, all audiovisual recordings on analogue media such as tape or cassette are under threat. Because of the specialist expertise needed to interpret linguistic material, the unique cultural content of collections in now endangered languages is doubly vulnerable. Digitisation of existing records on tape or cassette, and training new researchers in sustainable digital fieldwork methods has become urgent.

In 2003, Linkage Infrastructure Equipment and Facilities (LIEF) funding to a collaborative project between four universities established a pilot digital archive, PARADISEC, with the task of creating the architecture for a digital archive of materials from the region. In three years PARADISEC has created archivally stable copies with proper cataloguing information of 500 languages, totalling 1400 hours of sound recording, amounting to 2.87TB of digital material.⁶⁰ Now recognised as a world leader in sustainable digital archiving of cultural material, PARADISEC has collaborated with other institutions in Australia and internationally to establish best practice standards for archiving complex digital objects and together with APAC at the ANU, pioneered approaches to managing online access to large datasets of complex objects. PARADISEC's innovations in providing access, training, and archiving have evoked eager and positive responses from our neighbours in the region.

Vulnerability of PARADISEC

There is no mechanism for long-term funding for digital archives such as PARADISEC. And there is no other institution (library or archive) that has the resources to maintain large datasets of complex linked objects (sound, audio, video and text) and keep them accessible to researchers, let alone continue the work of preserving the botanical, zoological, ecological, ethnographic and historical knowledge of the region. Digital archives such as PARADISEC need funding for:

- maintenance and upward migration of digital data-sets;
- ingest of new digital files (and transferring analogue materials to digital formats);
- architectural development;
- cataloguing of digital data; and
- education and training of researchers so as to produce the best and most archivally stable documentation.

⁶⁰ This represents less than 1% of the documentation task. A minimum documentation standard for a language is 50 hours of audio and video recordings, and encyclopaedic definitions for 5000 words. To document the languages of the region we are therefore looking at a minimum of 150 000 hours of recording.

Case Study 4: Australian Digital Thesis Project

Introduction

The Australian Digital Thesis (ADT) project is a collaborative project among a majority of Australian and some New Zealand universities. This is the Australia-New Zealand version of worldwide movement to release research theses on the Internet for public access.

Aims

The aim of the ADT project is to establish a distributed database of digital versions of theses produced by the postgraduate research students at Australian universities. The theses will be available worldwide via the web. The ideal behind the project is to provide access to, and promote Australian research to the international community.

The initial project was funded by an Australian Research Council (ARC) Research Infrastructure Equipment and Facilities (RIEF) Scheme grant (1997–98). This was a collaborative project led by the University of New South Wales. The project received additional funding from DEST in 2004 through the Systemic Infrastructure Initiative as part of *Backing Australia's Ability*. DEST funding was provided in response to the recommendation of the Australian Research Information Infrastructure Committee (ARIIC).

The ADT concept was an initiative of seven Australian universities (listed below) in association with the Council of Australian University Librarians (CAUL).

The ADT model was developed by the seven original project partners during 1998–99. The program was then opened up to all CAUL members (all Australian universities) in July 2000. The original seven partners will continue to guide and advise the national group in their role as the ADT Steering Committee.

The original ADT membership group:

- University of New South Wales (lead institution)
- University of Melbourne
- University of Queensland
- University of Sydney
- Australian National University
- Curtin University of Technology
- Griffith University

ADT Project Overview

The ADT project is designed to improve access to, and enhance transfer of, the research information contained in theses by providing a full text version available from the desktop via the web. The retrieval is enhanced by the inclusion of metadata tags in the documents which are given a higher weighting by the more sophisticated search engines.

It is also designed to provide a new model for deposit and archiving of theses that takes into account the tools and technologies that students are now using to prepare their theses. The project has two major components, digitisation of theses as part of the deposit process and the digitisation of a selected number of frequently requested existing theses. As each University is responsible for maintaining an archival copy of the theses of their own institution, each participant in the project will mount their own theses on a server located in their respective institution. The participants will use the same database configuration, standards and metadata system to ensure compatibility. The document format will be Adobe Acrobat Portable Document Format (PDF) ensuring that the data is independent of the platform on which it is created. PDF ensures that a high quality printed version can be provided if needed. Acrobat is relatively easy to use, with a high quality, free reader readily available. PDF has also become an electronic publishing standard.

Member institutions currently consist of the following:

- Australian Catholic University
- Australian National University
- Central Queensland University
- Curtin University of Technology
- Deakin University
- Edith Cowan University
- Flinders University
- Griffith University
- La Trobe University
- Murdoch University
- Queensland University of Technology
- RMIT University
- Southern Cross University
- Swinburne University of Technology
- University of Adelaide
- University of Ballarat
- University of Canberra
- University of Canterbury (NZ)
- University of Melbourne
- University of New South Wales
- UNSW at ADFA
- University of Newcastle
- University of Otago (NZ)
- University of Queensland
- University of South Australia
- University of Southern Queensland
- University of Sydney
- University of Tasmania
- University of Technology Sydney
- University of Waikato (NZ)
- University of Western Australia
- University of Western Sydney
- University of Wollongong
- Victoria University of Technology

Conclusion

The ADT project would be a good example of a federated information management system in which each institution, because of their own regulations and requirements, would have its own policies relating to the release of information contained in the theses. This service is now well established, and is self maintaining in that each participating institution is responsible for its own information management.

Case Study 5: Species 2000 and ITIS Catalogue of Life

The Species 2000 and ITIS Catalogue of Life is an example of:

- a successful current approach to managing complex data in biology;
- the computational challenges in managing complex data from multiple sources; and
- the sociology of international collaboration between database projects in biology.

The world has about 1.75 million known and named living species of plants, animals, fungi and micro-organisms. People need to be able to access information about them for the management, conservation and sustainable use of biodiversity.

The international partnership of Species 2000 and the North American Integrated Taxonomic Information System (ITIS) is producing a unified, authoritative index of the world's species: the Catalogue of Life. This is a keystone knowledge set – the gateway to a digital library of biodiversity information on the Internet, using direct species links to other data systems on subjects as varied as specimen data, agriculture, pharmacognosy, and conservation uses. The Catalogue of Life is available on the Web (www.sp2000.org) and also as an annual checklist edition on a CD.

The Catalogue of Life uses scientific names to designate species because they are unambiguous and ideally have a 1:1 relationship with species, and therefore are best for locating and linking information about species. Common names are problematic because there are so many of them, in so many different languages, and rarely with a 1:1 correspondence to species. Frequently, there are many common names for one species, or conversely one common name is used for several species, leading to confusion as to which species is meant.

The genetic diversity inherent in living organisms means this is far from a simple exercise of listing names. This is a knowledge-gathering programme, involving taxonomic expertise in interpreting species and their relationships. The expertise needed to create and continuously enhance a global species database for a major group is the 'tip of an iceberg' below which lies layer upon layer of the taxonomic processes: from field observation and collections through to monographic revisions and phylogenetic analysis. Names are the mere tags by which this knowledge is accessed.

Compiling an index of species is further complicated by the fact that understanding of biodiversity is still far from adequate, resulting in many scientific names not yet being in a 1:1 relationship with species. For example, a widespread species that grows in several regions is likely to have been given a different scientific name in each of these regions over the past two centuries, as a result of taxonomic biologists working in isolation and not having ready access to specimens and publications from other regions. Taxonomic research by experts is needed to sort out such problems.

Sources of information (electronic and hardcopy) are scattered and, until recently, it was difficult for anyone to readily find out the names of whatever species were of interest to them, or to find further information about those species. One had to know where to find the scattered sources of information and then be knowledgeable enough to interpret what was found: for example, did species name A in country X represent the same species as species name B in country Y? And what about species name C from another region of country Y: was it the same as A and/or B, or a different species?

The complex nature of biological knowledge, as in other complex disciplines such as meteorology and oceanography, requires careful consideration of the structure of databases and data repositories trying to manage such data.

The last decade has seen a revolution in this area of research, as in most others, with the explosive development of e-Science. It is now much easier for biologists in different regions to collaborate on research projects, thanks to innovations such as email and video-conferencing, and aided by availability of analytical software and electronic images of specimens and publications, often on websites. Also, an increased ability for biologists to travel has allowed them to extend their research through fieldwork in relevant parts of the world.

So there is a renewed effort to sort out confusion about the delimitation of species and their scientific names. Many biological database projects have started around the world, making available electronically information about a particular group of plants, animals or micro-organisms. We call these Global Species Databases (GSDs), and they are key elements in the Catalogue of Life because they provide a comprehensive taxonomic snapshot of all the species in a particular group. Regional databases that include all the organisms in a particular region of the world are also important in adding details not covered in the GSDs.

A range of database projects, spread around the world, are collaborating to produce the Species 2000 and ITIS Catalogue of Life. This has involved developing software and data standards to link and merge multiple complex data-sets. However, the key component that marks this as being much more than a mere list of names is the expert input from taxonomic biologists in all parts of the world to validate the complex biological content. This has dictated a distributed model for the Catalogue of Life. Even though a centralised model is more efficient computationally, it is sociologically very important to keep the individual data-sets close to the taxonomists who provide the expertise to update the species information. Another advantage of the distributed approach is that the work of aggregating taxonomic knowledge is going ahead in a massively parallel way, rather than in a serial fashion as would happen with a centralised approach.

Besides interactions with individual taxonomic experts, this project interacts strongly with a wide range of international and national bodies, including the International Union for Biological Sciences (IUBS), the International Union of Microbiological Societies (IUMS), the Committee on Data for Science and Technology (CODATA) of the International Council for Science, the International Working Group on Taxonomic Databases (TDWG), the European Union, the Global Biodiversity Information Facility (GBIF), the Global Taxonomy Initiative of the Convention on Biological Diversity, and the National Institute of Environmental Studies in Japan. In Australia, the Royal Botanic Gardens Sydney has played a role from the early days of the project, and more recently the Australian Biological Resources Study and biologists from the Australian Museum and the Queensland Museum have become involved.

The success of this distributed approach is seen in the fact that, since 2001, more than 880 000 species have been added to the Catalogue of Life: about 50% of the world's known species (Figure 1). The aim is to add the other 50% by 2011, but these species mostly belong to poorly studied groups, especially among the insects, and so it will be a major challenge to reach 100% in that time frame.

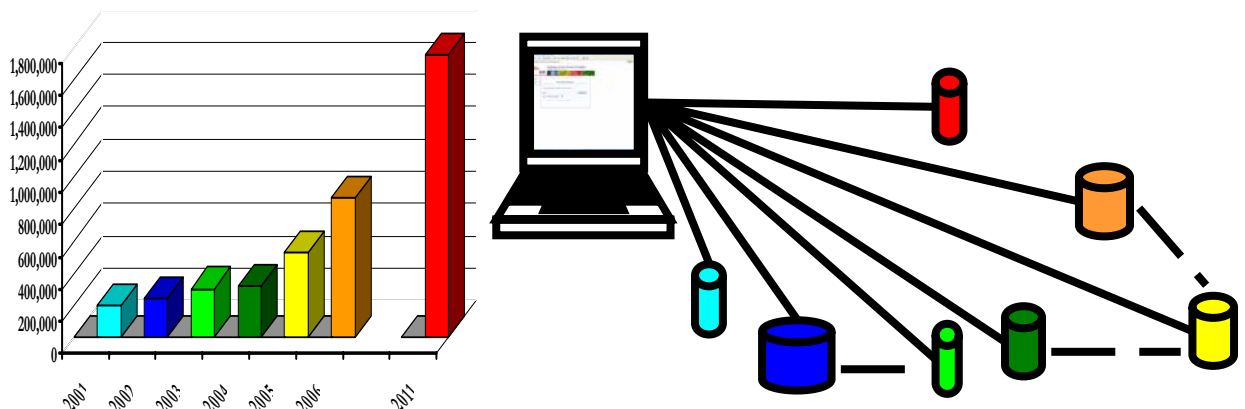
The Catalogue of Life is an index to the world's species, and its obvious application is to provide links to a wide range of information about the world's species: where they live naturally, where they are invasive, what they look like, what their uses for human food or medicine might be, etc. Even though the Catalogue is not yet complete, it is already proving useful as an index. For example, the Global Biodiversity Information Facility (www.gbif.net) uses it as the taxonomic backbone for its web portal.

The continuing challenge is to promote the study and naming of the living species that are not yet known and named: estimated at anywhere from 3 to 50 million species, most of them micro-organisms or small invertebrate animals such as insects.

Species 2000 is implementing an architecture that is capable of both creating a complete Catalogue of Life and of maintaining its taxonomic enhancement through time. At a superficial level, this programme is about creating databases and continuing to maintain them, but underlying this is a serious proposal for self-organisation within the taxonomic community and for rationalising and structuring taxonomic effort on a global and continental scale.

Figure 6. Progress with compiling the *Catalogue of Life*.

Currently, half of all known species (880,000) are included in the Catalogue of Life Annual Checklist, drawn from nearly 40 contributing species database projects. The aim is to include all species by June 2011.



Case Study 6: Biometric Technologies

Biometrics is the term applied to technologies used for recognising an individual based on their physical and behavioural characteristics. These may include face recognition, fingerprint recognition, iris recognition, palm print recognition, speech recognition, thermal imaging of vascular and skin patterns, gait recognition, hand geometry and ear shape.

In essence biometric technologies operate in three domains:

- Verification – Is the person who they claim to be?
- Watchlist – Is the person on a database? If so who are they?
- Identification – This person is on the database. How soon can he or she be found?

There is a strong argument for a national approach to biometric applications:

- It would make more efficient use of R&D and acquisition capabilities, especially where there is a need for Australian specific requirements.
- It would ensure that all Government agencies have access to current technologies and techniques.
- It would facilitate commonality and thus (as appropriate) sharing of datasets, tools and training.
- It would also enable sharing of personnel to meet operational emergencies.
- It would ensure common operational standards – including in such matters as privacy and evidentiary practices.

Considerable effort has been made to improve the accuracy, speed and reliability of biometric technologies however a number of significant challenges remain.

Their successful application, for example, requires systems to be trained on representative datasets that are often difficult to obtain. Current coverage of open source representative datasets is sparse and of variable quality.

Most biometric technologies involve the development of a model that will most likely be stored as a sequence of numbers and defines the individual characteristics to be used for recognition. For example, a face recognition system will transform the characteristics of a person's face in to a sequence of numbers, a feature vector that will be stored in the database to be compared against similarly transformed incoming test faces. Different face recognition systems will use a different set of features and hence use a different format for storing models. In order to effectively share these databases, it is important that agreed standards be developed that define the formats to be used for defining the structure of face models.

Another issue is the development of consistent metadata standards that support advanced query mechanisms and facilitate access and sharing of large volumes of multimedia information. Compression formats also need to be agreed upon and the format recorded as metadata for each file, and a variety of data issues need to be considered to ensure that information can be accessed and shared efficiently, for example data fields could be broken up according to a pre-defined structure.

As the demand for more sophisticated biometric technologies grows there is an increased requirement for implementation of nationwide protocols, standards and systems that provide effective creation of and access to structured datasets, as well as data storage facilities and computational capabilities to facilitate access and analysis. This will enable improved access by the research community to develop and test new biometric technologies and by defence and security agencies to implement these technologies into their operations.

Case Study 7: ICTVdB, Universal Virus Database built for the International Committee on Taxonomy of Viruses

Origins and objectives

In 1990 the American Type Culture Collection (ATCC; the US repository of microbial type species) organised a think tank to examine the need for an all inclusive virus database. Experts in virology and bioinformatics worldwide, and representatives of ICTV, EMBL/EBI, NCBI, NIH, and USDA participated in this think tank. The traditional, “bottom-up” meeting recommended construction of a Universal Virus Database (ICTVdB) under the umbrella of ICTV, based on the successful model of VIDE developed by AJ Gibbs at ANU, using DELTA (Description Language for Taxonomy) developed by MJ Dallwitz, CSIRO Entomology. that was then the accepted world standard.

Dr Cornelia Büchen-Osmond, then working on the plant virus database, VIDE, in the Research School of Biological Sciences (RSBS), ANU, was appointed to develop ICTVdB in Canberra supported by ANU and supplemented by NSF grants to ATCC. Until 2006, ICTVdB remained a single investigator project, mentored by visionaries in virologists and bioinformatics. It recognised that data is recorded, evaluated and published by individual researchers or team efforts after established processes of peer review, and that individuals and teams differ greatly in their creativity, styles, standards and research methods in diverse disciplines. The early objectives of ICTVdB were:

- to standardize the characters needed to describe all properties of all types of viruses. This effort was guided by the “ICTV code for the description of virus characters” (1983), an earlier ICTV computer database initiative which could not be realised with existing computer systems. All other available virus databases were scrutinised to capture the growing number of characteristics for virus identification.
- to accommodate the semantic complexity that had evolved in naming of viruses, and to handle renaming of entities in this young, rapidly developing branch of science. A unique decimal code identifier (now 19 digits), was developed that indicates the taxonomic level and relationships of all virus particles. This robust core of ICTVdB was built on diverse concepts such as a computer IP number, the enzyme code, and the Dewey system. It serves as a file name for descriptions generated in the database, and as a tag for linkage from other databases such as Genbank (genomics) and SWISS PROT (proteomics) and search engines.
- to use the Internet for easy access to, and distribution of the latest peer reviewed decisions on virus nomenclature. The Web presentation of ICTVdB was set out like a library catalogue, and prior to the emergence of sophisticated search engines (such as Google), was widely used by scientific and lay communities. Distinctively, images, references, and other frequently used resources were stored outside the database core but attached to the catalogue of species names and virus descriptions generated from the database.

Lessons and achievements

Having met the above core objectives, having anticipated the needs for flexible data structures and intrinsic interoperability, many other challenges emerged that had to be addressed, including links:

- to archival data from aging experts (especially EM images) now less “fashionable” but still indispensable;
- to huge new datasets (such as genomic proteomic sequences) as they emerged in the public domain;
- that facilitated real-time data exchange (work in progress) for ICTV peer review before entry in ICTVdB.

Some lessons are that interoperability needs to be intrinsic, not extrinsic, that protocols need to be designed and agreed by those who generate and use the data, and that resources need to be provided early, and at disciplinary and interdisciplinary levels (that is, distribute correct species names to other databases and taxonomy providers), to achieve these data management goals.

Although a small but long-term effort ICTVdB, has now attracted secure funding from the United States National Institute of Health (the NIH⁶¹) that potentially will see it transformed to accept, digest and present the huge amount of sequence, epidemiological and other data generated by virologists, using the most suitable and interactive new IT platforms. The initiative, initially the domain of those concerned to provide easy access to expert agreed, unambiguous nomenclature of viruses, subsequently found important applications “in the national interest”, including monitoring of biothreats⁶¹ and emerging diseases.⁶²

⁶¹ Ecker D J *et al.*, 2005, *BMC Microbiology*, **5**, 19: <http://www.biomedcentral.com/1471-2180/5/19>

⁶² Büchen-Osmond C., 2007. 'Taxonomy and Classification of Viruses'. In: *Manual of Clinical Microbiology*, 9th edition, Vol 2, Ch 76. ASM Press, Washington DC

Appendix B: Consultations

In developing this report, a range of individuals, organisations and agencies, both national and international, were contacted to seek their views and opinions on the issues being addressed by the Working Group.

The Working Group would like to thank the following, and to acknowledge the valuable contribution they made:

Dr Penny Allbon, Director, Australian Institute of Health and Welfare

Professor Warwick Anderson AM, CEO, National Health and Medical Research Council

Australian CODATA science team: Prof Ray Norris, CSIRO Australia Telescope National Facility, Dr Charles Barton, Research School of Earth Sciences, ANU, Dr Karen Wilson, Royal Botanic Gardens, and Dr Alex Held, Head, CSIRO Office of Space Science and Applications

Australian Government Department of the Environment and Heritage

Australian Government Information Management Office

Dr Linda Barwick, Director, Pacific and Regional Archive for Digital Sources in Endangered Cultures

Associate Professor Bob Beeton, Chair, 2006 State of the Environment Committee

Dr Cornelia Büchen-Osmond, ICTVdB Management, Columbia University, USA

Dr Markus Buchhorn, Australian Partnership for Advanced Computing

Professor Tom Cochrane, Deputy Vice-Chancellor (Technology, Information and Learning Support), Queensland University of Technology

Professor Len Cook

Ms Karen Curtis, Privacy Commissioner

Dr Michael J Coughlan, Head, Bureau of Meteorology National Climate Centre

Peter Crossman, Head, the Office of Economic and Statistical Research, Queensland

Dr Matthew Cuthbertson, CEO, Cooperative Research Centre for Advanced Automotive Technology

Defence, Science and Technology Organisation

Dr John Dodgson, CEO, Australian Academy of Technological Sciences and Engineering

Dr Sue Forrest, Director, Australian Genome Research Facility

Dr Rhys Francis, National Collaborative Research Infrastructure Strategy Facilitator, Platforms for Collaboration; Executive Manager, CSIRO e-science

Professor Malcolm Gillies, President, Council for the Humanities, Arts and Social Sciences (CHASS)

Professor Chris Goodnow, Director, Australian Phenomics Facility, John Curtin School of Medical Research

Professor Ian Gust, Director and Dr Ian Barr, Deputy Director, World Health Organisation's Collaborating Centre for Influenza

Professor Peter Høj, CEO, Australian Research Council

Ms Sophie Holloway, Manager, Australian Social Science Data Archive

Mr Laurence Lock Lee, Principal Consultant, Strategy, Research Associate, CSC Leading Edge Forum

Professor Tony McMichael, Director, Associate Professor Chris Kelman, and Dr Gillian Hall, National Centre for Epidemiology and Population Health

Mr Paul McNamara and Dr Adrian Burton, Australian Partnership for Sustainable Repositories

Dr Anthony Maedar, Research Director, and Mr Gary Morgan, CEO, e-Health Research Centre, Brisbane

Dr Phil McFadden, Chief Scientist, Geoscience Australia, and National Collaborative Research Infrastructure Strategy Committee Member

Mr Scott McTaggart, CEO, the Australian Computational Earth Systems Simulator (ACcESS) MNRF; Dr Joe Kurtz, ACcESS, Research School of Earth Sciences, ANU; and Prof Hans-Bernd Mühlhaus, ACcESS, Earth Systems Science Computational Centre, University of Queensland

Dr Reagan Moore, San Diego Supercomputer Centre, CA USA; Chief Architect, Storage Resource Broker, USA

Professor John O'Callaghan, Executive Director, Australian Partnership for Advanced Computing

Professor Bernard Pailthorpe, Chair, Computational Science, University of Queensland; CEO, Queensland Cyber Infrastructure Foundation Ltd

Dr Richard Price, Head, Intelligence Analysis Group, Defence Science and Technology Organisation

Professor Sue Richardson, President, Academy of the Social Sciences in Australia

Dr Mike Sargent AC, Chair, e-Research Coordinating Committee; Chair, National Collaborative Research Infrastructure Strategy

Professor Sue Serjeantson AO, Executive Secretary, Australian Academy of Science

Professor Andrew D Short, School of Geosciences, University of Sydney

Dr John Sims, Programme Leader and Dr Cliff Samson, Executive Director, Bureau of Rural Sciences

Mr Dennis Trewin, Australian Statistician, Australian Bureau of Statistics

Mr Paul Trezise, Chief Information Officer, Geoscience Australia

Mr Mike Tsykin, Systems Engineering Research Centre, Fujitsu Australia and New Zealand

Prof Graeme Turner, President, Australian Academy of the Humanities

Mr Warwick Watkins, Chair, Australia and New Zealand Land Information Council – the Spatial Information Council

Mr Peter Woodgate, CEO, Cooperative Research Centre for Spatial Information

Appendix C: Australian and International Initiatives

The challenges and opportunities discussed in Chapter 1 have been acknowledged by many different agencies, including government agencies and research institutions, as being fundamental to their individual missions of delivering productivity and effectiveness in the 21st century. There is also a clear indication that the overarching requirement for scientific data and its management has been identified at a national level in many countries. In many cases national initiatives have been drawn up and are underway that seek to provide solutions to this scientific data management issue and all of its subordinate requirements.

International forums, such as the OECD and the ICSU, have highlighted the benefits, outcomes and productivity impacts that are likely to occur if nations can put in place national policies and infrastructure for scientific data management. And they are actively encouraging cultural and behavioural change at all levels in order to stimulate progress towards a vision of research and scientific endeavour suited to the 21st Century and beyond.

The Australian and international initiatives outlined below are provided as an indication of the various activities taking place around the globe.

Australian Initiatives

Systemic Infrastructure Initiative (SII)

Funded under the Australian Government's "Backing Australia's Ability – An Innovation Action Plan for the Future" and the follow-on "Backing Australia's Ability – Building Our Future Through Science and Innovation" packages, the SII addressed key issues in Australia's national research information infrastructure. The breadth of SII funded activities has ranged from the establishment and enhancement of physical research infrastructure, including a major upgrade of Australia's broadband research and education network, maintaining and developing Australia's advanced computing capabilities, through to the research and development of software systems for digital repositories.

A significant proportion of SII funding has addressed key issues in the management and integration of large data sets, technical interoperability and accessibility, deployment including copyright and digital rights management, and the federation and sustainability of digital repositories.

In particular, SII projects have delivered a number of important information infrastructure outcomes of direct relevance to scientific data management, including:

- Repository management systems
- Expertise and advice on the selection, implementation, operation and maintenance of institutional repositories
- Robust technology solutions for an open standards approach to authorisation that can be adopted by a wide range of digital repository systems, enabling AAA⁶³ within the research and higher education sector
- Alignment of different authentication technologies to support varied requirements for the strength of security over data or other resources
- Tools to support deposit into, access to, and annotation by a range of actors, to digital libraries including publications, datasets, simulations and software

⁶³ Authentication, Authorization and Accounting

- Assistance for researchers in dealing with intellectual property issues during and after the research process
- Demonstrators of different approaches to capturing and managing large and disparate data-sets, including distributed data and issues of privacy in medical research

e-Research Coordinating Committee

The e-Research Coordinating Committee was established in recognition of the potential capability and benefits of modern information and communication technologies (ICT) in the conduct of research. Committee members were appointed on 21 April 2005 by the Minister for Communications, Information Technology and the Arts and the Minister for Education, Science and Training.

The e-Research Coordinating Committee's vision was to enable Australian researchers to achieve world class research endeavours and outcomes and to disseminate knowledge gained from research, through the use of advanced ICT. The Committee's objectives were 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 approach.

The Committee's interim report, published in 2005, set out the policy issues pertinent to Australia securing maximum benefit from the use of e-Research techniques; proposes strategic directions which should be pursued; and proposes further steps that would allow generation of an implementation plan.

The final report has been sent to the Minister for Communications, Information Technology, Senator the Hon. Helen Coonan, and the Minister for Education, Science and Training, the Hon. Julie Bishop MP for their consideration.

National Collaborative Research Infrastructure Strategy (NCRIS)

The Australian Government's "Backing Australia's Ability – Building Our Future through Science and Innovation" package has also provided funding for NCRIS. The NCRIS Strategic Roadmap identified the discovery, access, storage, management, and curation of data as a key platform required to sustain the standing of Australia's researchers and support the development of collaborative approaches to research.

Many of the capabilities identified in the Roadmap will produce or depend upon large sets of 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. The investment in generic information infrastructure will be conducted through the "Platforms for Collaboration" capability.

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.

The collaborative nature of NCRIS will necessitate the adaptation and evolution of current information infrastructure resources. Repositories have the potential to move beyond the

traditional approaches, namely for storing publications, to support innovative new forms of research data, collections and research output.

In order to manage these new forms research outputs and their application, 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.

To be fully exploited by search engines and data mining tools, much of the data to be made accessible through the linkage of databases (which will include experimental data) will need to be annotated with relevant metadata, providing information on provenance, content, conditions of use, and so on.

To enhance researcher effectiveness and facilitate easier access to research results and outcomes, it is also essential that electronic storage of research is consistent with internationally agreed technical standards.

National Data Network (NDN)

The NDN, currently in a Demonstration Phase, is a network of information resources (data, analysis/presentation tools or other related services) being developed by the ABS on behalf of a consortium of public sector agencies.

The NDN's vision is for the best utilisation of, and return on investment from, survey and administrative information resources by:

- providing Users (eg planners, researchers, policy analysts, project managers, evaluators) with an improved means of finding and accessing information resources, particularly those information resources that aren't publicly available, and
- providing Custodians of information resources with a means of giving greater visibility and wider access to their information resources, while continuing to give close attention to access and use conditions.

The NDN is an Internet-based distributed computer network comprised of:

- a central catalogue comprised of entries made by the Custodians of information resources,
- nodes through which Custodians make some information resources electronically accessible by Users,
- a facility for Users to search the catalogue and find/discover relevant information resources,
- facilities that provide Users with electronic access to information resources, subject to access rights having been granted by the Custodians concerned.

The NDN is being developed using open source software, available free of charge, thereby:

- avoiding software costs that might otherwise be a barrier to participation by Custodians and Users, and

- creating considerable potential for collaboration on the future development of the network.

NHMRC, ARC, and AVCC Australian Code for the Responsible Conduct of Research

The NHMRC, the Australian Vice Chancellors' Committee, and the ARC have established a Joint Working Group to review the NHMRC/AVCC Statement and Guidelines on Research Practice.

The purpose of the Australian Code for the Responsible Conduct of Research is to guide institutions and researchers in how to achieve and maintain responsible research practice. The code addresses the role of institutions in establishing research policies that promote high standards of research integrity and an environment in which research will be conducted responsibly.

While written particularly for public sector institutions that undertake and support research in Australia, this code will also have relevance to private sector institutions. These institutions are urged to adopt this code as far as possible within their operating environments.

Of particular relevance to the work of this Working Group are the Code's sections on the Management of research data and primary materials and on Publication and dissemination of research findings.

It is expected that the final version of the Code will be available in late 2006/ early 2007.

International Initiatives

OECD Committee for Scientific and Technological Policy

The OECD Committee for Scientific and Technological Policy (OECD/CSTP) aims at informing the policy debates on the contribution of science and technology to sustainable growth and societal needs in knowledge-based economies and at promoting international co-operation in scientific research. The OECD/CSTP have focussed on issues that are related to the accessibility of publicly funded research and more specifically those related to ICT, such as Access to Network Infrastructures, Virtual Libraries, Electronic Publishing, Virtual Laboratories, Intellectual Property and Data Resources.

The OECD/CSTP met at Ministerial level on in January 2004. The meeting was chaired by Australia's Minister for Science of Australia, the Hon Peter McGauran MP. In relation to access to data, the Committee's final communiqué noted that:

Access to research data

17. Ministers recognised that fostering broader, open access to and wide use of research data will enhance the quality and productivity of science systems worldwide. They therefore adopted a Declaration on Access to Research Data from Public Funding, asking the OECD to take further steps towards proposing Principles and Guidelines on Access to Research Data from Public Funding, taking into account possible restrictions related to security, property rights and privacy (Annex 1).

Annex 1

OECD DECLARATION ON ACCESS TO RESEARCH DATA FROM PUBLIC FUNDING

adopted on 30 January 2004 in Paris

The governments (1) of Australia, Austria, Belgium, Canada, China, the Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Luxembourg, Mexico, the Netherlands, New Zealand, Norway, Poland, Portugal, the Russian Federation, the Slovak Republic, the Republic of South Africa, Spain, Sweden, Switzerland, Turkey, the United Kingdom, and the United States

Recognising that an optimum international exchange of data, information and knowledge contributes decisively to the advancement of scientific research and innovation;

Recognising that open access to, and unrestricted use of, data promotes scientific progress and facilitates the training of researchers;

Recognising that open access will maximise the value derived from public investments in data collection efforts;

Recognising that the substantial increase in computing capacity enables vast quantities of digital research data from public funding to be put to use for multiple research purposes by many research institutes of the global science system, thereby substantially increasing the scope and scale of research;

Recognising the substantial benefits that science, the economy and society at large could gain from the opportunities that expanded use of digital data resources have to offer, and recognising the risk that undue restrictions on access to and use of research data from public funding could diminish the quality and efficiency of scientific research and innovation;

Recognising that optimum availability of research data from public funding for developing countries will enhance their participation in the global science system, thereby contributing to their social and economic development;

Recognising that the disclosure of research data from public funding may be constrained by domestic law on national security, the protection of privacy of citizens and the protection of intellectual property rights and trade secrets that may require additional safeguards;

Recognising that on some of the aspects of the accessibility of research data from public funding, additional measures have been taken or will be introduced in OECD countries and that disparities in national regulations could hamper the optimum use of publicly funded data on the national and international scales;

Considering the beneficial impact of the establishment of OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data (1980, 1985 and 1998) and the OECD Guidelines for the Security of Information Systems and Networks (1992, 1997 and 2002) on international policies for access to digital data;

DECLARE THEIR COMMITMENT TO:

Work towards the establishment of access regimes for digital research data from public funding in accordance with the following objectives and principles:

Openness: balancing the interests of open access to data to increase the quality and efficiency of research and innovation with the need for restriction of access in some instances to protect social, scientific and economic interests.

Transparency: making information on data-producing organisations, documentation on the data they produce and specifications of conditions attached to the use of these data, available and accessible internationally.

Legal conformity: paying due attention, in the design of access regimes for digital research data, to national legal requirements concerning national security, privacy and trade secrets.

Formal responsibility: promoting explicit, formal institutional rules on the responsibilities of the various parties involved in data-related activities pertaining to authorship, producer credits, ownership, usage restrictions, financial arrangements, ethical rules, licensing terms, and liability.

Professionalism: building institutional rules for the management of digital research data based on the relevant professional standards and values embodied in the codes of conduct of the scientific communities involved.

Protection of intellectual property: describing ways to obtain open access under the different legal regimes of copyright or other intellectual property law applicable to databases as well as trade secrets.

Interoperability: paying due attention to the relevant international standard requirements for use in multiple ways, in co-operation with other international organisations.

Quality and security: describing good practices for methods, techniques and instruments employed in the collection, dissemination and accessible archiving of data to enable quality control by peer review and other means of safeguarding authenticity, originality, integrity, security and establishing liability.

Efficiency: promoting further cost effectiveness within the global science system by describing good practices in data management and specialised support services.

Accountability: evaluating the performance of data access regimes to maximise the support for open access among the scientific community and society at large.

Seek transparency in regulations and policies related to information, computer and communications services affecting international flows of data for research, and reducing unnecessary barriers to the international exchange of these data;

Take the necessary steps to strengthen existing instruments and - where appropriate - create within the framework of international and national law, new mechanisms and practices supporting international collaboration in access to digital research data;

Support OECD initiatives to promote the development and harmonisation of approaches by governments adhering to this Declaration aimed at maximising the accessibility of digital research data;

Consider the possible implications for other countries, including developing countries and economies in transition, when dealing with issues of access to digital research data.

INVITE THE OECD:

To develop a set of OECD guidelines based on commonly agreed principles to facilitate optimal cost-effective access to digital research data from public funding, to be endorsed by the OECD Council at a later stage.

(1) Including the European Community

International Committee for Science

The International Committee for Science (ICSU) is a non-governmental organisation representing a global membership that includes both national scientific bodies (103 members) and international scientific unions (27 members). It is an international forum for scientific research and policy development. ICSU plays an important role in global data management initiatives – it is the founder of the current World Data Centre system, comprising facilities in over 40 countries that are engaged in collecting, managing and distributing geophysical, astrophysics and environmental data

Recognising that it has a lead role to play in guiding scientific data management policy in an international context, in December 2004 ICSU undertook a review of the needs and priorities for scientific data and information in terms of its production, management, access and dissemination. An analysis of existing ICSU activities and structures was also progressed.

Many recommendations were forthcoming from the review, covering issues such as:

- systems for data dissemination;
- interoperability;
- equitable access to data and information;
- intellectual property rights;
- metadata;
- data and information rescue;
- scientific publications;
- professional data and information management; and
- archiving.

ICSU has determined that it needs a long-term strategic framework for scientific data and information (policies, practices and infrastructure). The framework should build on existing data and information structures and services where it is advantageous to do so, but the ICSU is prepared to re-think, re-orient, and replace existing structures and bodies where it is necessary.

ICSU also intends to establish an international Scientific Data and Information Forum (SciDIF) involving all key stakeholders: ICSU members, interdisciplinary bodies, science funding bodies and other data providers and users. Through SciDIF, ICSU will aim to ensure that the full benefits of new data and information technologies and capabilities are extended to scientists throughout the world. It will also establish a Strategic Data and Information Committee to oversee the development of the long-term integrated framework for data and information and a Scientific Data and Information Forum (SciDIF). Membership of this committee should include representatives of relevant ICSU bodies and unions, and experts in information technology and professional data management.

UK e-Science Initiative

The UK e-Science initiative conceived in 2001 has changed the way the UK is approaching the management and exploitation of its scientific data. Prior to the UK e-Science initiative, scientific data management within the UK mirrored the situation in Australia. Some institutions were exemplars of best practice, but there was no strategic, whole of government approach to managing the nation's science data assets. While the e-Science initiative covers more issues than just sustainable data asset management, data management pervades most of the e-Science activities.

This initiative is now in its second phase and is addressing some of the perceived weaknesses in its first phase. The main goal of the UK e-Science project is to enable better research in all science disciplines. This is expected to be achieved through developing collaboration supported by advanced distributed computation

With this goal in mind, the second phase includes six major activities:

- A national e-Science Centre linked to a network of Regional Grid Centres

A National e-Science Centre has been established in Edinburgh, managed jointly by Glasgow and Edinburgh Universities. Eight other regional centres have been established (see <http://www.nesc.ac.uk/centres/>) and are helping to establish a UK e-Science Grid (see <http://www.grid-support.ac.uk/etf/>). Subsequently, these 9 Centres have been augmented by seven Centres of Excellence.

- Support activities for the UK e-Science Community

The UK's National Grid Service (NGS - <http://www.grid-support.ac.uk/>) provides a core e-Infrastructure that underpins UK e-Research, providing standardised access to data management and computer resources and supporting collaborative computing across the UK. The NGS is being built through re-focusing existing work of the Grid Operations Support Centre (GOSC).

The NGS also provides a national "gateway" to international initiatives through collaboration with related e-infrastructures internationally. The National Grid Service, funded by the Joint Information Systems Committee (JISC) and Council for the Central Laboratory of the Research Councils (CCLRC), was created in October 2003. The service entered full production in September 2004 at which point the GOSC was created with support from the Engineering and Physical Sciences Research Councils (EPSRC) core e-Science programme.

- An Open Middleware Infrastructure Institute (OMII),
OMII-UK (<http://www.omii.ac.uk/>) provides a web service infrastructure for building grid applications. They are focused on providing an open source system that addresses the user requirements of combining ease of use with a secure environment.
- A Digital Curation Centre (DCC)
Working with other practitioners, the Digital Curation Centre (<http://www.dcc.ac.uk/about/>) will support UK institutions who store, manage and preserve data to help ensure their enhancement and their continuing long-term use. The purpose of the centre is to provide a national focus for research and development into curation issues and to promote expertise and good practice, both national and international, for the management of all research outputs in digital format.
- New Exemplars for e-Science,
- Participation in International Grid Projects and Activities.

US National Science Board

In May 2005 the US National Science Board (NSB) issued a report entitled *Long-lived digital data collections: enabling research and education in the 21st century*.

The NSB had recognised the growing importance of digital data collections for research and education, and the substantial funds that are being invested in creating them. The Board was concerned to establish a considered policy framework for the management and use of these collections. It was concerned in particular with “long-lived” collections, which it defined as those which will be needed for a period of time long enough for there to be a concern about the impacts of changing technology.

In its report, the NSB called on the US National Science Foundation to establish a clear technical and financial strategy which would ensure the maximum possible return on the investments made by the Foundation in developing data collections. Among other things, it called on the Foundation to analyse the appropriate balance between the investment in data collections and the investment in the research that exploits the collections. It identified the issue of “proliferating collections” and noted that, as research becomes more interdisciplinary, policies and standards need to be harmonised, which is an increasing challenge as the number of digital data collections grows.

The report recognised that many organisations that manage digital data collections take on what it called “community-proxy functions”: they make choices on behalf of current and future user communities concerning the way that these data collections are made accessible and sustained for the long term. The report raised a concern that these decisions may be inconsistent and uncoordinated.

The report recommended that the National Science Foundation require that research proposals for activities that will generate digital data of long term significance include a data management plan that can be evaluated by peer reviewers.

The report identified the importance of “data scientists”, among which it included computer scientists, disciplinary experts, librarians and archivists who are crucial to the successful management of digital data collections. It recommended that the National Science Foundation be proactive in advancing programs that educate and reward data scientists, and that it work with other stakeholders to develop and mature the career path for data scientists.

National Aeronautics and Space Administration Space Science Data Services

The National Aeronautics and Space Administration (NASA) coordinates a system of “data” centres that operate according to different models, depending on the type of data being collected and analysed and the collaborators involved. In 2005 a new system was proposed that introduced the concept of “Resident Archives”. This was to ensure the long-term preservation of both NASA related data and derived products. All NASA related missions require the development of a data management plan covering the period from collection to permanent archiving.

The NASA system includes a permanent, funded Data Centre for long-term archiving of data – the National Science Space Data Centre (NSSDC). Affiliated agencies may also have Centres which act as permanent archive facilities for some types of mission data (by agreement). In some cases these facilities are outside of the USA.

“Resident Archives” sit close to the project and its personnel and may be organised around an instrument, a theme or an entire project. It would also be possible to have a Collaboration of Resident Archives (CRAs). A Resident Archive is envisaged to have a finite life-span, centred around the requirements for which it was established, but can morph into a permanent archival facility if funded, desirable and practical.

A Resident Archive would assemble intermediate and final data products and make them accessible to a permanent archive at the end of the Resident Archive’s life, as well as serve the data and products to the community in the intervening period. A resident archive therefore performs the following role:

- produces as complete a set of data products as possible (either new or improved, comprehensive, high time resolution, high quality) to the stage where they can be served to users;
- ensures that the mission (or project) data are served to the general space and solar physics community in an efficient and scientifically useful interoperable manner consistent with community data and infrastructure standards (e.g. those for virtual observatories) using readily sustainable, automated software;
- maintains the integrity of the data by safeguarding against data loss which could be effected by providing a mirror site or other mechanisms,
- documents (metadata) concerning the data and products (including mission, pi and other types of information) as required to maintain independent usability;
- obtains community feedback on the services of the archive to ensure its success;
- makes sure that the data will be archived after the Resident Archive is no longer needed (e.g., preserved by transferring to another Resident Archive, the NSSDC or other approved long-term repository).

The Resident Archive is expected to have two modes of operation; a startup phase and an operational phase. In the former phase, the above six functions are the prioritized set, while in the latter phase, the development of new/improved products becomes secondary to the continual serving of the data to the community.

In the NASA model it is anticipated that there are funds for the establishment of Resident Archives and the allocation of grant monies associated with this is managed by the NSSDC. The NSSDC also creates a Resident Archive User Group whose function it is to refine the criteria used in the peer review process (of Resident Archives), share common problems and solutions such as the methods to develop metrics for usage of data, evolution of services such as virtual data products, and the priorities for distribution of funds.

References

1. Academy of Medical Sciences (UK) *Report on Personal Data for Public Good: Using health information in medical research*. January 2006
2. Dr Penny Allbon, Australian Institute of Health and Welfare, submission to the Working Group, 6 September 2006
3. ARC, NHMRC, AVCC *Australian Code for the Responsible Conduct of Research – Second consultation Draft, February 2006* Revision of the Joint NHMRC/AVCC Statement and Guidelines on Research Practice [online] available from http://www.nhmrc.gov.au/publications/_files/acrcr.pdf; Information about the Review process is available [online] at: <http://www.nhmrc.gov.au/funding/policy/code.htm>
4. Australian Academy of Science, *A Submission to the Prime Minister's Science, Engineering and Innovation Council Working Group on Data for Science* 11 August 2006
5. Australian Social Science Data Archive [online] available at <http://assda.anu.edu.au/>
6. Australian Water Resources 2005 Information and System Gaps Limitations [online] available at http://www.water.gov.au/Keymessages/InformationAndSystemGapsLimitations/index.aspx?Menu=Level1_1_9
7. Australian Water Resources 2005 Status of Information [online] available at http://www.water.gov.au/Keymessages/StatusOfInformationIn2004_05/index.aspx?Menu=Level1_1_2
8. Bibbo M, Haenszel WM, Wied GL, et al. *A twenty-five-year follow-up study of women exposed to diethylstilbestrol during pregnancy*. N Engl J Med 1978;298(14):763-767.
9. Bishun NP, Smith NS, Williams DC, Raven RW. *Carcinogenic and possible mutagenic effects of stilboestrol in offspring exposed in utero*. Journal of Surgical Oncology 1977;9(3):293-300
10. Bramble D. Annotation: *The use of psychotropic medications in children: a British view*. J Child Psychol Psychiatry 2003;44(2):169-179.
11. Brook E, personal communication with Professor Fiona Stanley 2006
12. Brook EL, Rosman DL, Holman CDJ, Trutwein B. *Summary report: research outputs project, WA data linkage unit (1995-2003)*. WA data linkage unit, Department of Health 2005
13. Markus Buchorn and Paul McNamara, *Sustainability Issues for Australian Research Data: the Report of the Australian e-Research Sustainability Survey Project*, APSR Publications, October 2006 [online] available from <http://hdl.handle.net/1885/44304>; accessed 20 October 2006
14. The Hon Peter Costello MP *Address to the Australian Bureau of Statistics Centenary Celebration, Canberra* delivered 8 December 2005 [online]; available from <http://www.treasurer.gov.au/tsr/content/speeches/2005/019.asp>; accessed 26 September 2006

15. Council for Science and Technology (UK) *Better use of personal information: opportunities and risks* 2005 [online] available from <http://www2.cst.gov.uk/cst/reports/#10>
16. The Data Linkage System [online] available at <http://www.publichealth.uwa.edu.au/welcome/research/dlu/linkage/system>
17. Defence Science and Technology Organisation, submission to the Working Group
18. Department of Communications, Information Technology and the Arts, submission to the Working Group
19. Department of Education, Science and Training (DEST) and the Department of Communications, Information Technology and the Arts (DCITA), *The e-Research Coordinating Committee's Interim Report* September 2005: Chapter 4, p14-17 [online] available from: www.dest.gov.au/NR/rdonlyres/B6F765A7-DD2C-432B-9064-2F9CD4E17E66/10518/InterimReport2.doc; accessed 1 May 2006
20. Department of Environment and Heritage. submission to the Working Group
21. Department of Environment and Heritage (DEH), *State of the Environment Report* 2006, December 2006 [online] available from <http://www.deh.gov.au/soe>
22. eSecurity Framework [online] available at <http://www.esecurity.edu.au/>
23. Dr Rhys Francis, NCRIS Platforms for Collaboration Facilitator, submission to the Working Group, July 2006
24. Giusti RM, Iwamoto K, Hatch EE. *Diethylstilbestrol revisited: a review of the long-term health effects*. *Annals of Internal Medicine* 1995;122(10):778-88.
25. The governments (including the European Community) of Australia, Austria, Belgium, Canada, China, the Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Luxembourg, Mexico, the Netherlands, New Zealand, Norway, Poland, Portugal, the Russian Federation, the Slovak Republic, the Republic of South Africa, Spain, Sweden, Switzerland, Turkey, the United Kingdom, and the United States, *Declaration on Access to Research Data from Public Funding*, 30 January 2004 in Paris: Appendix XXX [online] available from <http://www.codataweb.org/UNESCOmtg/dryden-declaration.pdf>; document provided in the Academy of Science submission to the Working Group. Full text available at Annex 1
26. Tony Hey and Anne Trefethen, *The Deluge of Data: An E-Science Perspective*, UK e-Science Core Programme. Ch36 in *Grid Computing* Edited by Fran Berman, Geoffrey Fox, Tony Hey; in *Wiley Series in Communications Networking & Distributed Systems*, edited by David Hutchison; 29 May 2003: Digital Object Identifier: 10.1002/0470867167.ch36; [online] available from <http://www3.interscience.wiley.com/cgi-bin/summary/104535645/SUMMARY>; accessed 15 September 2006
27. Holman CDJ, Bass AJ, Rouse IL, et al. Population-based linkage of health records in Western Australia: development of a health services research linked database. *Aust N Z J Public Health* 1999;23:453-459.
28. Imaginova *Space.com* [online] available from <http://www.space.com>

29. International Council for Science (ICSU), *ICSU Report of the CSPR Assessment Panel on Scientific Data and Information 2004* [online] available from http://www.icsu.org/1_icsuinscience/DATA_Paa_1.html; accessed 23 August 2006
30. Lenz W. *Thalidomide and congenital abnormalities*. *Lancet* 1962;1:45.
31. Michael Lesk, <http://archiv.twoday.net/stories/337419/> 2004. [Accessed 4 December 2005] quoted in Microsoft Research Cambridge, *Towards 2020 Science* [online]; available from http://research.microsoft.com/towards2020science/downloads/T2020S_ReportA4.pdf; accessed 30 October 2006
32. Richard Macey, *One giant blunder for mankind: how NASA lost moon pictures* The Sydney Morning Herald, August 5, 2006 [online] available from <http://www.smh.com.au/news/national/one-giant-blunder-for-mankind-how-nasa-lost-moon-pictures/2006/08/04/1154198328978.html>; accessed 6 September 2006
33. MAMS is the Meta Access Management System. It was funded by DEST's Systemic Infrastructure Initiative (part of BAA 1) in 2003. This project allows for the integration of multiple solutions to managing authentication, authorisation and identities, together with common services for digital rights, search services and metadata management. More info [online] at <https://mams.melcoe.mq.edu.au/zope/mams>
34. McGettigan P, Henry D. *Cardiovascular risk and inhibition of cyclooxygenase: a systematic review of the observational studies of selective and nonselective inhibitors of cyclooxygenase 2*. *JAMA* 2006;296(13):1633-1644.
35. Microsoft Research Cambridge, *Towards 2020 Science* [online]; available from http://research.microsoft.com/towards2020science/downloads/T2020S_ReportA4.pdf; accessed 30 October 2006
36. The Molecular Medicine Informatics Model (The Bio21 : MMIM platform) [online] <http://mmim.ssg.org.au>
37. National Aeronautics and Space Administration (NASA), Jet Propulsion Laboratory, *Mars Exploration Rover Fact Sheet* [online] available from <http://mars.jpl.nasa.gov/missions/present/2003.html>
38. National Health and Medical Research Council (NHMRC) Section 95A Privacy Act 1988
39. New Scientist, *Dicing with Death*, 29 July 2006
40. Open Access (OA) [online] available at <http://www.eprints.org/openaccess/>
41. Organisation for Economic Co-Operation and Development (OECD) *Principles and Guidelines on Access to Research Data from Public Funding*, Science, Technology and Innovation for the 21st Century. Meeting of the OECD Committee for Scientific and Technological Policy at Ministerial Level, 30 January 2004 - Final Communiqué [online] available from: http://www.oecd.org/document/0,2340,en_2649_34487_25998799_1_1_1_1,00.html; accessed 14 June 2006 Annex 1 available at Appendix B
42. Open Access <http://www.eprints.org/openaccess/>
43. Queensland University of Technology (QUT) *Measuring the research impact of your publications: citation indexes and alternatives, Strategies to increase citations to your*

publications [online] available from
http://www.library.qut.edu.au/subjectpath/citation_indexes.jsp#strategies; accessed 24
October 2006

44. Square Kilometre Array [online] available at <http://www.skatelescope.org/>
45. Trutwein B, Holman CD, Rosman DL. Health data linkage conserves privacy in a research-rich environment. *Ann Epidemiol* 2006;16(4):279-280. N.B. the box on the Mars Rover's not referenced