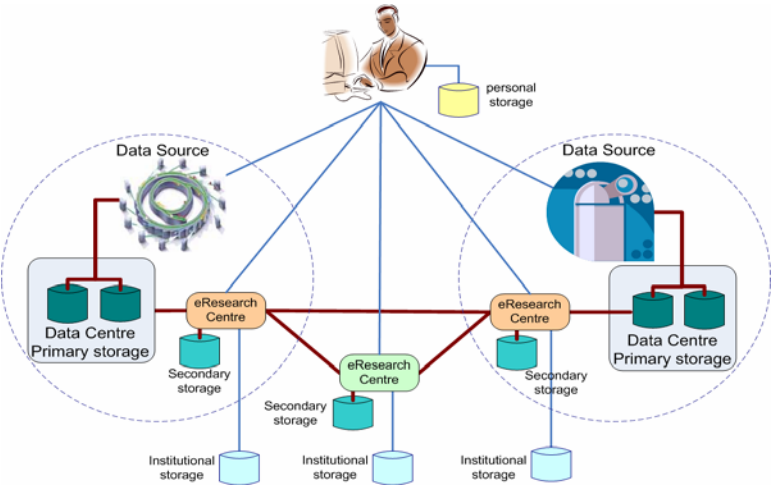


A National Data Architecture for Australian Research

Victorian Perspective



Prepared by Multimedia Victoria

July 2005

© The Victorian Government – 2005

This publication is copyright. Other than for the purposes of and subject to the conditions prescribed under the Copyright Act, no part of it may in any form or by any means (electronic, mechanical, microcopying, photocopying, recording or otherwise) be reproduced stored in a retrieval system or transmitted without prior written permission. Inquiries should be addressed to –

The Executive Director
Multimedia Victoria
Department of Infrastructure
Government of Victoria
Melbourne

Reliance and Disclaimer

The professional analysis and advice in this report has been prepared by Multimedia Victoria for the exclusive use of the party or parties to whom it is addressed (the addressee) and for the purposes specified in it. This report is supplied in good faith and reflects the knowledge, expertise and experience of the personnel involved. The report must not be published, quoted or disseminated to any other party without Multimedia Victoria's prior written consent. Multimedia Victoria accepts no responsibility whatsoever for any loss occasioned by any person acting or refraining from action as a result of reliance on the report.

In conducting the analysis in this report Multimedia Victoria has endeavoured to use what it considers is the best information available at the date of publication, including information supplied by the addressee. Unless stated otherwise, Multimedia Victoria does not warrant the accuracy of any forecast or prediction in the report. Although Multimedia Victoria exercises reasonable care when making forecasts or predictions, factors in the process, are inherently uncertain and cannot be forecast or predicted reliably.

For information on this report

Please contact:

Alastair Crow

Multimedia Victoria

Telephone (03) 9651 9275

Email alastair.crow@mmv.vic.gov.au

Table of Contents

Executive Summary	5
1. Introduction	6
1.1 Background to the Report	6
2. The Global e-Research Agenda	9
3. Data Sources for Australian Researchers	10
4. Data Production	12
4.1 Data Produced at Major Research Facilities	12
4.1 Data Produced by Compilation	12
5. Grid Services Identified by Researchers	14
6. The Need for a National Data Architecture	16
7. Elements of the Architecture	18
7.1 Accessibility Grid	19
7.2 Data Grid	23
7.3 Semantic Grid	24
8. Development of Skills Sets for e-Research	29

Executive Summary

The e-Research agenda for Australia requires the establishment of an integrated strategic approach to linking the researchers' activities in research and development, and its consequent data production and access needs, with major scientific instruments and data sets. The evolving Australian research and education network and computing environments require focussed attention on developing secure, authenticated access to, and storage of, information.

This report summarises the Victorian Government's extensive investigation of the needs for a strategy for a national data architecture for Australian researchers.

The key elements of this architecture are the accessibility grid, the data grid and the semantic grid. While the main technical features of these elements are presented in this report, and the required services, further work is required to establish a strategy and implementation plan for the development of these grid elements.

The need for outreach and skills development in the conduct of these actions is identified as a critical enabler.

This report has been prepared by Multimedia Victoria, (an agency of the Department of Infrastructure within the Victorian Government), at the request and direction of The Hon Marsha Thomson, MLC, Minister for Information and Communication Technology in the Victorian Government.

The Minister has submitted this report to the Minister for Communications, Information Technology and the Arts, Senator Helen Coonan to be passed onto the recently established e-Research Coordinating Committee (established jointly with the Federal Minister for Education, Science and Training) for its consideration and implementation.

The Victorian Government would like to thank Dr Robert Hobbs, Professor John O'Callaghan, Dr Don Robertson, Dr Paul Carr, Dr Richard Garrett, Dr Nick Hauser, Professor Ah Chung Tsoi, Robert O'Connor, the Council of the Australian University Directors of Information Technology (CAUDIT) and those many scientists/researchers who contributed to this report. The Victoria Government would especially like to thank Dr Mike Sargent and Paul Davis for their major contribution to the development of this report.

1. Introduction

This report proposes the establishment of a national data architecture for Australian research and details the possible approaches for the implementation of such a project.

This document is structured as follows:

Section 2 briefly examines the global e-Research agenda and key components of e-Research infrastructure required to enable collaboration between Australian researchers and their colleagues world-wide.

Section 3 explores Australian researcher production and utilisation of data sources in their research. Examples of the sorts of data sources and data sets include: synchrotrons, research reactors, telescopes, microscopes, and medical, humanities and earth sciences data sets.

Section 4 explores the different data storage requirements and challenges for data produced at major research facilities in comparison to data produced by compilation.

Section 5 and 6 highlights the range of ‘grid’ services important to researchers and the need for a comprehensive, flexible and collaborative approach and program to serving the future data management needs of Australian research.

Section 7 describes the integrated elements that form the proposed national data architecture for Australian Research.

Section 8 highlights that a critical element in the development of a national data architecture will be in the implementation of a comprehensive programme of skills development of information and communications technology professionals and researchers at undergraduate and postgraduate level.

1.1 Background to the Report

The Australian Government has recently established initiatives (for example the National Collaborative Research Infrastructure Strategy or NCRIS) to develop strategy and policy to coordinate support for research utilising advanced research infrastructure. The goals of such initiatives include achievement of national research outcomes, capacity building to participate in national and international research programmes, and improved collaborative research activities.

The recently announced e-Research Coordinating Committee (eRCC) as an initiative to coordinate e-Research in Australia, and inform e-Research investment decision made under other programs such as NCRIS.

The key elements of national research grid infrastructure are currently being implemented, following significant investment of systemic infrastructure funds by governments and research institutions. These include the Australian Research

and Education Network (AREN) and the regional research networks, the Advanced Network Program (ANP), the Australian Partnership for Advanced Computing (APAC) and the regional nodes such as VPAC, and the Australian Research Information Infrastructure Committee (ARIIC) initiatives.

The Victorian Government released its Broadband Framework in April 2005 with one of its key elements to encourage the transformative use of broadband technology in key sectors of the economy.

The 'grid' has been identified as a key component of transformative use of broadband in sectors such as health, education, primary industries and research. For example, in the research sector, grid will enable the connecting up of dispersed resources and the provision of services for researchers to access and share infrastructure such as specialist research instruments (synchrotron, microscopes, telescopes, research reactor, etc.), high performance computing power and data storage capacity. This involves a move to a new e-Research paradigm, requiring investment in new forms of software, applications and business process re-engineering.

Science, particularly that based on large instruments, increasingly involves distributed, global collaborations enabled by the internet and using very large scale data collections, high performance computing resources, tele-science (remote access and control of instrumentation) and collaborative visualisation. Collectively this is referred to as e-science. The grid is a new-generation information utility that brings together middleware, software and hardware resources from different administrative domains to access, process and store huge quantities of data. As such, the grid is enabling infrastructure for e-science.

The Victorian Government is scoping a program of grid activities and projects relevant to Victorian issues and needs. These activities will assist in developing the local skills and capabilities important for widespread deployment as the grid technology and its use become more embedded in government and commerce.

The Government has recently committed support for the Victorian Partnership for Advanced Computing (VPAC) and its regional education and research network which would be likely to be involved in grid projects supporting Australia's e-Research agenda.

The Government also considers national research facilities, such as the Australian Synchrotron and the Lucas Heights Research Reactor, to be a logical catalyst for a national grid project as they are major sources of data, have dispersed (national and global) digitally advanced user communities and are major infrastructure investments.

In October 2004, MMV convened a meeting of national science, research and technology organisations to discuss the issue of grid and its potential impact on research facilities such as the synchrotron.

It was agreed that a 'working group' comprising representatives from MMV, the

Australian Synchrotron Project, the Australian Partnership for Advanced Computing (APAC), Australia's Academic and Research Network (AARNet), and Australian Synchrotron Beam-line Advisory Group, be established to develop the case (this report).

MMV formed a 'project team' with members including: Dr Mike Sargent, Director, Mike Sargent & Associates; Paul Davis, Director, GrangeNet and Alastair Craw, Project Manager, Multimedia Victoria.

This survey of researchers clearly indicated that there was a critical need to address the issues of data management and accessibility for Australian researchers to national research instruments/data sets.

This report summarises the conclusions reached following a recent extensive process of consultation with researchers who are potential users of facilities such as the Australian synchrotron and of various stakeholders associated with such research (such as governments and industry), and a survey of developments associated with similar major research facilities overseas.

2. The Global e-Research Agenda

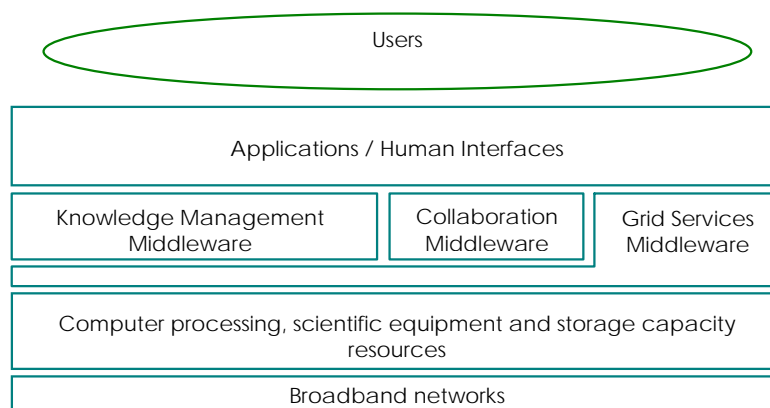
The rapid evolution of digital, information and communications technologies has created an environment in which the paradigms of research have changed. This change has been recognised internationally, and significant resources have been devoted in many countries to ensure that their researchers can remain globally relevant and competitive in this new environment. Over the past few years a number of initiatives in Australia and overseas have been taken to ensure that these opportunities can be effectively exploited – The Australian Partnership for Advanced Computing (APAC), the Australian Research and Education Network (AREN), and the Australian Research Information Infrastructure Committee (ARIIC).

At the same time the evolution of the capability and sophistication of scientific instruments and facilities has seen an explosion in the quantum of data produced by experimentation, and the complexity of analyses conducted using this data. The cost of leading edge research facilities is such that there is an increasing imperative to facilitate their use on a collaborative basis, and to provide remote access, both nationally and internationally, to the facilities. The establishment of the National Collaborative Research Infrastructure Strategy (NCRIS) recognises this need.

Research, particularly that based on large instruments, increasingly involves distributed, global collaborations enabled by the Internet and using very large scale data collections, high performance computing resources, tele-science (remote access and control of instrumentation) and collaborative visualisation. The enabling mechanisms for this e-Research are known under the generic title of ‘grids’. The ‘grid’ is a new-generation information utility that brings together middleware, software and hardware resources from different administrative domains to access, process and store huge quantities of data.

The interaction of these elements is shown below.

Figure 1: Key Components of e-Research Infrastructure



3. Data Sources for Australian Researchers

Australian researchers produce and utilise diverse data sources in their research, ranging from major data sourced from a limited number of discrete research facilities or data repositories, to extensive holdings of data of smaller dimension but held in many diverse and distributed ‘stores’ – for example, at national and international facilities, research institutions, government departments and authorities.

Examples of these sources and data sets include:

Synchrotrons

Australian researchers have used synchrotrons at many international locations over a number of years in a collaborative research activity funded by Australia. The construction of the Australian Synchrotron will provide increased access to synchrotron facilities for Australian researchers over a broad spectrum of research activity, covering, for example, areas such as structural and conformational analysis of proteins, advanced materials science, imaging of cell and plant tissue, minerals characterisation, lithography, etc.

Research Reactor

HIFAR (High Flux Australian Reactor) and the replacement research reactor OPAL (Open Pool Australian Light water reactor) - which will come on-line in 2006 – will provide more than 7000 hours per year in neutron beam time to scientists and students from Australia and overseas.

The research reactor is a major research tool for Australian scientists. Studies of the chemical structures and magnetic properties of materials are undertaken using neutron diffraction or scattering techniques and the research leads to outcomes such as the creation of stronger, lighter, more heat resistant materials for industry, and more advanced pharmaceuticals. The research activities will produce similar volumes of data to that produced at the synchrotron

Telescopes

Australia has a unique galactic view and very low levels of electronic and light pollution. Optical and radio telescopes across the continent engage in a variety of research activities including eRCBI (very long baseline interferometry), the Sloan digital sky survey, and studies of the distant universe (high energy astrophysics, high precision observation of stars, solar physics and solar system astrophysics). These activities are all capable of producing terabytes of data and use large amounts of super computer time.

Microscopes

Australia hosts several large instruments for atomic level and microstructural characterisation and manipulation. These include high voltage transmission electron microscopes, scanning electron microscopes, cryo electron microscopes,

electron probe micro analysers and microprobes, atomic force microscopes, Advanced Focussed Ion Beam Platform, x-ray microtomographs, Raman spectrometers and a range of light microscopes. Some of these instruments are already connected to tele-presence and remote control facilities, others will follow this trend and contribute to the mass of data produced by this class of instrumentation.

Medical Data Sets

Clinical MRI and Tomography has the potential to generate hundreds of terabytes of data per year particularly when coupled to bright light sources such as the Australian Synchrotron (at synchrotrons, data volumes will increase with increasing sophistication in sensor technology). Other data sets, from medical instruments, x-ray (eg the reference x-ray collection at Westmead), MRI and clinical experiments and trials are all candidates for data federation to enhance collaboration and data dissemination.

Australia should have a mirror of the Genbank (Los Alamos) bio data-set and widely accessible copies of other key bioinformatics reference sets. The Austrian Medical Grid (distributed applications for heart simulation, virtual lung biopsy and virtual eye surgery) and the Medical Grid in Japan (real-time, functional MRI data) are examples of data intensive grid activities.

Humanities Data Sets

The Humanities and the Arts are not normally associated with large data sets however, each State Library maintains substantial image collections, the PARADISEC Endangered Language project is amassing terabytes of data and ADFA houses the finest on-line collection of Australian literature. Researchers elsewhere now talk in space-time-language coordinates – federation of Australia's substantial holdings in this context would benefit from a National storage resource and the proposed Semantic Grid.

Earth Sciences

Geosciences and sensor arrays are rapidly expanding. Australia is part of the global EcoGrid project (supported by PRAGMA) with the US, China, New Zealand, Taiwan, Thailand, Japan etc. This project will improve global understanding of the interaction of climate, the oceans and lake systems and in doing so will collect and analyse huge amounts of data from thousands of remote sensors. This is one project typical of several others that involve substantial data streams from very large sensor arrays.

Within all of these categories, but with different emphasis, data sets are derived broadly in two ways, data:

- produced at major research facilities/instruments; or
- produced by compilation.

4. Data Production

4.1 Data produced at major research facilities

Where data is produced at a major research facility, researchers generally expect that there will be available data storage facilities (short-to-medium term storage and archive of raw experimental and pre-processed data). There is general recognition that the data will need to be transferred to other data centres for long term storage and archival, and that copies of the data and collaborative engagement with data sets will be necessary.

The requirements for this storage are:

- experimental data should not be lost or corrupted;
- data must be secured against unauthorised access;
- raw data should be archived for a number of years;
- there must be reliable mechanisms for making copies of the data for remote and collaborative use;
- copies of the applications software used to pre-process raw data should be archived;
- it must be possible to access data at any time using simple, intuitive tools; and
- it must be possible to incorporate data from many sources (such as experimental logs), including those external to the facility (such as synchrotron and reactor data).

Researchers also generally anticipate (and or would value) an interaction with researchers at the data generating facility(s) and with other collaborators at other locations using grid based video conferencing tools.

For example, researchers can envisage projects where the primary data collection is managed remotely and the principals in the research team interact with a subgroup of the team at the research instrument who are responsible for reasonably automated data collection. That is, researchers would value real-time video interaction with experimental process, real time remote receipt of experimental output, and real time collaborative analysis services such as data visualisation, as part of the remote control of experiments.

4.2 Data Produced by Compilation

Major data sets are being produced by compilation, or by federation of smaller data sets. This is typical of social science and humanities research, where for example, major data repositories of anthropological data are developing, the size

of which rivals those physical sciences research. Similarly, the development of extensive sensor networks (e.g., Ecogrid), and geosciences data, is resulting in the accumulation of large data sets in the natural science of a similar scale.

Generally these repositories are distributed in nature, and require significant collaborative effort to maximise utility.

In contrast to the data of the preceding section, which is typically ‘owned’ by individual researchers or research groups, the data of this type is ‘owned’ by many collaborating researchers and institutions. The challenges for this type of data storage therefore are not only the development and maintenance of standards and security, but also the management of data distributed across many institutions, and the management of collaborative access to the data. There is therefore a high dependence on systems management of data.

The requirements for this storage are similar to the first group, viz:

- original data should not be lost or corrupted;
- all data must be secured against loss, corruption and unauthorised access;
- data should be archived for a number of years, perhaps even decades or centuries;
- there must be reliable mechanisms for making copies of the data for remote and collaborative use;
- copies of the applications used to process data should be archived;
- it must be possible to access data at any time using simple, intuitive tools;
- it must be possible to incorporate data from many sources and many repositories; and
- preservation and curation of data are required.

5. Grid Services Identified by Researchers

A number of surveys and reviews have been undertaken over the past few years to identify the key issues and needs of researchers in relation to the creation, management and analysis of data sets. These include work undertaken under the auspices of ARIIC, APAC and GrangeNet, and more recently an extensive survey and consultation with researchers who are users of synchrotrons and research reactors.

Given the diverse nature of research undertaken on even one instrument such as a synchrotron, it is not surprising that there is a broad spectrum of stated needs and priorities for research grid services. These reviews indicate some broad common themes in respect of data management and access.

The lack of access of Australian researchers to extensive grid architectures, nationally and internationally, has led to scientific processes that often are oriented towards:

- personal storage of data;
- analysis based on use of personal equipment;
- data transfers by physical means (with delays between production and analysis measured in terms of days or weeks); and
- involving collaborations that are static or pre-ordained in nature.

The aforementioned development of modern grid architectures in Australia means that more dynamic, timely and sophisticated data management and analysis is possible, and more dynamic collaborative activities will be possible.

The range of grid services important to researchers are detailed in the table below.

Table 1: Important Grid Services

Grid Services	Functions
telepresence	<ul style="list-style-type: none"> – connecting sites/sources – experimental recording and annotation – remote interaction with instrument scientist/operator from virtual instrument laboratory/desktop for management/collaboration, training, etc
remote operations	<ul style="list-style-type: none"> – instrument is an ‘online device’
data transmission	<ul style="list-style-type: none"> – security – speed – integrity
data storage	<ul style="list-style-type: none"> – raw data from experiment/test

	<ul style="list-style-type: none"> - derived data (eg from simulation/analysis) - instrument data (eg calibrations, characteristics) - experiment logs, including video data, audio data, annotations - standards and format - time to retain raw data, derived data, etc
data management	<ul style="list-style-type: none"> - original, back-up, archival, retrieval - retrieval parameters of active data - level of back-up, media used, time to hold - archival media, time to hold, retrieval notice required
linking data from diverse and multiple sources	<ul style="list-style-type: none"> - links to data from diverse science instruments/facilities
data manipulation and annotation	<ul style="list-style-type: none"> - 'standard' analytic software - version control - archival of analytic software - annotation tools
data security and audit trail	<ul style="list-style-type: none"> - researchers ownership of data - flexible control of access by owner - time/source/destination tags for original data and its copies/derivatives
user authentication	<ul style="list-style-type: none"> - unique user key, globally accepted Public Key Infrastructure (PKI) - authorisation of access by data owner - management of access by storage provider
virtual organisation management	<ul style="list-style-type: none"> - flexible, secure collaboration tools - session recording and annotation
computational and simulation capability	<ul style="list-style-type: none"> - real time analysis, particularly for the purposes of assuring/adjusting experiment
easy access to resources (data, communications, computation capacity, computational software).	<ul style="list-style-type: none"> - High amenity, low cost, production level services with low barriers to use (eg. web services)

Researchers require production quality data management processes, with support by the service provider. Simple data storage, access and retrieval processes, and resource (data stores, analytic and visualisation software, network allocations) acquisition and utilisation systems are essential. A one-stop-shop approach to this is seen as desirable for the researchers, so that they do not need to sequentially marshal resources, or need to search for resource location.

Cost of the data management service must be low - at present most researchers use portable hard disc drives and laptop computers to handle their data needs, and this sets the price expectations.

6. The Need for a National Data Architecture

From this it is clear that there is an exigent need to assist the scientific community with the development of significant data storage facilities as a component of Australia's strategic research infrastructure.

These needs can be grouped as:

- **at the point(s) of data acquisition:**
 - tele-presence;
 - remote instrumentation;
 - data transmission;
 - primary data storage;
 - data management;
 - data manipulation and annotation;
 - data security and audit trail; and
 - user authentication.
- **at the location(s) of data storage:**
 - primary data backup;
 - data management;
 - linking various data/instrument sources of data;
 - data manipulation and annotation;
 - data security and audit trail;
 - user authentication; and
 - virtual organisation management.
- **at the location(s) of data analysis:**
 - data access;
 - data security and audit trail;
 - user authentication;
 - virtual organisation management;
 - computational and simulation capability;
 - derived data; and

- visualisation of data and results.

Obvious concerns associated with the exponential growth of data volumes and complexity are those of describing and cataloguing the data, rights management, resource discovery and curation. Traditional methods do not scale and are an ill fit to the needs of managing very large databases. The emerging solution to these problems is the Semantic Grid.

“The Semantic Grid is an extension of the current Grid in which information and services are given well-defined meaning, better enabling computers and people to work in cooperation”

The Semantic Grid: Past, Present and Future

D. De Roure, N.R. Jennings & N.R. Shadbolt

Proceedings of the IEEE, VOL. 93, No. 3, March 2005.

Practically the Semantic Grid provides a methodology for describing data, managing its storage, intelligently linking it with related information and delivering context-aware decision support and inference services. This enables information...

“...to be presented to users at the right time, in the right format, on the right device and with the right level of intrusiveness.”

The Semantic Grid: Past, Present and Future

D. De Roure, N.R. Jennings & N.R. Shadbolt

Proceedings of the IEEE, VOL. 93, No. 3, March 2005.

The breadth of need for Australian researchers to acquire, access and analyse data, and to collaborate nationally and internationally in research activities, indicates that a comprehensive, flexible and collaborative approach and program to serving the future data management needs of Australian research is essential. This program needs to be seen in the context of similar international activities and Australia needs to engage with and participate in those international activities.

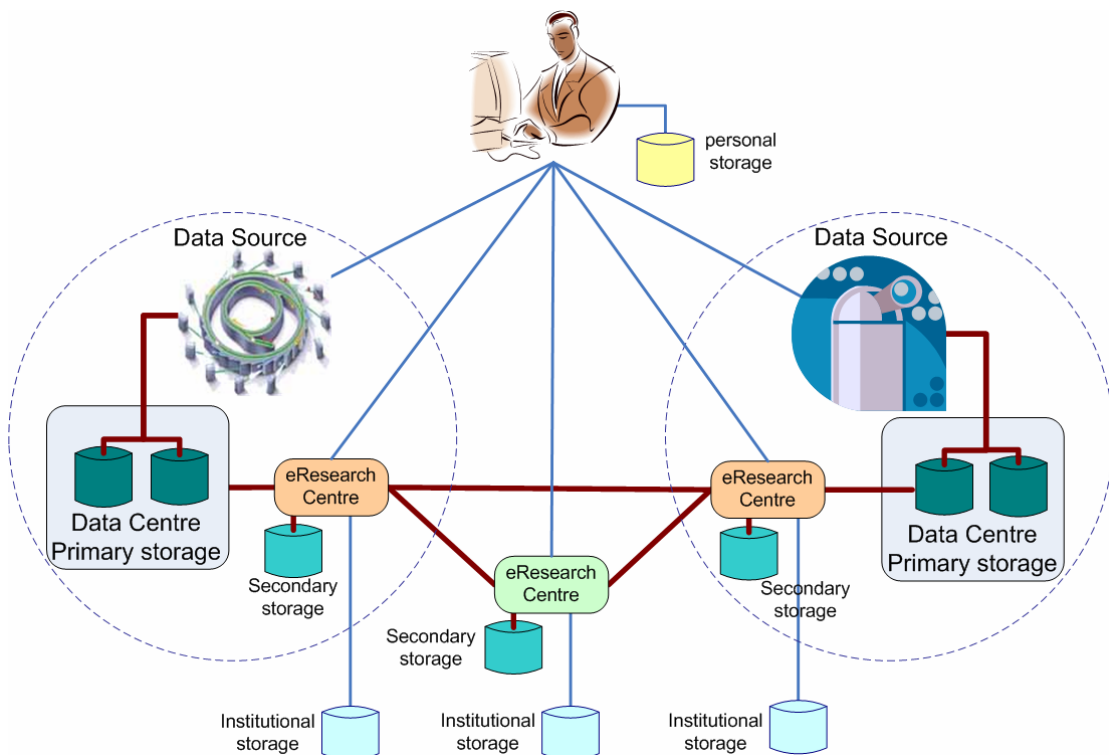
7. Elements of the Architecture

The above categorisation indicates that there are three broad but integrated elements to the National Data Architecture:

- an **Accessibility Grid** which manages access to data and facilities, and collaboration between researchers, in a secure and flexible manner; this conceptually is based on e-Research Centres (eRC) incorporating virtual laboratories (VL), and use of personal Access Grid nodes providing a secure access environment;
- a **Data Grid** which manages data storage, transfer, backup, archiving, etc; this will be based on an architecture with four main storage elements; and
- a **Semantic Grid** which facilitates access to and collation of resources within the Data Grid and associated computing and instrument grids, and which organises the data into “experimental containers” managed and maintained by the Data Grid.

The diagram below highlights the interaction of physical components underpinning these grids.

Figure 2: Physical Elements of the National Data Architecture



Components within the circles would be typical of a major data node (either a major research facility or a major data repository). The institutional data stores represent distributed and diverse data sources held by researchers and institutions. The red lines represent dedicated, very high bandwidth networks links; the blue lines represent normal network links. The closed 'private' network represented by the red lines also serves as a security perimeter, designed to ensure that there is no user/intruder access to this network. Operationally, secure agents invoked through portals will move files around the network. This design feature is common in secure data centres.

The user's local computing environment is enhanced by access to the secondary storage resources on the grid at the virtual laboratories within the e-Research Centres, where advanced resources – visualization, computing, data copying, room sized Access Grid Node and higher network speeds – are available. It is envisaged that there will be e-Research Centres and/or virtual laboratories at major data sources and one or more serving each region.

The user, through a collection of portals and web tools, can:

- use virtual organisation management tools from the desktop to control or monitor remote experiments and interact with other researchers located either at their desks or at an e-Research Centre or virtual laboratory;
- move data from the Data Centre to the secondary storage attached to the e-Research Centres;
- read and write data from the e-Research Centre storage system to local storage; and
- make copies of data sets locally or using the output media resources at an e-Research Centre.

The Semantic Grid keeps track of who has accessed and copied which files and when, how data has been processed and what applications were used.

7.1 Accessibility Grid

The **Accessibility Grid** manages access to data and facilities, and collaboration between researchers, in a secure and flexible manner. It allows many-to-many interactions and communications simultaneously, and thus will enable researchers from a number of sites to share and discuss data and experiments in real time. This ability for instantaneous communication among researchers will facilitate improved collaboration, and improved accessibility to research facilities.

The key requirements for effective establishment of the Accessibility Grid are:

- development of access and authentication standards and frameworks;
- development of data and software standards;

- development of frameworks for virtual organisation management; and
- implementation of accessibility features such as remote working and telepresence.

Authentication and Access

Security is clearly a concern. All researchers consulted noted that security of data, as well as flexible control of access is essential. As the grid services architecture will be a national (and an international) architecture, there will be a need for institutions to collaborate in fitting this environment into the established IT infrastructure and policies of research institutions (firewalls, access controls, capacity limits, charging).

In the proposed architecture there are two security domains – the major primary data centres and the e-Research Centres. At the major primary data centres the main security functions are that of physical security and data integrity. Scientists expect that the data they collect will be safely stored and secured against loss and corruption and will be available on demand. At the e-Research Centres and beyond the expectation is that authentication and authorisation of users will protect data from harm and unauthorised access.

The physical security of the major primary data centres and the integrity of data stored therein will be one of the primary foci of the data centre management. Security in this domain is enhanced by virtue of the ultra-fast private network interconnecting the instrument data sources, the data centres and the e-Research Centres.

The security aspects of the e-Research Centres and user interfaces require careful design and deployment if they are to meet the requirements of the users without being unduly obstructive.

It is anticipated that the work commenced under the auspices of ARIIC, and the CAUDIT PKI project will form an internationally recognised authentication and access regime and Certificate Authority for Australian researchers. It is expected that certificates issued under this authority will form the basis of authorisation and access and that portals providing access to the e-Research Centres and secondary data stores will be certificate based.

Locational Flexibility

There is a demand for locational flexibility in the conduct of research – that is researchers should not have necessarily travel to data, computational or instrument resources to conduct their research. Achieving this will require provision of a range of tools including remote operation and management of experiments, tele-presence tools, and virtual organisation management tools.

The extent to which remote control and monitoring is implemented will essentially be the province of the researchers through the various networks and communities of interest, such as the Australian Synchrotron Beamline Advisory

Panels, the Molecular and Material Structure Network, etc. These groups will need to design and deliver tools and procedures for remote access and control of instruments. It is expected that commercial instrumentation will come with proprietary interfaces for remote access.

There are a number of components to such a capacity:

- network access;
 - The network must have sufficiently high bandwidth, which can be allocated if high bandwidth is required. This is best achieved using quality of service facilities, which come with network backbone switches. However, much work will need to be performed to determine how such quality of service can be enabled, as well as controlled, perhaps under user control. This is largely uncharted territory, and much work remains to be performed in order that remote control of instruments can be enabled effectively.
- integration of remote interaction tools with an overall workflow management system under the control of the user;
- portals designed to set up access grid sessions and standard hardware/software platforms need to be developed; and
- virtual organisation management tools, which must be integrated with the workflow management system.

Data and Software Standards

The diversity of research to be conducted using major research facilities and major and distributed data sets indicates that it would not be realistic or feasible to attempt to develop a single set of data and software standards that would apply to all such research activities. However, increasing collaboration amongst researchers requires that standards for data storage and software (including version management) will need to be established to serve cohesive groups of researchers.

While certain research groups have some data and software standards defined, there will be a need to have cohesive groups of researchers to develop such standards.

The grid services should include a software management facility, which maintains and supports both standard software and experimental software.

e-Research Centres (Virtual Laboratories)

The e-Research Centres provide:

- a focal point for facility activities remote from the facility;
- a tele-presence / Access Grid environment for collaboration, remote control of data collection, training and pre-visit preparation;

- remote instrumentation;
- a site where advanced visualisation and computing resources can be shared;
- the connection point for managed storage resources;
- data transcription where scientists can make copies of critical data on removable media;
- dedicated (IT) staff to manage facility and provide user support;
- outreach activities for researchers skill development;
- a focus for research associated with support of e-Research activities; and
- a focus for skills development of IT professionals to support e-Research activities.

The e-Research Centres will be modular in nature, allowing for systematic and targeted development. As an indication of cost, it is expected that a basic facility with a room-based Access Grid node, a 100TB storage unit, servers and graphics terminals, DVD and tape writers would cost of the order of \$1M to build and \$0.5Mpa to run. Research support would be additional to this.

A major function of the e-Research Centres is to provide a nexus for the connection of secondary storage, institutional and personal storage. The network links into the e-Research Centres would be part of a private high speed network obviating the need for firewalls and facilitating the transfer of data from the major primary data centres to the local storage system. As for security and integrity reasons, individuals researchers would not be permitted to write data to the Data Centres, the secondary storage at the e-Research Centres provides researchers with storage close to their laboratory that they can both read and write to. It is anticipated that universities will attach their own storage systems to the e-Research Centre private network further enhancing the data grid.

Many of the prospective users of the major research facilities expressed a desire to change the way in which they conducted their experiments. Most agreed that it was essential for students and new facility users to visit the facility as a learning experience, however, they wanted to be able to mentor the data collection and to interact with collaborators during data collection without having to travel. The e-Research Centres cater for this approach.

It is anticipated that the universities in major centres remote from the major research facilities (eg Perth, Adelaide, Brisbane, Sydney, Hobart and Canberra) would jointly fund, with State and Federal government assistance, a communal facility using a model similar to a Regional Network Organisation (RNO).

7.2 Data Grid

The Data Grid comprises four main components:

- major primary data centres;
- secondary storage at the e-Research Centres;
- federated data sets held by individual research institutions; and
- personal data sets held by individual researchers.

It is expected that data-sets under the stewardship of the research community for local copies of key databases such as the Brookhaven Protein Database (PDB) and The Cambridge Structural Database (CSD) would be held either at (one or more of) the major primary data centres or e-Research Centres.

Major Primary Data Centres

In addition to any facilities provided to store data at the major data sources, raw experimental data and data pre-processed at the data source facility, or acquired from other data sources, would be written to the major primary data centres' storage systems. Users would have read-only access to major primary data centres using portals to identify file sets and move copies of the archived data to secondary storage and local computers. The Data Centre is the key interface between the data sources and the data grid.

The Data Centres must be professionally run data resources for storing and archiving facility data. Based on established Hierarchical Storage Management (HSM) practice, data would be retained on the Data Centre Storage Area Network (SAN) for a period of time in which there is high activity in accessing the data, after which it will be automatically migrated to a side-line tape silo. After a further period, the data tapes will be warehoused at a commercial data vault. While the SAN has to be close to the major data source, the tape silo and data vault could be connected to the SAN via high-speed data links and sited anywhere on the Australian mainland to which there is a high performance, high reliability gigabit network. Tape silos exist elsewhere within the scientific community (APAC, QPSF, CSIRO) and could be incorporated into the storage hierarchy. With the network infrastructure under development by AARNet, it should be possible to recover data from side-line (tape silo) storage within tens of minutes. Commercial data vaults are available in all capital cities – most offer various tape recovery services and next-day services are relatively inexpensive.

To achieve the very high reliability and availability demanded by the users, the Data Centres will operate as a storage area network (SAN) distributed over two sites with identical storage elements, each a mirror of the other and interconnected by dedicated fibre. Where the Data Centre services a major research facility, such as the synchrotron or research reactor, one of the storage elements should be sited at the facility for reliability, the other placed in a suitable

site within a 50km radius of the major data source. The distance constraint is imposed by the gigabit interfaces used to terminate the optical fibre.

Preliminary estimates indicate an initial capacity of major primary data centres should be approximately 300TB. Two pairs of dark fibre should be provided to establish the SAN.

The indicative cost for a major primary data centre, including servers, disc units, SAN fabric, network equipment and output devices and HSM software is \$5M, with an annual operating cost of approximately \$2M (including recurrent costs for the HSM software licenses).

It is recommended that the data centres be professionally managed by a not-for-profit company limited by guarantee. The management company, which as a national service should be established with the support of the Australian Government, needs to provide enduring service sustained by the academic/research sector who could be shareholders of the entity.

The company should have a board selected from recognized experts in the area and vendors and a high profile chairman. Personnel would include a CEO, CTO, Finance manager and two or three technicians. The Board and management of the data centres management company would be supported by standing committees, such as:

- a research committee that meets annually and elicits the user community's needs; and
- a technical committee to interface with vendors.

7.3 Semantic Grid

Major scientific facilities such as sensor networks, telescopes, Hadron Large Collider and synchrotrons are capable of producing massive amounts of data and many scientific communities are struggling with the issues of long term archive and the accessibility of large data sets. The Virtual Observatory Alliance provides a good example of the difficulties associated with retro-fitting metadata and data management to a large data set.

The data centric (e.g. Storage Resource Broker) and XML (e.g. the Metadata Catalogue Service) approaches to data management are limited and there is an increasing body of work emerging on the use of *semantics* to describe data grids and data grid services.

“The Semantic Grid is an extension of the current Grid in which information and services are given well-defined meaning, better enabling computers and people to work in cooperation”

The Semantic Grid: Past, Present and Future

D. De Roure, N.R. Jennings & N.R. Shadbolt

Proceedings of the IEEE, VOL. 93, No. 3, March 2005.

The UK Digital Curation Centre (DCC) is developing a layered approach to the curation of scientific data based on an Open Archival Information System (OAIS) model where ‘containers’ are used to package semantic information with data and metadata. This approach enables the management and maintenance of large scientific data sets using semantic grid technology. The resulting structure provides a method for encapsulating links to the raw data, its associated contextual and provenance metadata, processing steps and derived knowledge.

The semantic web offers the means for keeping track of who has accessed and copied files, of when and how data has been processed, of what applications were used, and of individual contributions to collaborations.

Extensive data logging facilities and audit trail will be necessary to keep track of who has accessed and copied which files and when. A workflow engine as part of the set of middleware incorporated into the system, will allow users to have access to an electronic log book, in which a record of how the data has been accessed, the types of analysis procedures which have been carried out, the databases accessed, etc.

Such information will allow scientists to have a record of what they have done, and can repeat the experiments at a later date if necessary. The data will be enhanced with extensive annotation and metadata. Such annotation and metadata facilitates the deployment of semantic grid technology. This technology will allow extraction of knowledge from the data.

In the context of an experiment at a facility such as the synchrotron, a “container” might encapsulate:

- the data from a series of experiments;
- metadata describing the experiments and raw data;
- the beamline characteristics and rig parameters for each data set;
- a video record of the experiment;
- the software used to reduce the data; and
- links to related data sets.

Every operation on the data – reading, copying, processing – would be recorded forming part of the audit trail for the data set.

Separate ‘containers’ would hold copies of all of the required software so that it is possible to recover both data and the applications needed to process and analyse detector output at any future time.

Research Program

Implementing a semantic grid is not a trivial exercise and research will be necessary to:

- investigate and recommend data standards for metadata and data formats;
- understand how the Web Ontology Language (OWL/OWL-S) will interface to the Globus Web Services Resource Framework (WSRF);
- develop ontologies to describe all grid components (instruments, people, computers, software, networks and storage facilities) and top level ontologies to harmonise these semantic descriptions;
- formulate rules to enable the optimum assembly of grid components for a given collaboration;
- design agents capable of dynamic discovery, invocation, composition, and choreography of most appropriate grid resources and services;
- develop mechanisms for determining when new data is available which impacts on their existing "scientific models" and notifying relevant person or agent; and
- build mechanisms for linking facility data to publications or on-line learning objects.

The development of the semantic grid will require a strategic and coordinated research program, utilising established Australian research strengths (there are Australian researchers with international reputation and collaborations in this area capable of carrying out this development) and strong international collaboration.

This program should be developed incrementally through a number of phases. Phase 1 provides research support for measures that can be implemented today using available technologies. Phases 2, 3 and 4 provide the broad phased research program for progress to semantic grid.

Table 2: Semantic Grid Research Program

Semantic Grid Research Program (Phases)	
Phase 1:	<ul style="list-style-type: none"> • diagnostic checking of data obtained. This requires checking all instrument states, the quality of the obtained data, and notification of any possible errors • storage of the validated raw data obtained • establish a mechanism for secure access to the data by users • establish the "data grid" infrastructure first using available technologies such as the Storage Resource Broker

- ensure sufficient metadata is captured at the time of data capture to support the knowledge mining and management services to come later – this will require tools that minimise effort involved, eg. :
 - automatic capture of machine settings, session details (participants, date/time, etc)
 - pull down menus using pre-set controlled vocabularies - these can be pre-configured for each collaborative group and use basic (but extensible)
 - ontologies for describing/structuring the data
 - experimental design (parameters) driven by statistical analysis of data but with the ability for scientists to override/intervene
 - mechanisms for defining persistent unique identifiers for instruments, people, samples, experiments, etc
 - one central metadata repository but with rules for where to store the actual data files based on size, type, user preferences, etc
 - capture of associated remote videoconferencing sessions (many prefer simple Polycom to access grid nodes)

Phase 2:

- provision of supportive tools and services - both generic and domain-specific. These services should be made available through general or domain-specific registries which can be searched and browsed by people or agents - so the optimum service can be discovered and invoked. Communities should be able to publish/advertise their own domain-specific data processing services on registries
- methods for publishing raw data sets to enable sharing/re-use - users may want to provide different levels of access to different types of users
- services for analysing, processing, transforming, migrating, and preserving data
- methods for linking data and other information into 'scientific model' containers
- annotation services for raw data sets, derived information, scientific models'
- methods for tracking life cycle/audit trails of data that undergoes processing, interpretation, derivation
- methods for structuring, compressing and extracting salient information from videoconferencing recordings

Phase 3: (Knowledge/Semantic Grid)

This involves more blue-sky research which could be investigated in parallel with Phase 1.

- participate in the development and use of the emerging international standard for incorporating semantic grid technology to WSRF to define how OWL/OWL-S will interface to WSRF
- ontologies to describe all Grid components (instruments, people, computers, software, networks, storage facilities) and top level ontologies to harmonise these semantic descriptions
- rules to enable the optimum assembly of grid components for a given collaboratory
- agents capable of dynamic discovery, invocation, composition, choreography of most appropriate grid resources and services
- mechanisms for determining when new data is available which impacts on their existing "scientific models" and notifying relevant person or agent

- Mechanisms for linking facility data to publications or on-line learning objects

Phase 4: (Data Curation and Preservation)

This involves the archiving and preservation of data for long term archival of the knowledge developed.

- data archive using an international standard, eg, OAIS. This will involve the implementation of an archival architecture which allows data to be ingested, and facilitate the capture of their intermediate results
- data curation – not all data obtained will lead to enhanced scientific knowledge. Data curation is to examine the knowledge generated by the obtained data, and to archive data which leads to enhanced scientific knowledge, or improvement in our understanding of the basic science facilitated by the synchrotron
- data preservation – this is to facilitate the preservation of data for long term storage. This will include its own metadata, and its own description of reasons why the data should be preserved for long term storage and access

It would be preferable to conduct the research program in a collaborative manner through a ‘virtual’ research centre with an agreed core strategy and program of activities. A Centre Director and a research leader with a staff of four research assistants will be required to commence the program. A three-year program would cost approximately \$3M if associated with a host university. There are Australian researchers capable of carrying out this development, and resources could be made available through existing Australian Government research programs, such as the ARC.

8. Development of Skills Sets for e-Research

The evolution of e-Research as a key facet of modern, globally competitive research has revealed a need to address key skill shortages in the provision of the enabling technologies. This has been recently identified by the UK eScience Programme and by the National Science Foundation:

‘To harness the full power of cyberinfrastructure and the promise it portends for discovery, learning and innovation requires focussed investments in the preparation of a science and engineering workforce with the knowledge and skills needed to create, advance and exploit cyberinfrastructure ‘

NSF CI-TEAM solicitation NSF 05-560, May 2005 – see <http://www.nsf.gov/funding>

A critical element in the development of a national data architecture for Australian researchers will therefore be a comprehensive programme of skills development of information and communications technology professionals and researchers at undergraduate and postgraduate level. The e-Research Centres should have as a key component of their activity the development of these capabilities at the postgraduate and post-doctoral level, and specific targeted funding of professional and research training is required.