Sustainable paths for Data-intensive Research Communities at the University of Melbourne:

A Report for the Australian Partnership for Sustainable Repositories.

Anna Shadbolt, University of Melbourne
Dirk van der Knijff, University of Melbourne
Eve Young, University of Melbourne
Lyle Winton, University of Melbourne

http://www.apsr.edu.au

August 21 2006

Executive Summary

The Australian Partnership for Sustainable Repositories (APSR[1]) is a DEST funded project that brings together a group of committed stakeholders in the field of higher education digital sustainability and preservation. In 2005 APSR proposed to offer a "sustainability consultancy" service to selected research communities whose work is data-intensive and who have an identified need to store and share that data. The Melbourne-based APSR project forms a subset of a national (AERES[2]) project which aims to make APSR's resources relevant to the day to day work of these data-intensive research communities with core outcomes of that project being the development of a base framework for analysing a research community's systems and procedures for data management and archiving. Findings from the Melbourne project will feed into the national AERES project report.

The purpose of this report is to present the local project findings with a view to identifying how these findings may add to the knowledge base for informing an e-research strategy for the University of Melbourne. It also provides important considerations for how major Government initiatives in research policy and funding might impact on research data and records management requirements.

Eleven research communities from diverse disciplines were consulted for an audit of their data management practices. Researchers from these communities represent a number of diverse disciplines: Applied Economics; Astrophysics; Computer Science and Software Engineering; Education; Ethnography; Experimental Particle Physics; Humanities informatics; Hydrology and Environmental Engineering; Linguistics; Medical informatics; Neuroscience and the Performing Arts.

In addition to the specific findings for each group audited, the project findings also provide information about sustainability issues around research data management practices at the university.

Meeting the needs of the researchers interviewed will take resources, and at present much is left to academic departments; often leading to either no or limited action or to 'reinventing the wheel'. These findings point to a number of issues that can help to inform an e-research strategy for the university. Eight recommendations have been formulated for consideration by key stakeholders.

ISSUE 1: The importance of an institution-wide strategy for eResearch.
The findings from this project reinforce the work of Professor Geoff Taylor, Ms Linda O'Brien and the eResearch Advisory Group, identifying the need for an institution-wide strategy to progress and manage eResearch engagement and support. In particular, the findings demonstrate that when it comes to digital data management, there is variable capability among our research communities to comply with the University's Policy on the Management of Research Data and Records[3] and the (consultation draft) Australian Code for the Responsible Conduct of Research.[4] Data management, including access, discovery and storage, must be a fundamental component of such an institution-wide strategy. A broad eResearch strategy can also position the University to meet the challenges of the Research Quality and Research Accessibility Frameworks[5].

Recommendations three to eight below provide some of the essentials for such strategic planning. The Research and Research Training Committee (R&RT) would provide the governance for enabling its implementation.

> ***<u>Recommendation 1</u>: That the University develops a strategy that broadly addresses the policy, infrastructure, support and training needs of eResearch.***

[1] http://www.apsr.edu.au
[2] Sustainable Paths for Data-intensive Research Communities, APSR AERES Project Proposal 2006 http://www.apsr.edu.au/currentprojects/sustainable_paths.htm
[3] http://www.unimelb.edu.au/records/research.html
[4] http://www.research.unimelb.edu.au/hot/current.html#draft
[5] http://www.research.unimelb.edu.au/hot/current.html#quality

<u>Recommendation 2</u>: That the University's R&RT Committee consider forming a subcommittee to provide governance for enabling eResearch at the university. This committee should have broad representation and include Information Services and eResearch leaders.

ISSUE 2: A lack of information policy and guidelines.

There is a lack of best practice guidelines and policy statements available to support researchers with their data management decision making processes. Areas of need include: implementation of research record keeping principles and requirements; data management for short term sustainability and long term preservation; metadata standards, principles and systems, standards for authentication and authorization, and systems for access and storage of scholarly IP.

<u>Recommendation 3</u>: That Information Services initiate a consultative process for the development of appropriate guidelines and, where relevant, policy statements, to support researchers with the management of their research data and records.

ISSUE 3: Absence of a coordinated data management infrastructure for research.

The findings suggest a need for centrally supported flexible data management, authentication and access systems. Groups audited were found to be managing their own data and developing their own access and presentation systems. The need was also identified by several groups for managed data storage facilities. Groups are supporting a variety of software. Needs around authentication and access differed; requiring a variety of public, local, national and international collaborator access. The need for a data management capability that is internationally interoperable; allowing for local storage and collections to federate internationally was highlighted. There will also be a need to promote among the University research community, our capacity for digital data management. This emphasis on developing and marketing 'platforms for collaboration' through ICT within and across institutions is a key aspect of the National Collaborative Research Infrastructure Strategy (see capability area 16).[6]

<u>Recommendation 4</u>: To review ICT infrastructure for research, paying urgent attention to data management infrastructure.

ISSUE 4: Capabilities needed by e-Researchers.

The audit identified expertise used in the conduct of eResearch across a variety of disciplines. The findings show that an eResearch consultation service needs to include at a minimum, information and access to expertise in: Database management; Middleware development, management and support (Data management systems, Grid and other distributed systems, Authentication and Authorisation management); XML advice and expertise; Metadata advice: metadata systems, schema and taxonomy development; Curation and Preservation advice and support for raw data and scholarly output (Business case development advice and support, Discipline based advice and support around sustainable data format selection, and Obsolescence planning).

<u>Recommendation 5</u>: To establish a structured consultation process for eResearch support

[6] See http://www.dest.gov.au/sectors/research_sector/policies_issues_reviews/key_issues/ncris/platforms_for_collaboration.htm

ISSUE 5: Difficulty accessing information about eResearch activity and capability.

This project has identified problems around access to information about eResearch activity, capability and support; with much information exchange occurring fortuitously. It is recommended that an information exchange strategy be established to increase the dissemination of information about support for eResearchers. A springboard to this process could be the delivery of an eResearch Expo in December 2006 to showcase university-wide activity and resources for eResearch.

Recommendation 6: ***To establish an Information Exchange Strategy around eResearch.***

Part of the information exchange strategy is a registry of research capability across the university which would facilitate the dissemination of this information. The feasibility of linking such a registry to the Themis Research Management System should also be established; minimizing the need for duplication of data entry by our researchers.

Recommendation 7: ***To establish a Registry of e-research expertise.***

ISSUE 6: Implications for education and training.

Project findings reinforce the view expressed by the Australian Government's e-Research Coordinating Committee that "The ultimate success of the implementation of a strategic e-Research framework will be dependent on people with attitudes, skills and an understanding of the benefits that the framework can deliver":

> Three groups of skills development are needed to hasten the adoption of e-Research methodologies. Firstly, researchers need easy and structured ways of acquiring basic e-Research skills. Secondly, researchers need a researcher/skilled IT interface, to provide them with day-to-day support. Thirdly, researchers need high level ICT and information management professional support.[7]

We need to look at how we can assist University researchers (staff and students) to acquire and develop skills in e-Research to facilitate their research and to ensure compliance with data management requirements (incl. University Policy on the Management of Research Data and Records). This would include an understanding of research data policies, responsibilities, collections, curation, preservation, copyright/IP, metadata and standards. We need then to ensure they know to access skilled support and high-level infrastructure.

Recommendation 8: ***To review the implications of project findings for researcher education and training.***

[7] An e-Research Strategic Framework: Interim Report of the e-Research Coordinating Committee, 30 September 2005 – see http://www.dest.gov.au/sectors/research_sector/policies_issues_reviews/key_issues/e_research_consult/

CONTENTS

1. Introduction

The Australian Partnership for Sustainable Repositories (APSR) is a DEST[8] funded project that brings together a group of committed stakeholders in the field of higher education digital sustainability and preservation[9.] APSR's scope covers aspects of data management, including the establishment, usability, curation, governance, and sustainability of repository and preservation environments used in research and higher education generally. Through its university partners, APSR is interested in tracking the data management needs of data-intensive e-research and the role that repositories may play within the infrastructure support for these research communities.

This project grew out of APSR's proposal to offer a "sustainability consultancy" service to selected research communities whose work is data-intensive and who have an identified need to store and share that data. The national AERES[10] project aimed to make APSR's resources relevant to the day to day work of these data-intensive research communities with core outcomes of that project being the development of a base framework for analysing a research community's systems and procedures for data management and archiving.

The University of Melbourne project provided a coordination base for the research communities based at University of Melbourne[11]. Its goal is to ensure that APSR's resources are relevant to the day to day work of these data-intensive communities by working in a manner that is "embedded" within the communities to document their research data infrastructure; incorporating their systems and procedures for data management, dissemination, and archiving.

The purpose of this report is to present the local project findings with a view to identifying how these findings may add to the knowledge base for informing an e-research strategy for the University of Melbourne. Results from this project will also be reported to the National APSR AERES project.

1.1 Objectives

The objectives of this project were:

- to conduct an audit of ten to twelve data-intensive research communities across a variety of disciplines;
- to document the data management issues for each community as identified during the audit;
- to develop a framework for progress for each community;
- to work closely with three to six communities to implement change, and
- to draw broader conclusions from these sample communities

1.2 Steering Committee

The project was guided by a steering committee comprised of a number of key stakeholders:

- Ms Nicki McLaurin Smith, (Chair) Director, Information Management and Project Sponsor.
- Ms Anna Shadbolt, (Secretary) Project Manager.
- Dr Angela Bridgland, Deputy Principal, Information (Services).
- Mr. Andrew Yeoh, Director, IT User Services.
- Ms Sally-Anne Leigh, Director, Information and Education Services.
- Ms Martine Booth, Manager, Information Planning and Architecture, Information Management.
- Mr. Devendra Nambiar, Manager, Infrastructure Planning and Architecture, Information Infrastructure.
- Dr Glenn Swafford, Vice Principal (Research).

[8] Department of Education, Science and Technology, Federal Government.
[9] More information about the project is available at: www.apsr.edu.au
[10] Sustainable Paths for Data-intensive Research Communities, APSR AERES Project Proposal 2006 http://www.apsr.edu.au/currentprojects/sustainable_paths.htm
[11] Please note that several of the proposed research communities are multi-institutional, however all include the University of Melbourne as one of the core collaborators.

- Mr. Gavan McCarthy, Member, e-Research Advisory Group and member of participant research community.

2. Methodology

2.1 Participating projects

Thirteen research communities from the University of Melbourne were approached between January and July 2006. Eleven projects or groups were selected based on diversity, availability and project resources.

- Experimental Particle Physics, School of Physics, Faculty of Science.
- Australian Science and Technology Heritage Centre, Faculty of Arts.
- Learners' Perspective Study, International Centre for Classroom Research (ICCR), Faculty of Education.
- Neuroscience MRI Computational Facility, Howard Florey Institute.
- Australian Virtual Observatory, Astrophysics Group, School of Physics, Faculty of Science.
- Hydrological Measurement and Monitoring, Faculty of Engineering.
- Molecular Medicine Informatics Model (MMIM). Bioinformatics, Bio 21 Institute.
- PARADISEC and EthnoER, Department of Linguistics and Applied Linguistics, and the Language Technology Group, Department of Computer Science and Software Engineering.
- Household Income and Labour Dynamics in Australia (HILDA), Melbourne Institute of Applied Economic and Social Research, Faculty of Economics and Commerce.
- Australian Sound and Design Project. Centre for Ideas, The Australia Centre, Faculty of Arts.
- The Kidneyome project, Faculty of Medicine, Dentistry and Health Sciences and Department of Information Systems, Faculty of Science.

2.2 Procedure

The project was intended to have two distinct parts: phase one which entailed an audit of the data management issues of the research groups; and phase two, which would identify gaps around data management practices for a selection of discipline-based and multi-disciplinary research communities participating in the initial audit process. It was anticipated that the outcome of this gap analysis by the local APSR team may lead to changes in practice. The reality found that these two phases were not so discreet, but rather merged as part of an ongoing consultative relationship between the local APSR team and the research groups. Ongoing contact with a number of the groups audited enabled a more detailed description of their data issues and their responses to the input and recommendations from the local APSR project team. Consequently, the methodology was a reflection of this active engagement where participation in the audit had some influence on practice and in some cases resulted in change. The local APSR team documented much of these processes.

The information gathered during the audit and subsequent meetings with research project teams is intended to feed into the national AERES[12] project.

2.2.1 The audit process

The audit process was loosely modelled on that used by Bradley and Henty (2005)[13]. Some amendments were made. The Data Process Classifiers[14] was developed by the local APSR and National AERES project teams to assist in the data process mapping.

[12] More information about the AERES project is available on the APSR website at: http://www.apsr.edu.au/currentprojects/sustainable_paths.htm

[13] Refer Appendix 5.1 for a copy of the survey. Report by Bradley, K. and Henty, M. (2005) Survey of data collections: a research project undertaken for the Australian Partnership for Sustainable Repositories, available at: http://www.apsr.edu.au/publications/data_collections.htm

[14] The ***Data Process Classifiers*** tool is located in Appendix 5.2.

2.2.2 Phase two consultations

A number of criteria were developed for selecting projects interested in further consultation with the project:

- Ensuring diversity of discipline areas.
- Ability to resource the intervention/support.
- Willingness of the research group to participate in the process.
- Reusability of intervention/resource within other research communities.
- Identifiable public knowledge component (i.e. research knowledge that has external value and can be shared).

The process of identifying communities moving into further consultation occurred both as an intentional project decision to work with a group and as a natural progression from the team engagement with the researchers during the audit process. Both approaches were considered valid and represented a service delivery strategy for Information Services. Notably the timeframes of this current project did not match those required for real time consultation around the complex technology and sustainability needs of our eResearch communities. Consequently these Phase Two consultations are ongoing and this report identifies the activities commenced and not necessarily their outcomes.

Four projects participated: Learners' Perspective Study (ICCR-Education); the Molecular Medical Informatics Model (MMIM) Bio 21; the Hydrology Research Group (Environmental Engineering), and the Australian Sound Design Project (Centre for Ideas – The Australia Centre).

3. Project Findings

The findings are presented in three parts; the first (3.1) provides specific information about each project, the second (3.2) highlights the specialist eResearch capabilities identified during the audit, and the third (3.3) focuses on the infrastructure implications of the consultations during the project.

3.1 Overview of participating projects

This section provides a brief overview of each project audited. It includes the names of researchers consulted, project partners, funding sources, general project information and details of their data management processes. Links to more detailed information are also provided if deemed appropriate.

3.1.1 Experimental Particle Physics[15]

The focus of this audit was the Large Hadron Collider (LHC) project, the ATLAS experiment[16] located with the Experimental Particle Physics Research Group in the School of Physics. This experiment is still at the preparation/build up stage and is expected to commence in 2007 and continue until 2017. It is based in Geneva; the location of a very large instrument (LHC Accelerator) that will, once the experiment begins, produce tens of petabytes of data per year for distribution around the world across multiple collaborator sites. The project at Melbourne is part of Australia's participation as a Tier 2 facility in this experiment. The Melbourne ATLAS project has also contributed by participation in the detector technology development and constructing on a number of components and has participated heavily in the operation of the ATLAS test beam (scaled down detector test) and in a number of aspects of software development.

[15] http://epp.ph.unimelb.edu.au/EPP/AtlasActivities
[16] More information about this global experiment is available at: http://atlasexperiment.org/

The Melbourne Project Leader is Professor Geoff Taylor. This team includes 19 researchers and four were consulted during the audit:

- Associate Professor Martin Sevior
- Dr Glenn Moloney
- Dr Lyle Winton
- Dr Marco La Rosa

Project Partners/Collaborators in Australian Tier 2 data processing facility[17]

- Falkiner High Energy Physics group, University of Sydney.
- The ATLAS collaboration (includes 4682 participants from 100s of institutions)[18].
- CERN (European Organisation for Nuclear Research[19]) and the LCG (LHC Computing Grid[20])
- APAC and VPAC.

Funding sources are numerous and include funding agencies and institutions around the world associated with each member group. At the local level funding has been sourced over the years from:

- ARC grants.
- Departmental/University funds.
- External grants, including DEST, APAC and VPAC.

Data Management Processes

This project represents a highly specialized technology and data management skills set.

Data Acquisition: Experimental data are archived at the Tier 0 in CERN, distributed to a global federation of ten Tier 1 sites, and then partially distributed or accessed from Tier 2/3 sites within the ATLAS collaboration. Simulation data are generated throughout the collaboration and collected and archived by Tier 1 sites. This transfer of data requires advanced file transferring services, and high speed, well tuned network connections at every site. A key problem in this transfer which is unique to Australia is bandwidth and overcoming the big international latencies that are involved in the data transfer.

IP/Copyright of Data and scholarly output: Prime responsibility for the data remains with the global collaboration. Post-processing, including analysis of the data at Tier 3, that is, the places where Melbourne physicists will be doing their data analysis as individuals, remains with the individual and the local team. This is essentially where researchers try and compete with the rest of the collaboration to be the first to make these discoveries that come about. Where individual researchers generate output from shared data it is considered individual IP, but publications coming from this data must include the collaboration on the author list. So the researcher can never truly claim this as solely his/her own work.

Data Quantities: This information is provided to illustrate the current[21] and projected storage levels required by the local team.

Current usage at the local level fluctuates depending on where Melbourne ATLAS researchers are located. For instance, in July 2006 there is little data for the ATLAS experiment with one student running locally: has downloaded 60 GB from the collaboration (simulation data set) and is processing this data down to about 30 GB. This researcher is conducting some private simulation of data, currently of the order of a GB or two, but this could become 10 GB. Two other main software users (a post-doc and another student) are at CERN so are currently using disk space offsite. This current usage does not reflect the storage needs when all our ATLAS researchers (19) are on site in Melbourne.

---

[17] Probably hosted at an APAC type facility

[18] This global collaboration comprises 4682 participants from 100s of institutions; 3 being in Australia with 24 participants (http://graybook.cern.ch/programmes/experiments/lhc/ATLAS.html)

[19] CERN is the European Organization for Nuclear Research; the world's largest particle physics centre. It sits astride the Franco-Swiss border near Geneva. More information about CERN is available at: http://public.web.cern.ch/Public/Welcome.html

[20] The LHC Computing Project (LCG) is building and maintaining a data storage and analysis infrastructure for the entire high energy physics community that will use the LHC: http://lcg.web.cern.ch/

[21] This is based on information provided in July 2006.

The projected data quantities are based on meeting dual obligations as part of the collaboration at both Tiers 2 and 3. Once the experiment has commenced[22] the data quantities will continue to increase over the ten to fifteen year life of the experiment. For individual researchers at the Tier 3 level, there is a data storage allocation of 1 terabyte per user per year. This will start off at 40 terabytes for 2007; increasing to 300 terabytes by 2012. The Tier 2 facility is aiming at starting in 2007 with 100 terabytes going up to 1.5 petabytes by 2012; and the Melbourne group must contribute to this requirement.

Data storage and Backup: Most of the existing collaboration data are stored at CERN; the host institution for the ATLAS experiment. But much is also stored across institutions throughout the ATLAS collaboration and is available via RLS, LFC, and DQ2; all somewhat integrated Grid enabled replica catalogues. This agreement will continue when the experiment commences in 2007. Backup is through local facility resources and replication. Local TIER 3 data will be backed up using departmental and institutional servers and remains the responsibility of the individual researcher.

Data formats: Open standard or locally produced (de facto community standard) formats are utilized for all stages of the data life cycle. This is Root/IO (C++ based) and includes tables of information with a data/field dictionary. Tables conform to the ATLAS EDM (event data model). The data formats are summarised or processed to varying levels with different EDM versions. The output remains in the Root/IO format. These processes are coded and these various code versions are kept indefinitely.

The Root format is used to access data; providing a fast and flexible framework for the researcher to construct any analysis with access to all available/converted/derived data. The Gaudi/Athena tools contain the data dictionary or meaning of the data. Athena is the preferred analysis tool. Different versions of the software are stored using the CMT computer system based on Concurrent Versions System[23]. The data are available in different versions which correspond to the software versions of Athena; the analysis tool.

Metadata: The collaboration uses the Pool Catalog (XML) standard but unclear about a fixed schema as such. The metadata is used to describe the technical and structural attributes of the data. It is recorded around the files. The primary creator (the global collaboration) has responsibility for generating this metadata.

Data Access, Authentication, Authorisation and Security: Membership of the ATLAS collaboration entitles researchers to data access. Each institution or group has a head/contact. Membership approval must come from a group head. Membership size of a group determines the group's annual maintenance and operations contribution cost which can be offset by the quantity of 'in-kind' contribution over the years. Collaboration members contribute towards the cost of detector development, deployment, operation and infrastructure as well as investing years[24] of effort. Without such fees and effort the experiment would not exist and if access to research data were allowed without contribution, this would actively discourage the levels of contribution required to make the experiment possible. Melbourne currently has a membership of 19 researchers.

For data management only traditional authentication (UNIX username password) exists. Some Grid (PKI) authentication is being used. Interpretation of data can be difficult so security is not a major concern. It is possible to get tapes from CERN if you do it yourself. However, when data becomes larger and more distributed this may be difficult. Access to tape making facilities is pass worded.

There may be some risk associated with the protection of work in the Grid context. Where there is data which is not open to others in the collaboration, it would need to be placed on a storage resource (Grid enabled) that is secure. Access policy can be determined using the Grid tools, but this can be difficult with the current ATLAS Grid middleware as it is designed for massive data production. The user analysis part is still being worked out.

[22] The experiment will start an estimated 10 PB data production per year in 2008 (both experimental and simulation) that will be managed at Tiers 0 and 1 levels, with limited production commencing in 2007. This is when both the LHC and ATLAS detectors are operating at full capacity.
[23] CVS – open source version control
[24] The Melbourne team has been involved in the preparation of this experiment for the past 17 years.

3.1.2 Australian Science and Technology Heritage Centre[25] (AUSTEHC)

The focus of this audit was the AUSTEHC itself which is not a research project as such but a Centre for Cultural Informatics and Humanities Computing that partners with research communities and projects.

The Director of the centre, Mr. Gavan McCarthy, was consulted during the audit.

The Centre is based in the Faculty of Arts but has many associations with other Universities, including ANU and Latrobe University, as well as strong associations in the United Kingdom and has been working for the International Atomic Energy Agency[26].

Examples of projects partnering with the Centre include:

- Indigenous Studies Unit and Faculty of Law - 'Agreements, Treaties and Negotiated Settlements'.
- Departments of History and Philosophy of Science, History, and the Australian National University, Australian Dictionary of Biography unit – 'Australian Dictionary of Biography Online[27]'.
- Department of Linguistics – PARADISEC and the archival collections of various Australian linguists.
- Department of History - Australian Women's Archives Project (National Foundation for Australian women).
- Centre for the Study of Health and Society - Diane Barwick Archival collection.
- Australian Venom Research Unit, Department of Pharmacology, Faculty of Medicine Dentistry and Health Sciences – 'Australian Venom Compendium'.
- Department of History - Encyclopaedia of Melbourne – *Encyclopaedia of Melbourne online*.

Funding sources vary across projects but in an archival sense the funding of the original materials (data) could have been from any number of historic sources (this is part of the contextual framework surrounding the data). However, the "archiving" of the data – which is the contemporary activity is generally funded by a body external to the Centre and its project partners, for example the ARC, government and industry.

Financial sustainability is an issue for this project with most staff maintained on short term annual contracts. This places pressure on the centre's ability to guarantee ongoing employment risking loss of expertise that will jeopardize ongoing operations.

Data Management Processes

The Centre provides project management, software, digitization advice, resources and repository/web publication platform for users. A core part of the work of the Centre is to support research and development around the systems and tools that support their own activities and those of the partners.

Data Acquisition: The starting point for most projects of the Centre is an archival collection – so therefore archival records, in all their variety provide the foundation. The source of the data varies depending on the collection or project. These collections are generally donated to an archive but not always (they may be loaned for imaging). In the process of documenting the archival materials new data are created and new digital information objects are created.

IP/Copyright of Data and scholarly output: It is generally considered that the IP of the collections belong to the contributors. As a rule, these collections are made available for free and have no commercial bases. Nevertheless the moral rights of creatorship expect that researchers accessing and using data in scholarly works acknowledge the creators. However, AUSTEHC accept that there are complexities with some of their projects and that this area has not been dealt with systematically to date. Until recently, there were no copyright statements provided to users accessing the collections, nor any statement of moral

[25] Centre website is located at: http://www.austehc.unimelb.edu.au/
[26] Report produced for the International Atomic Energy Agency on sustainable knowledge management can be located at: http://www.austehc.unimelb.edu.au/
[27] Available at: http://www.adb.online.anu.edu.au/adbonline.htm

expectations surrounding the use of data held within them as these roles were deemed to be the responsibility of the managing archive.

Data Quantities: The number of digital objects in the various collections managed by the centre is uncertain but is in the vicinity of 200,000 items. The Centre server has a capacity of one terabyte. The tape archive (digital tape magazine) of the full dataset has a capacity of 2.7 terabytes and is held off-site.

Data storage and Backup: All active data are stored on the centre server (capacity 1TB). The system is only just managing its current demands and it is projected that it will not have the capacity to meet growing demands of centre projects. The server undergoes daily incremental backups with a weekly complete backup on a three week re-use cycle. The most recent backup is stored off-site (at the Director's home) on a digital-tape magazine with a capacity of 2.7 terabytes and is on weekly changeover (3 week cycle) - this backs up the Centre's whole data set. Most of our web publications are backed up by PANDORA[28] at the National Library of Australia. The PANDORA backups seem to be about once each six months. There is some concern about the potential risks surrounding the current offsite backup storage arrangements as the backup model used is geared to a dynamic data environment and it does not seem to be well-suited to backing up a store of predominantly static objects.

Data formats: The data could be provided to the Centre in any form or format and cover a range of digitized and non-digitised materials including:

- Sound/audio with some transcripts;
- Film and video;
- Text – biographies;
- Archival materials, and
- Photographic images.

Digitally imaged materials are generally from archival collections – most materials are digitized by the centre to improve access and are not imaged to replace the original (the original is not at risk of destruction). Materials are documented and managed in systems tools developed locally by the Centre – the OHRM[29] and the HDMS[30].

Both the OHRM and the HDMS are open source relational database systems which use MS Access as the data collection and management platform. In both cases data are exported from the MS Access form into web-technologies suitable for web publications.

Data are made available to users via the web so it is predominantly in html or other standardized file types that are handled by web browsers. For imaged records this is generally jpg files but PDF files may also be used. The outputs from the OHRM and the HDMS currently have three standard forms. These are outlined below.

OHRM:

- Programmed Static html – Basic level and advanced level (xhtml and css) implemented.
- Open source relational database – Postgresql implemented.
- XML – compliant with the Encoded Archival Context (EAC) XML Schema but not implemented; output to MAPS ands MODS schema for the National Library of Australia tested.

HDMS:

- Programmed Static html – Basic level implemented.
- Open source relational database – Postgresql tested but not implemented.
- XML – Encoded Archival Description (EAD) Schema version 1 implemented.

---

[28] PANDORA is the National Library of Australia web archive. More information is available at: http://pandora.nla.gov.au/

[29] The Online Heritage Resource Manager (OHRM) - locally developed open source context based resource discovery and access system. More information available at: http://www.austehc.unimelb.edu.au/ohrm/

[30] The Heritage Documentation Management System (HDMS) is locally developed open source tools for an archivist to process and manage any collection or grouping of records. More information at: http://www.austehc.unimelb.edu.au/HDMS/findingaids.html

Metadata: is originally collected and managed in the relational databases but can be exported in a variety of forms for different purposes. The metadata schemas used by AUSTEHC are primarily based on the standards produced by the International Council on Archives[31]; namely the ISAD (G) and ISAAR (CPF). There are two XML schemas that have been produced and are widely accepted as embodiments of these standards (EAD and EAC – referred to above).

The metadata is used, to varying degrees, to describe: rights and permissions; provenance; technical metadata; administrative/management; bibliographic/descriptive and structural attributes of the data.

Generally, metadata creation falls to the AUSTEHC staff working on a particular project. In practice, the Centre has found that it is very rare for the primary creators of the data they manage take responsibility for metadata creation and as a rule tend to leave out the common knowledge; key aspects that future users really need to know.

Data Access, Authentication, Authorisation and Security: As highlighted earlier, the AUSTEHC collections are public access requiring a standard web browser to use to the collections. Users remain anonymous and little is known about who the users of the collections are.

3.1.3 Learners' Perspective Study, ICCR

INITIAL AUDIT

The focus of this audit was the Learners' Perspective Study[32]; an ongoing and growing project of the International Centre for Classroom Research[33] (ICCR) in the Faculty of Education. The eResearch initiative for the project is the investigation of the viability of using Grid-enabled technology to share and extend existing practices in video analysis on an international scale and to evaluate the relative efficacy of this approach in comparison with web-mediated streamed data sharing and analysis using non-Grid approaches.

Two key members of the project team were consulted:

- Professor David Clarke, Project Leader and Director of ICCR.
- Mr. Cameron Mitchell, Technical Manager for ICCR.

Project Partners/Collaborators in the project are increasing annually. At the time of the initial audit there were 15 countries in the collaboration and by July this had increased to 19. Countries currently include: Australia; China; Czech Republic; Denmark; Germany; Israel; Japan; Korea; Norway; The Philippines; Portugal; Singapore; South Africa; Sweden; Taiwan; the Netherlands, the United Kingdom, and the United States.

Funding sources include:

- ARC grants.
- Departmental/University funds and by team members' universities.
- External grants, including The Spencer Foundation, USA, The South African National Research Foundation; The Japanese-Australian Collaborative Research Fund; The Hong Kong Research

[31] More information about this standard is available at: http://www.icacds.org.uk/eng/home.htm

[32] Commencing in late 1990s, the project documents patterns of participation in well-taught eighth grade mathematics classrooms in a comprehensive and integrated fashion using multimedia technologies that enable the documentation of both the obvious social events that might be recorded on a video as well as the participants' understandings and constructions of those events, including their memories, feelings, and the mathematical and social meanings and practices which arose as a consequence of those events. More information about the project and study design is available at: http://extranet.edfac.unimelb.edu.au/DSME/lps/subabout.shtml

[33] The ***International Centre for Classroom Research*** at the University of Melbourne provides a facility for the storage of the project data, its dissemination to team members, and a site for collaborative data analysis. Funds have been provided by the ARC to support the accommodation of overseas team members during their periods of collaborative work in Melbourne.

Grants Council; The Swedish Council for Research in the Humanities and Social Science, and The Swedish Foundation for International Cooperation in Research.

Data Management Processes

Data Acquisition: The Research Team in each participating country is responsible for contributing and collecting a full set of data[34] for that country. Project partners are at different stages of completion of this process; with some yet to commence. It is estimated that the cost per country for creating their own dataset is currently in the vicinity of AUS$300,000. This investment provides each country to the data of all participating countries in the project.

Current ingestion and distribution of data between project partners is via CD, DVD or mobile pocket (80GB) drives. The goal of the current project is to change this to a distributed international network for ingest, access, storage and analysis.

IP/Copyright of Data and scholarly output: The IP of each country's dataset remains with the data creators, the contributing research team, and it is the responsibility of the project leader in each country to authorise access to the IP associated with the raw data, including approval for analysis proposals for that data. It is assumed that researchers within the collaboration using the data of another country duly acknowledge the authorship and origin of that data in any scholarly output generated from that data.

Data Quantities: It is estimated that each country's raw dataset is 50GB. This will build up to around one terabyte of raw data for the current 18 collaboration partners. This does not include the data associated with any analyses or annotations of that data or any other scholarly outputs.

Data storage and Backup: The ICCR storage capacity is 11TB and adequately meets current project needs. In preparation for projected increase in storage needs an additional 14TB of storage space to be held within the Faculty has been requested for back up of ICCR data.

Once acquired, data are compressed or converted into the current format standards (currently MPEG 4, text and PDF) and stored on the centre server. Offsite back up of this version is made every time a new batch of data arrives and maintained at ICT Building, 111 Barry St. Local project output at Melbourne, namely analyses of data and scholarly outputs are backed up nightly and stored locally and offsite at the ICT Building, 111 Barry St.

The archiving of raw data are essentially the storage of original data as acquired from project partners – as DV tape, DVD, CD and to a lesser degree, analogue VHS and audiotapes; all held onsite in filing cabinets. There is no cataloguing of this data. It is assumed that originating countries also maintain copies of their own raw data. Raw Classroom datasets are re-distribution to other sites in same digital format as acquired. There is no updating of file formats or standards of this data; it remains as acquired.

The project team believes that this collection is one of local, national and international significance which should be preserved for re-use and re-purposing by future researchers. Current raw data storage formats suggest that the collection is not sustainable long term. There is a need to investigate the need for format migration to protect the collection from digital format obsolescence, particularly where commercial formats have been used.

Data formats: A combination of open source and commercial formats are used by the project. The centre provides partners with documented guidelines[35] for preferred formats for data:

- Video data - preferred format is MPEG 4, but also submitted as: .mov, .mp4, MPEG1.
- Text data - preferred format is .txt files or PDF, but also submitted as: .txt; .doc; TIFF and JPEG.
- Image data - preferred format is PDF, but also submitted as: TIFF and JPEG.

[34] The dataset for each country includes all materials associated with the delivery of a collection of thirty classroom lessons (ten lessons for each of three teachers), including videoing of these 30 lessons from three camera perspectives, post lesson interviews, transcripts, translations, general information about the school and other contextual information.

[35] Please contact the Technical Manager for further information about the ***Technical Guidelines for LPS Data Processing***

Video data analysis in Melbourne (and for some of the project partners) is conducted using the commercial software *StudioCode*[36]; developed in conjunction with the Australian software company Sportstec. The project is currently in negotiations with Sportstec for making this software "Grid-enabled" or distributed for access to all partners. This would enhance opportunities for collaborative analyses of project data. Analyses using *StudioCode* are not accessible without the software limiting some types of research collaborations across project sites.

Long term use of this commercial software will require the storage of the software. This process has been necessary for the preservation another analysis tool used by the project - *VPrism;* discontinued software that was previously used by the Melbourne team and remains in use for some data and by some project partners. This software has been archived with the OS9 Mac operating system needed to run it.

Metadata: There is no systematic cataloguing of the data collection currently held within the project. Much of the information exists within a complex files classification system[37] for the raw datasets but is not currently extracted and managed within a relational database and is not easily discoverable. Key project information is held within the heads of key personnel or in emails or within other project papers and correspondence.

The project plans to develop a metadata schema for information around rights and permissions; provenance; technical metadata; administrative/management and bibliographic/descriptive attributes of the data will be extracted and/or developed. It is anticipated that research teams will take responsibility for the development, maintenance and management of this metadata. Advice about metadata schemas for data management and collection preservation will be sought at that time to ensure should this be deemed desirable by the collaboration.

Data Access, Authentication, Authorisation and Security: The current protocol is that data are stored centrally in Melbourne for all project sites. Arrangements for access of data by researchers within the project collaboration are manually managed, requiring the Melbourne facility to act as intermediaries for the data access requests. A more distributed model of data access and storage may streamline some of these processes.

A condition of access to a country's data is that the requesting researcher must gain approval for his/her proposed analysis from the originating country (the owners of the data IP). This acts to safeguard against duplication of the same research outputs from different sites and ensures that first option for particular analyses goes to the data creators.

PHASE TWO CONSULTATION

This project engaged in further consultation to identify and implement change around their data management processes. The focus of this follow up was particularly on middleware selection and testing. This activity has commenced and will continue as part of the services provided by Information Services personnel at Research Computing and Information Management.

Activity included:

The exploration of various data infrastructure (middleware) models to address the distributed data sharing and analysis across the global collaboration:

- Establishment of an SRB[38] testbed and testing suitability for project needs.

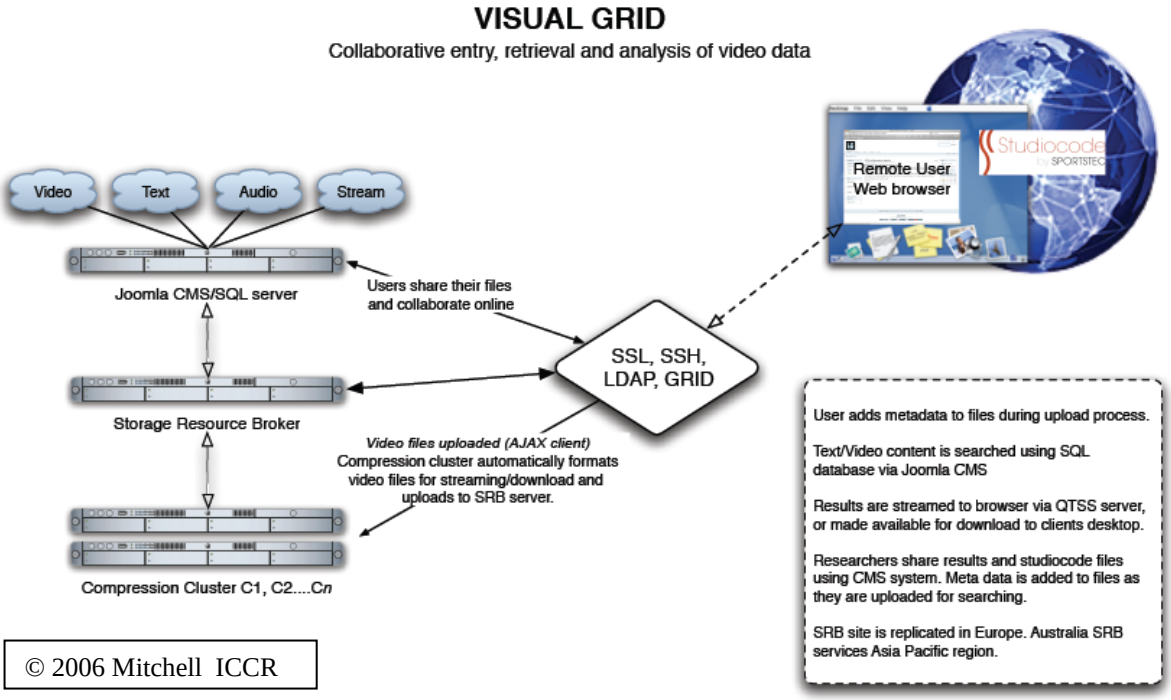
[36] Currently using version 2.0.5 of StudioCode. More information about the software is available at: http://www.studiocodegroup.com/sbg/studiocode-education.html

[37] This is outlined in the ***Technical Guidelines for LPS Data Processing*** for the raw data compilation.

[38] SRB – The SDSC Storage Resource Broker (University of San Diego) – supports shared collections that can be distributed across multiple organizations and heterogeneous storage systems. The SRB can be used as a Data Grid Management System (DGMS) that provides a hierarchical logical namespace to manage the organization of data (usually files). More information is available at: http://www.sdsc.edu/srb/index.php/Main_Page

- Sakai[39] – participation in Melbourne collaboration to explore this platform to meet the needs. Initial investigations suggest that Sakai may provide a possible solution for the second phase of their project: improving the interface and providing collaborative tools.

The diagram below illustrates the proposed amended architecture for the collaborative infrastructure, incorporating the SRB layer. This is still evolving.



Discussions about metadata schema development to meet increasing need for archiving of local project and collaboration assets. Current practice is not sustainable. Consultation with the collaboration is needed to consider the development of appropriate schema for the group.

Discussions about data archiving, preservation metadata and consideration for the long term preservation implications of current practice have commenced. This is not a current priority of the project but will need to be raised with the collaboration.

3.1.4 Neuroscience MRI Computational Facility

The focus of this audit was the computational facility and repository for neuroimaging data (established March 2006) and the sharing of computing and information resources among neuroscience researchers.

Two key members of the project team were consulted:
- Associate Professor Gary Egan, Project Leader
- Dr Neil Killeen, Facility Manager

Project Partners/Collaborators include:
- Howard Florey Institute.
- University of Melbourne (Electrical & Electronic Engineering, Computer Science & Software Engineering).
- National ICT Australia (NICTA).
- University of Queensland (Biomedical Engineering).
- Flinders University (Psychology/Cognitive Neuroscience).

[39] More information available at: http://www.sakaiproject.org/

- Silicon Graphics (SGI).
- UNSW (Prince of Wales Medical Research Institute).

Data Management Processes

The core objective of this facility is to generate a distributed neuroscience facility. The current model for the facility consists of a primary node and two secondary nodes: University of Melbourne (primary node); University of Queensland (secondary node), and Flinders University (secondary node). This structure may change over time; nodes may expand and more nodes may also be added. Currently, it is expected that each node will provide both computational and some data storage capabilities, however, the primary node will be substantially more capable. The current phase of the facility set up is focusing on the functionality of the primary node at Melbourne (the team with whom we have consulted) and the establishment of the repository infrastructure.

Data Acquisition: Data are current research and legacy MRI data acquired from a number of scanners.

IP/Copyright of Data and scholarly output: Human MRI scans are de-identified subject (research participant) data. Generally, the IP of data belongs to the contributing researcher, though facility policy around this is yet to be documented[40]. Data reuse policy is also under discussion and consequently, only non-human data are re-purposed within current guidelines.

Data Quantities:
Currently: 3 TB of existing data are being transferred into the facility.
Projected: An upper limit of approximately 30 TB within 2 years is possible based upon expected data acquisition and potential collaborations. This includes one time legacy transfers. The current expected rate of data growth is <10TB/annum.

Data storage and Backup:
Current storage capacity is 62Tb (actually 33Tb as data are stored redundantly on tape).
Data back up and storage processes are:

- Data are migrated transparently between disk and tape via the Data Migration Facility (DMF) software. From the user's perspective, data are always 'online'.
- Data are migrated redundantly (2 copies) to tape by DMF, but not all files (e.g. small files) migrate to tape.
- Data that are mistakenly deleted can be retrieved for up to 7 days.
- There is currently no other copy of the data managed by the Facility. Data originating at hospitals will usually have a copy there (but not easily accessible to users).
- The Facility plans to make offsite copies of high-value assets from the data storage facility.

The facility needs to establish a reliable off-site store for back up storage and has begun exploring the possibilities. The University Data Centre is a resource that might meet this requirement. However, the new University Queensberry St Facility will not be available until second quarter 2007; its capacity may be insufficient and the business model is not yet clear.

Data formats: Data formats and software are both open standard and commercial. Images formats are DICOM[41], NIFTI[42] and ANALYZE are de facto standards.

The Archive is built on top of a commercial Digital Asset Management system: *MediaFlux*[43]. Data are pushed to and pulled from the Archive via *MediaFlux*. A DICOM server has been built into *MediaFlux* providing tight integration between remote MRI scanners and the Archive. Thereafter data are retrieved

[40] A number of these issues have been raised in facility issues paper: *Asset Privacy Requirements and Implications*, Killeen (May 2006). Policies are yet to be formalised.
[41] DICOM is the MRI clinical standard. At times it does not meet the needs of the researcher.
[42] NIFTI is a new data format standard for analyzing Neuroscience imaging data. Processed data uploaded into the Archive will be in NIFTI format.
[43] The facility manager has worked closely with the software provider who has custom built the system for the MRI facility. More information about this software is available at: http://www.arcitecta.com/mediaflux/technology.html

from the Archive, with optional format conversion (e.g. DICOM-NIFTI) and various Neuroscience analysis software packages are then used to analyze the data. The metadata may be retrieved from *MediaFlux* into interoperable XML files and the data format itself is not changed by MediaFlux.

The facility is developing its own custom interfaces to *MediaFlux* that will manage the user's workflow (ingest. Egest, processing pipelines, etc).

Metadata: Currently, the project is only holding technical metadata harvested from the DICOM metadata. It is anticipated that this will broaden in the future. The metadata schema by the facility is loosely based on the BIRN[44] schema and can be retrieved into other schema. The metadata are stored in a PostGres data base managed by MediaFlux. In the future, MediaFlux will have a self-contained database.

Associate Professor Gary Egan states that the Australian neuroscience research community is currently working on developing an ontology for its data, as are other such communities in the US and Europe. This process foresight into how future researchers will want to describe their data which needs a strong conceptual framework; it needs to describe what the data are and how they might relate to other data that are potentially of interest in multi-discipline analyses. Participation in global collaborations like the INCF[45] enables such work.

Data Access, Authentication, Authorisation and Security:
Currently data are only available to users within the project collaboration. Access to the collection is via password authentication and does not have an open IP address.

Data from the Royal Children's Hospital scanner are transferred to the Facility through a public IP with a controlled Access list. In the future the facility will move this to a VPN for greater security. Access between Howard Florey Institute and the Facility is currently via a VPN.

3.1.5 Astrophysics - Australian Virtual Observatory

The focus of this audit was the Australian Virtual Observatory (Aus-VO)[46] and the Australian Astronomy Grid (Aus-VO APAC Grid) which is being developed to handle the data storage and access needs for the research community. This project is building a distributed high bandwidth network of data servers[47]. The project is located within the Astrophysics Research Group[48] in the School of Physics.

Three researchers were consulted during the audit:

- Professor Rachel Webster, Lead Investigator
- Dr Katherine Manson, Grid Research Programmer
- Dr Randall Wayth

Project Partners/Collaborators

- The Aus-VO partners with the IVOA[49] (International Virtual Observatory Alliance)
- University of Melbourne
- Monash University
- Swinburne University of Technology
- University of Sydney

[44] Biomedical Informatics Research Network US based network developing standards and tools for neuroinformatics. More information available at: http://www.nbirn.net/index.htm
[45] The Global Science Forum (GSF) of the OECD initiated the International Neuroinformatics Coordinating Facility (INCF) to further the development of Neuroinformatics as a global effort with the support of all ministers of research within OECD. More information available at: http://incf.org/
[46] http://aus-vo.org/
[47] Refer to Webster slide 8 URL: http://astro.ph.unimelb.edu.au/~rwebster/MU_mar05.ppt
[48] http://astro.ph.unimelb.edu.au
[49] http://www.ivoa.net/

- University of New South Wales
- University of Queensland
- Australian National University and Mount Stromlo Observatory
- Australia Telescope National Facility
- Anglo-Australian Observatory
- Australian and Victorian Partnerships for Advanced Computing (APAC & VPAC)

Funding sources include:

- ARC grants
- Departmental/University funds
- External grants, including NSF, APAC

Data Management Processes

This research community requires a highly specialized technical skill set. The collaboration involves many astronomical surveys to be conducted between 2003 and 2008. Each survey involves dataset sizes ranging from 10-100 terabytes and 10 to 100 researchers. "It is no longer feasible to function effectively in Astrophysics as a solo researcher." (Webster)

Data Acquisition: The primary creator of data is the observatory. In the majority of cases observatory data are virtually inaccessible to the outside world. To get data you actually have to go to the telescope with your tape and download it (Webster). Each observatory has its own system for archiving data with some not archiving their data at all. The Aus-VO collaboration is currently working on establishing a distributive model to increase data accessibility among partners. Once data are acquired by the research community, the primary responsibility for maintaining the data lies with the researcher or project group.

Aus-VO data, sourced from observatories, is pre-processed prior to ingest. This includes quality screening (E.g. *"Do we want to keep this?"*). There is also a reduction process that occurs on the raw data. Some of this may be done by the instrument but most is done by the researchers. Researcher confidence in automated reduction processes are generally low; with most preferring manually process/calibrate the raw data (or their trusted PhD/Post docs).

IP/Copyright of Data and scholarly output: this currently sits with the Observatory and the Chief Investigator (CI). Observation time is booked with the observatory by the research project team. Data collected over these times (which may run for a number of days/weeks) remains the IP of that project (CI) for 12-18 months, depending on the observatory policy. After that time the data must be made publicly available. In all cases the observatory would be acknowledged as the data source in scholarly output.

Data Quantities: The current store of around 40TB of Aus-VO data is quite scattered. All Melbourne data are stored and managed by APAC and held at the ANU Peta-Store. Swinburne has about 12 TB of data on tape that is around 10 years old and sustainability of this collection is unclear.

Projected data store requirements are set to increase dramatically from 2010 when the new Australian telescope which will be generating 1-2.5 TB of data per day. The Melbourne team has bid to manage the data from this telescope.

Data storage and Backup: As mentioned above, Melbourne's data archiving is currently maintained by APAC. Back up of this data is managed as per APAC protocols with the Peta-Store which is a fully managed storage facility. Individual researchers at Melbourne also maintain their own personal research data and scholarly output on their own PCs which are backed up by Faculty servers.

Long term storage for this research community is still up for discussion. IVOA established an interest group on Data Curation and Preservation[50] which is working towards identifying both mechanisms for the long-term preservation of astrophysics collections and sustainability procedures to ensure continued

[50] More information about the IVOA Data Curation and Preservation interest group available at: http://www.ivoa.net/twiki/bin/view/IVOA/IvoaCP

access to astrophysics collections that are at risk. Making decisions about what should and should not be preserved remains debatable and parameters will need to be set by the community.

Professor Webster sees that the collaboration in Australia which is part of the global community would be well served by a federated national facility. Sites within the federation could easily be an institutionally based well managed data centre or at some other trusted site.

Data formats: All data formats and software used are open standard or locally developed. In some cases, telescopes may develop some of their own unique file formats. The IVOA facilitates collaboration on standards to ensure the interoperability[51] of the different Virtual Observatory projects.

FITS[52] and its various extensions is the open standard (NASA) for this research community. This standard has remained very stable and is used for archiving and transporting astronomical data. It is also backward compatible for around 20 years. Raw and processed data are accessed and stored using FITS.

Data occurs in different versions:

Observational data that comes from the observatories

- **Raw data** – as it comes out of the telescope in raw state
- **Processed data** – raw data that has had some cleaning by the instruments
- **Science ready data** – Observational data that is processed/calibrated by researchers. In some cases this can be an automated process. In general however, astronomers prefer to implement these calibrations on their own observations (or use graduate students here) rather than trusting any automated process.

Simulations – data generated by researchers using some sort of modeling software/process

Data Analyses of any or all of the above versions are generated by researchers using various software/code for analysis, e.g. Myriad, Iraf.

Metadata: The amount of metadata is variable. FITS captures some metadata. There is no current schema standard being used though this may change with IVOA data curation initiatives. For ground based telescopes astronomer log books document important data like the ambient temperature, the moisture, whether it's raining. This would need to be entered manually during data archiving but tends not to happen.

Data Access, Authentication, Authorisation and Security:
The data has restricted access for 12- 18 months (depending on the observatory) and then becomes openly available. Access is managed by the observatory and varies accordingly. The APAC Grid project is looking at a distributive model for transferring and accessing data. Security issues associated with data transfer and still being developed.

3.1.6 Hydrological Measurement and Monitoring

INITIAL AUDIT

The audit was focused on the Hydrological Monitoring Network projects[53] of the Hydrology Research Group, Department of Civil and Environmental Engineering. This data collection has national and international significance. Soil moisture data has been collected from parts of regional Victoria and New South Wales since 2001. It is currently the only Australian collection of soil moisture data that is presented as an accessible website containing background information about monitoring sites, instruments and drilling right through to the actual data. The goal, subject to funding, is to preserve the collection and continue to collect data long term.

[51] More information about interoperability of standards: http://www.ivoa.net/twiki/bin/view/IVOA/WhoIsWho
[52] http://fits.gsfc.nasa.gov/fits_home.html
[53] Project websites with data presentation are located at: http://www.civenv.unimelb.edu.au/~jwalker/data/oznet/ and http://www.civenv.unimelb.edu.au/~jwalker/data/sasmas/moisture.htm

Researchers consulted during the audit:

- Dr Jeffrey Walker, Project Leader
- Mr. Rodger Young
- Dr. Olivier Merlin
- Mr. Rocco Panciera
- Ms Clara Draper

Funding sources are

- ARC grants
- Departmental/University funds

The financial sustainability of this group is an issue with it relying on small annual grants to maintain the collection and management of the data. It relies heavily on access to limited faculty resources and post-graduate student involvement to ensure the maintenance of the dataset.

Data Management Processes

This project collects and manages data from a network of sensor instruments at multiple locations in the field. Significant resources are involved in this data collection including the maintenance of the instruments. The data collected is spatial soil moisture and ground-based soil moisture data; assembling it in different regions and in different spatial resolutions and scales.

Data Acquisition: Raw data are acquired via ground-based sensor instruments, field based weather instruments and aircraft surveys. Sensors take very intensive ground-based measurements of soil moisture on a continual basis. Satellite data and data from Bureau of meteorology

Instruments are mostly maintained manually and data are uploaded at the remote field through a phone modem and transferred back to computers for downloading. For half the sites (18) researchers have to physically go to the instruments to download data while remainder is phoned in remotely, which still takes around 10 hours to complete for remaining sites. This download occurs every 20 to 30 days and is managed manually with no automated or scripted processes; something the project is keen to explore to increase project efficiency.

Air campaigns are usually an annual event involving around 30 people from Europe, the US and Australia helping with the aircraft and collecting ground-based data.

IP/Copyright of Data and scholarly output: All IP is owned by the researchers and post graduate students conducting the research. Property owners where instruments are maintained are provided with detailed monitoring data from their properties on request.

Data Quantities: Long term monitoring data generates about 1 GB of raw data per year. Airborne studies generate about 100GB of raw data. All scholarly outputs from this data are additional.

Data storage and Backup: Sensor instrument setups are able to store between 36 and 90 days of data depending on the instruments and configurations. Once downloaded, raw data and associated project scholarly output is currently stored on Faculty servers.

There are multiple copies of the data. Faculty web server has a weekly tape backup and a copy on web server, our group's server and on researcher PCs. For the airborne data there are 2 copies on DVD and on 2 hard drives. Uncertainty about the back up processes for other project output was evident and the need for strategies for more regular archiving of data onto different media was raised.

Data formats: Commercial formats and software is used by this project. Raw data off the instruments is in DB4 and CSV. Data are stored and accessed in *Microsoft Excel* spreadsheets with macros. Some data are archived using zip files. Images are stored and accessed in JPEG format.

Metadata: Information about the data, including technical, environmental and provenance information, is contained within the excel file that stores the data. The team is currently looking at ways to incorporate metadata software with the project's image collection.

There are no set standards for metadata schema for this research community with little coordination of efforts around data management at a national or international level.

Data Access, Authentication, Authorisation and Security: Data are accessed via the project website. The top layers are open access but the data are currently in a secured passworded area. The intention is to open this up once the site is working properly.

Authentication process is a rudimentary password control on this. And there are two levels of control: the contact details and locations of the monitoring sites and the actual data itself. Only core project staff is able to access the location and contact details for the properties and property owners. Researchers who want to use the data are issued with a password for access. Access to airborne data is closed for two years and will then by made available to other researchers.

PHASE TWO CONSULTATION

This project expressed interest in further consultation around the areas of:

- The telemetry underlying the sensor data transfer from APAC
- Web publishing of integrated multimedia data
- Digital asset management systems

Activity included:

- Dr Gerard Borg[54], ANU/APAC has made some initial communications with the project team regarding the project's telemetry set up at the field site. Some investigation of their software is underway. Delays have been caused by the heavy field demands on the project team resources.
- Information about DigiTool will be made available when the software because available for university users.

3.1.7 Molecular Medicine Informatics Model – Bio21

INITIAL AUDIT

The focus of this audit was the Bio21: Molecular Medicine Informatics Model (MMIM)[55]. MMIM is a platform which provides clinical researchers with access to data from disparate existing databases across multiple disease types at multiple institutions, co-located in a virtual repository and which can be linked with publicly available research and genetic profiling data. The data has been collected by participating researchers over 20-30 years, but most is more recently acquired. During the pilot the data included data from researchers in the fields of Epilepsy, Diabetes, Oncology, and the Tissue Bank. The next phase of the project will see expansion in these datasets; particularly for oncology and neuroscience as well as the new disease set of Cystic Fibrosis.

Project team members consulted were:

- Dr Marienne Hibbert, MMIM Senior Project Manager
- Mrs. Naomi Rafael, Senior Database Administrator
- Mr. Henry Gasko, Image Project Manager and Research Data Coordinator
- Mr. Frank Devuono, Melbourne Health.

[54] Plasma Research Laboratory, Research School of Physical Sciences and Engineering
[55] http://mmim.ssg.org.au/

Project Partners/Collaborators:

- The initial MMIM project was a pilot collaboration with: Melbourne Health; Western Hospital; Austin Health; Peter McCallum Cancer Centre; The Alfred Hospital; The Ludwig Institute for Cancer Research; Bio21 Institute; The University of Melbourne, and Victorian Partnership for Advanced Computing (VPAC).
- Successful outcomes from the pilot resulted in securing federal funding to expand partnership to: Royal Children's Hospital; St Vincent's Hospital; Monash Medical Centre; Box Hill Hospital; Cabrini Hospital; Royal Hobart Hospital; Menzies Centre for Population Research; Murdoch Children's Research Institute; South Australia (Flinders RAH, QE) and APAC.

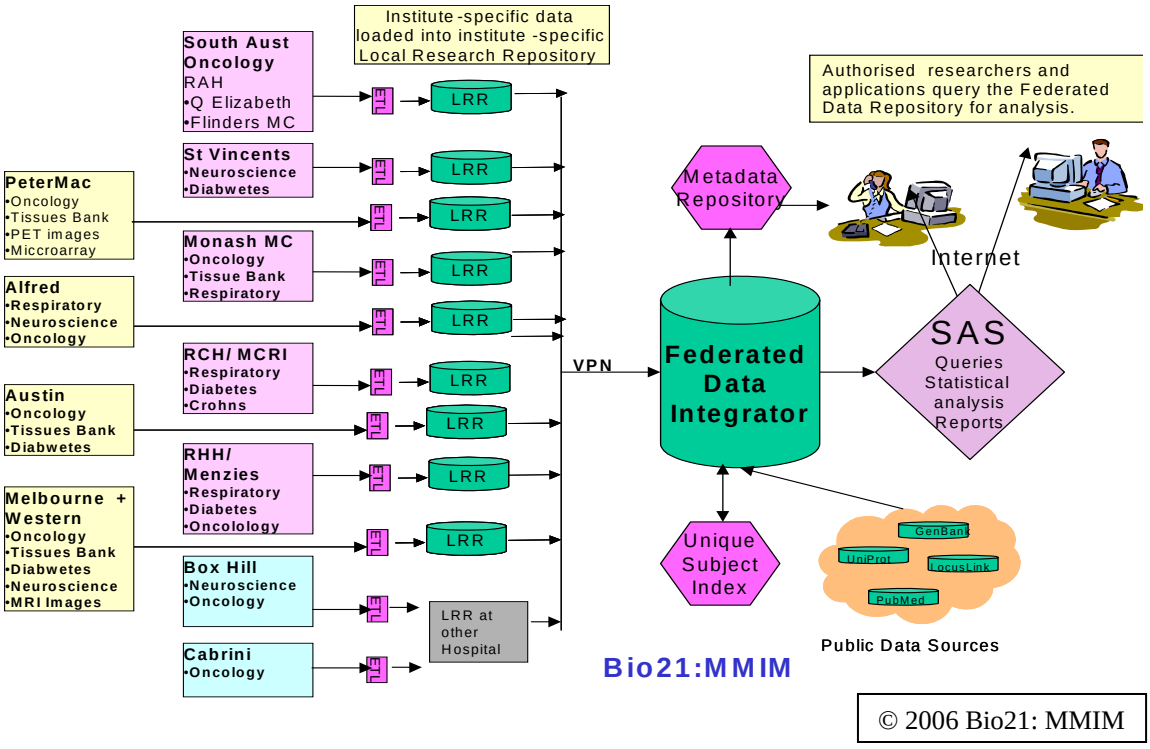
Funding sources:

- Phase 1: Science, Technology and Innovation (STI) Initiative, Victorian Government
- Phase 2: Department Education, Science and Technology (DEST), Australian Government
- Phase 3: STI/DIIRD (Department of Innovation, Industry and Regional Development, Victorian Government

Data Management Processes

Data are essentially patient/subject records and a variety of supporting information: Patient histories, including demographics and health issues; Pathology and radiology test results; Tissue sample analyses; Genomic data – biomarkers microarray, and various Images. The MIMM system maintains a database of these records which can be integrated across each LRR (Local Research Repositories). The data on the LRR is in the original format as per the institution of origin. The FDI (Federated Data Integrator) converts for access mode when interrogated but this is not stored by the system. The output of the FDI query would be stored by the researcher in that version, and may undergo further conversion post hoc.

This architecture is illustrated below and represents the MMIM model for the second phase the project. The LRRs are physically co-located with the original data source (e.g. hospital).



Data Acquisition: The form and content of the data ingested is determined by the host institute. Researchers access data by interrogating the system with a query at the Federated Data Integrator (FDI). All the data are based on research outputs from individuals participating in various research

projects conducted at the partner institutions. Subjects participating in this research must agree to have their data re-used before it can be included in the MMIM FDI.

IP/Copyright of Data and scholarly output: IP/Copyright of data is a complex area with current legislation stating that the hospitals are custodians of data collected from subjects and that information belongs to the subject. The MMIM model was structured to ensure that data remains on hospital property. There are currently no processes for tracking scholarly output produced using MMIM data.

Data Quantities: Rough estimates of the number of data records in the MMIM Federated Data Integrator (FDI) over a 12 month period are:

Pilot, 2005: 2,129.919
April 2006: 2,192,439

Essentially the growth amounts to 70,000 new records added to the ~80,000 patient records that were originally loaded. A lot of the records in the FDI are public data records and not patient records. These data estimates reflect the static nature of MMIM so far after the initial loading in 2005. The growth in 2006 will be directly related to the new databases to be federated into MMIM. After that, organic growth is expected as existing databases increase their volumes.

The FDI has approximately 286 MB of data and the capacity in the present configuration is 119,003 MB (116 GB). The Local Research Repository (LRR) at Melbourne Health is 1,075 MB (1GB) volume with a 77,497 MB (76 GB) capacity. Each partner LRR is currently of similar capacity.

Data storage and Backup: This is an ongoing federated collection that alters on a daily basis (updating and adding) which creates some problems for dataset access. Data that may have been accessed at a particular point in time cannot actually get replicated by the system as these data are not stored by the system. It is reliant on the researcher who uses this dataset to maintain this version of the data used – particularly if this has been used for some scholarly output.

More specifically, LRRs are maintained on the site of the host organization, i.e. each hospital. The FDI is maintained and hosted at Melbourne Health (MH). Data are updated on a daily basis and as such a new version of all data is created at the LRR daily. The FDI output is not kept – except by the researcher who would be expected to save this to his/her remote site. All servers are backed up nightly by MH who manages the server.

Data formats: All software and formats used directly by the project are Commercial. *IBM DB2 UDB* is the MMIM database but the institutional data could be in any type of database as the source data are extracted into a *DB2* repository at each site. The system can determine what this is when initially integrating a new database. Data transfer is then automatic. Data are stored however it was originally stored it does an extract transform load (ETL) between the source and site's and database.

IBM Websphere Information Integrator is the federator. The databases at the sites use whatever they like whether *DB2, Sybase, Oracle* etc.

Metadata: The current phase of the project is focusing on the common standards across all the databases to be accessed within MMIM; mapping to standards, getting the glossaries and metadata dictionary right and making it searchable. Existing standards in health will be used as much as possible. SNOMED-CT[56] is one that's been accepted as the standard for health federally, but each medical discipline may have different standards.

Technical metadata items will be annotated and the hierarchy is being developed to enable item discovery using MeSH[57] terminology. The metadata is stored in a relational database and stored in a metadata repository which is currently under development.

Metadata creation is the responsibility of data owner in conjunction with the MMIM team

[56] More information about this International standard is available at: http://www.snomed.org
[57] National Library of Medicine, US. More information about these medical subject headings is available at: http://www.nlm.nih.gov/mesh/

Data Access, Authentication, Authorisation and Security: Access to data is via a secured web interface to enable access to the *IBM – DB2* database using a SAS (terminal server) as the query and analysis tool. The project is currently looking at Grid technologies to assess if they can provide increased services.

Security of subject identity/privacy is managed within the system by the *IBM Websphere Data Integrator*. It's a two-stage process. It's a virtual repository. The critical stage is the first part when the identifying information first passes through a virtual private network through the federator to a record linkage program that checks if the individual exists. If not, a unique number gets sent back to be stored on the local research repository at that site, so the data query that a researcher does against the system, only has data attached to a unique subject index and no identifying information. This also allows for possible re-identification of that subject if ethically you need to do so.

Authorisation is required for researchers to use MMIM data. Researchers apply via an Access Request Form[58] providing details of investigators, clear description of the Research Project for which the data are to be used, including the science behind it, the databases required and how the data accessed will be stored and archived for the term of the project and beyond (including when it will be destroyed). These forms go to the MMIM Management Committee and to the Ethics committee if required.

PHASE TWO CONSULTATION

This project expressed interest in further consultation to identify and review their data management processes. The focus of this follow up was particularly around middleware selection, metadata schema frameworks and open access publications. This activity has commenced and will continue as part of the services provided by Information Services personnel at Research Computing and Information Management.

Activity included:

Consultation about various middleware for supporting distributed and Grid technologies

- Involvement with the SRB testbed at Research Computing
- Exploration of MAMS project outputs around identity and access management for web applications/interfaces including open source software, Shibboleth.

Consultation about metadata schemas

- Connecting with Experimental Particle Physics team to look at how they are using the Multi-disciplinary Scientific Metadata Management[59] framework developed by CCLRC in the UK and to assess utility for MMIM metadata framework development.
- Working with Digital Repository Coordinator to identify how preservation metadata frameworks like PREMIS[60] can be used within the MMIM database to ensure sustainability.

Consultation about Open Access Publication

- Working with Digital Repository Coordinator to develop strategies for increasing exposure of the MMIM data via publications of scholarly outputs using this data into UMER, the University of Melbourne ePrints repository.

[58] Access Request Form can be viewed at: http://mmim.ssg.org.au/join.htm
[59] Discussions held with project leader, Dr Kerstin Kleese van Dam at the UK eScince Centre has facilitated collaboration with Melbourne scientists wishing to use this framework. More information about the schema is available at: http://www.e-science.clrc.ac.uk/web/projects/scientific_metadatamgnt
[60] Maintenance of activity around this international standard is managed by RLG-OCLC. More information available from Library of Congress site: http://www.loc.gov/standards/premis/

3.1.8 PARADISEC and EthnoER

The focus of this audit was the two projects, PARADISEC[61] and EthnoER[62]. PARADISEC is an archive that offers a facility for digital conservation and access for endangered materials from the Pacific region. EthnoER is particularly focused on supporting secure and distributed collaborative research based on digital media and data repositories. Part of the project scope is to support interoperability of datasets and research tools.

Researchers consulted during the audit:

- Dr Nick Thieberger, Project Manager, Department of Linguistics and Applied Linguistics
- Associate Professor Steven Bird, Department of Computer Science and Software Engineering

Project Partners/Collaborations:

PARADISEC:

- University of Melbourne
- University of Sydney
- University of New England
- Australian National University
- GrangeNet

EthnoER

- University of Melbourne
- University of Sydney
- Australian National University
- Macquarie University
- Queensland University
- CSIRO
- Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS)
- Galiwin'ku Indigenous Knowledge Centre, NT
- School of Oriental and African Studies, London
- University of Alaska at Fairbanks
- University of Texas at Austin

Funding sources are

- ARC grants
- Departmental/University funds
- Grangenet

Financial sustainability is an issue for this research community. "With a secure bloc of funding over five years the current staff could establish processes that would secure the collection, and workflows could be disseminated through the research community to enable automated data ingestion. Current annual grants place pressure on the project and do not allow us to guarantee staff longevity of employment, risking a loss of skills that jeopardize ongoing operations." (Thieberger)

Data Management Processes

Issues across both projects will be highlighted though more emphasis will be placed on those associated with PARADISEC.

Data Acquisition: Data are deposited by contributors for conservation. Acquiring data from remote locations, particularly when not digitized, e.g. tapes, are typically sent with a trusted agent someone is

[61] Pacific and Regional Archive for Digital Source in Endangered Cultures, http://www.paradisec.org.au/
[62] Ethnographic Eresearch, http://ethnoer.unimelb.edu.au

traveling to these locations. Items have also be transferred by registered mail when necessary but this is not the preferred option given risks of damage to items.

The collection is three years old but original data are typically as old as 50 years. Non-digital data are digitized; a backlog has resulted in part from our high profile. Priceless tapes and transcripts from researchers in other countries have been deposited with no similar facility available elsewhere in the region. There are problems with interoperability of formats particularly in relation to metadata.

There are insufficient resources to meet the current and growing demand for digitization prior to data ingestion into the collection. AUSTEHC[63] has assisted with some of this.

IP/Copyright of Data and scholarly output: For the PARADISEC collection there are moral and ethical issues associated with the deposit and ownership of materials. The overall mission is to ensure preservation of the item by the legal custodian who may not necessarily by the 'owner' of the IP or cultural IP of the item as such[64]. Researchers using data are expected to acknowledge the owners and the collection as outlined in the Conditions of Access[65] policy.

Data Quantities: The PARADISEC collection is an ongoing and growing collection which currently holds about 3813 items comprising 2.6TB. Recently the collection has increased the amount of video data deposited which will increase the data storage demands.

Data storage and Backup: The PARADISEC archive is a collection of National and International significance which intended for long term preservation.

The server in Sydney is treated the official storage point of the definitive copy (on disc). The back up data are held by APAC at the Peta-Store at ANU. Ideally Melbourne would also have a mirror of the data, and funds are being sought from the ARC-LIEF scheme in 2006 to establish such a server at Melbourne.

Data are backed weekly onto tape with a full set of tapes in a fireproof cupboard in the PARADISEC lab in the Transient building at the University of Sydney. If the tape robot in Canberra failed there would still be the original sitting on backed up tape from the server, Azoulay in Sydney

In Sydney, when Azoulay:Archive partition gets to 60% full it triggers an automotive archive process to remove the oldest files in order, to take capacity down to 20%. Archive process stores batch of files on 200GB digital tape under a single archive tag. The tag and file names are emailed to project members.

Data formats: The preferred formats for both projects are for open source however there is some reliance on commercial software and formats within the broader researcher community. With PARADISEC, formats vary for different data. Digitisation is a major part of this project and current formats are limited to the following:

- **Text:** RTF, PDF, TXT, XML (Text character content - ASCII, Unicode)
- **Images:** TIFF, JPEG, PDF
- **Video:** .dv, MPEG3, MPEG 4
- **Audio:** PCM, MPEG3

With EthnoER the focus is more on the collaborative platforms and tools for the conduct of the research; including working with data in the field and held within digital repositories like PARADISEC. Links to various annotation and editing tools are provided on the project website including: Elan, CLAN, and Annodex. Supporting researchers in the use of these tools is a core objective of this project.

EthnoER is developing an online presentation and annotation system for archiving data; aimed also at providing a workflow for researchers that ensures they are using annotation tools to provide standards

[63] Refer 3.1.2 for more information about AUSTEHC.
[64] More information about the IP/Copyright policy is available at: http://www.paradisec.org.au/PdsCdeposit.rtf
[65] More information about the Conditions of Access policy is available at: http://www.paradisec.org.au/PDSCaccess.rtf

conformant and interoperable outputs. A problem with current work methods is that many tools are relatively new and each has its own characteristics; not all of which constrain the user to provide well-formed data as an output.

Metadata[66]: Current standards are based on those established by Open Languages Archives Community[67] (OLAC), the Broadcast Wave Format (BWF) metadata, and the National Library of Australia (NLA) metadata set. BEXT is the metadata encapsulated in Broadcast Wave Format (BWF) audio files.

EthnoER is mapping relationships between parts of the data as Dublin Core struggles with this aspect of data. The project is also looking at RDF metadata schema and Semantic-web technologies to manage archival objects in multiple parts.

The collection catalogue distinguishes three levels: the collection, item and file. Due to time and funding limitations PARADISEC records the bare minimum of metadata focusing on: rights and permissions; technical metadata; administrative/management and researcher annotations. The metadata is managed in an open source relational database - MySQL/PHP.

There are regular exports of parts of the metadata providing XML encoded data for an OLAC-compliant static repository, a generic catalogue listing for the APAC data store, and header information that is encapsulated into the BWF audio files.

Data Access, Authentication, Authorisation and Security: APAC manages the web presentation of the collection which is where the data access occurs. Anyone with access to a web browser is able to anonymously conduct a "Quick Catalogue Search" by clicking on this icon on the login page[68]. This search provides a listing of the entire collection with the following information:

- the item's unique identifier
- the item title (variable detail)
- the collector's name
- source language (as given)
- the country
- whether the item is digitized
- date when last modified
- details of the item (varies across items with most items allowing access to this description)

Researchers/users who wish to access the items apply for access to the collection via an access request downloaded from the PARADISEC website.

3.1.9 Household Income and Labour Dynamics in Australia (HILDA) Survey[69]

The focus of this audit was the HILDA Survey, a nation-wide household panel survey that focuses on issues relating to families, income, employment and well-being. The survey began in 2001 with a large national probability sample of around 14,000 individuals in almost 8000 Australian households and aims to provide, on an annual basis, longitudinal panel statistics describing the ways in which people's lives are changing in Australia[70].

The Project Director of the HILDA is Professor Mark Wooden. Two members of the project team were consulted during the audit: Ms. Nicole Watson, HILDA Deputy Director, Survey Management and Mr. Simon Freidin, HILDA Survey Research Database Manager and Analyst.

---

[66] Detailed information about project metadata is located at: http://www.paradisec.org.au/PARADISECMetadataset.rtf
[67] More information about the Open Language Archive Community and standards is located at: http://www.language-archives.org/ and http://www.language-archives.org/tools/search/
[68] PARADISEC login at: http://azoulay.arts.usyd.edu.au/paradisec/login.php
[69] Project website is located at: http://www.melbourneinstitute.com/hilda/
[70] Headey, B., Warren, D., and Harding, G., (2006) *Families, Incomes and Jobs: A Statistical Report of the HILDA Survey*. Melbourne: Melbourne University Press. Available at: http://www.melbourneinstitute.com/hilda/statreport/statreport2005.pdf

Project Partners:

- Melbourne Institute, Faculty of Economics & Commerce, University of Melbourne
- Australian Council for Educational Research
- Australian Institute of Family Studies
- The Australian Government Department of Families, Community Services and Indigenous Affairs (FaCSIA)[71]

The project is funded solely by the Australian Government Department of Families, Community Services and Indigenous Affairs (FaCSIA).

Data Management Processes

The dataset associated with this project is not typical of 'data-intensive' projects as such. However, from a social science perspective, the size of this dataset is significant, particularly as it is expected to be an indefinite life panel of the Australian population with a current funding commitment of $40 million.

Data Acquisition: Raw data are collected by ACNielson (ACN) via face-to-face interviews in participant homes; a process taking up to 8 months. Most of the data collected is manually entered on paper questionnaires by these interviewers and later punched into *SurveyCraft* and later converted to *SPSS* data files. One questionnaire is scanned and converted into TIFF files via *Hands and Eyes*, then imported into *SurveyCraft* and later converted to *SPSS* data files. Data from ACN is ingested into the HILDA server as *SPSS* files. Data are cross-checked and cleaned

IP/Copyright of Data and scholarly output: All IP and Copyright of the data belongs to the Australian Government. Scholarly output from the data is owned by the researcher.

Data Quantities: For each annual Wave around 40,000 questionnaires are completed with the project currently completing its sixth Wave. Data are pre and post processed resulting in a current data store, including raw data, analyses and associated programs, of about 200GB.

The datasets going to users, including all Wave Releases and some programs, can be managed on a single CD.

Data storage and Backup: All completed paper questionnaires are archived by ACN until the end of the life of the project. ACN also maintains copies of the raw digital data and a database of the personal information of participants including their contact details. Other project digital data are stored on the HILDA server which is maintained behind a firewall to meet the "In Confidence" security requirements stipulated by the Government contract. The work area for the data must also be located in an environment with secure access only to HILDA team members.

Data are updated with each Wave using cross Wave checking techniques and where errors or discrepancies are found datasets are updated and the previous Waves are re-released. Data are therefore not constant which may have implications for researchers who use older versions of the data for scholarly output.

Storage and back up procedures:

- All raw data are stored on Database Manager's removable hard disk.
- All original CD/DVDs from ACN are locked in secure cabinet.
- All the programs are backed up locally by the Institute using their back process.
- One external firewire drive used to backup the data several times a year.
- Full back up made on DVD annually and stored off-site with Information Services.

Disaster Recovery:

- Access to individual team member backups, including off-site copies of the data and programs.
- Access to ACN data (in Sydney) of original raw data sent to the project.

[71] http://www.facsia.gov.au/

Long term preservation of the data:

- Current project funding is sufficient to sustain the existing data for the life of the project.
- When the project ends the Government will decide whether the data should be preserved. Current contract states that once the project ends and there is no future for further surveys that all paper base questionnaires would have to be shredded and that all the data would be returned to the FaCSIA. No decision has been made regarding that long term. It is possible that some version of the Confidentialised dataset may go into a National Archive like the Social Science Archive at ANU.

Data are available to approved users in one of two versions:

- ***Confidentialised Dataset*** – some information is removed from the dataset to make it less likely for someone to be identified as a respondent e.g. taking away CD level, geographic information, SLA, postcode, detailed occupational and industry codes. Only have State or part of State identified.
- ***Unconfidentialised Dataset*** – still does not include names and addresses of participants but all geographic information, postcode, CD, top level coding on income and wealth and puts back 4 digits occupation and industry coding.

Data formats: All formats and software used by the project are commercial. The data formats include *SAS* datasets, *SPSS* datasets, PDF of the Code Frames, a set of frequencies, and PDF of the marked up questionnaires which are produced in *QUARKexpress.*

Metadata: *SPSS* creates the metadata framework and it is stored with the numeric data and the numeric context data so we can manage the whole data dictionary simultaneously with the whole collected data. *MSAccess* database is used to manipulate the metadata and produces PDF Code Frames which are different ways of looking at the metadata.

Data Access, Authentication, Authorisation and Security: There are currently about 650 direct users of the data. To gain authorization to access the data prospective users must submit application[72] to gain access the data. If the application is approved the user must sign contract/deed of confidentiality[73] with FaCSIA, Commonwealth Government. This will provide access to the Confidentialised version of the data. Very few users are able to access the unconfidentialised dataset and they must provide a secure facility to access that data and sign a different deed. Audits of security arrangements must be made on an annual basis to a selection of the users. The Melbourne facility has also been audited by security consultants.

3.1.10 Australian Sound and Design Project[74]

INITIAL AUDIT

The focus of this audit was the Australian Sound Design Project; an integrated archive of Australian interdisciplinary sound design of public space. This is a collection of national and international significance, being the first auditory archive of sound design in public space. An important feature of this collection is that many of the items within it represent *ephemeral, spatial architectures in time*. The project works to bring these installations via an integrated multi-media web presentation to expose and preserve them.

Project team members consulted during the audit were:

- Dr Ros Bandt, Project Leader
- Mr. Iain Mott, Technical Manager
- Mr. Gavan McCarthy, Director AUSETHC

[72] The Application form can be viewed at: http://www.melbourneinstitute.com/hilda/data/OrderAustR4Confid.doc
[73] The *Deed of Confidentiality* can be viewed at: http://www.melbourneinstitute.com/hilda/data/DeedAustR4Confid.pdf
[74] The project website is located at: http://www.sounddesign.unimelb.edu.au/site/index1.html

Funding sources are:

- ARC grants
- Departmental/University funds
- External sources: Melbourne City Council; City of Yarra; Australia Council New media Arts Board; Move Records, and The National Library of Australia

Financial sustainability is an issue for this research community. The need to constantly seek external funding via annual grants place pressure on the project and does not allow for staff longevity of employment, risking a loss of skills that jeopardize ongoing operations.

Data Management Processes

Data Acquisition: Data are provided to the collection by the Designer/Artist/Composer/Curator in a variety of formats. There are currently no restrictions on the formats accepted by the collection. Contributions undergo a variety of processes to enable the presentation of the integrated work in a web ready format.

IP/Copyright of Data and scholarly output: The IP of the works in the collection remain with the artist however they provide the Project with the license use and display the work[75].

Data Quantities: The archive is an ongoing and growing collection. Currently the archive consists of 345 entities with references to 340 published resources; 140 of the works are in multimedia products. The web based data constitutes only around 2GB of data. But when all the back ups and data stored on CDs and DVDs are included and software etc it would ten times this.

Data storage and Backup: This project aims to preserve the collection indefinitely. Currently, the web-based collection is managed using OHRM[76]. The dynamic content (PostgeSQL) with php scripts) and the site free-text indexing (htDig) is located on the AUSTEHC server and the static pages are located on the Arts Faculty server. Data on AUSTEHC server is backed up weekly with offsite tape backup. The Arts Faculty server also has a tape back up protocol. There are essentially three versions of all of the data that is in the OHRM at different levels of preservability with one being proprietary.

All other digital project data are stored on desk top computers. Back up copies are made when new data are contributed to the collection; usually on CD and DVD and kept on site. Administrative records and information about the project and the collection is stored in a variety of digital and non-digital formats. Some of this data is possibly at risk of loss should staff leave the unit.

The preservation of core digital data was identified as a top priority for the project. Data that needs preserving includes MPEG files, PDF, Quick Time movies, MPEG movies, the OHRM database and administrative records.

Data formats: Data formats and software are a combination of Open Source and Commercial. The OHRM, an open source tool from AUSTEHC, is the core management tool for the collection.

Data are acquired from contributors in a variety of digital and non-digital formats.

- Images: architectural plans, slides, prints, diagrams, sketches, TIFF, JPEG, PDF
- Audio: CD, DAT
- Video: MiniDV, VHS, MPEG, Quicktime movies
- Text: Word, RTF, .txt
- Codes

The data exist in different versions. To illustrate, a high resolution copy of media may be contributed on either a DV tape or a 44.1 kilohertz sound file, but it is published using 128 bits per second MPEG version

[75] Information about the *Deed of Licence* is available at: http://www.sounddesign.unimelb.edu.au/site/contribute.html
[76] Online Heritage Resource Manager, - open source context based resource discovery and access system developed by AUSTEHC. More information available at: http://www.austehc.unimelb.edu.au/ohrm/

of this sound file. Both versions are stored and at a later date this can be reassessed. The artist is informed and consulted when reformatting of their work is necessary.

Metadata: The metadata schema around the collection has been built locally by the project team and focuses on describing the relationships between the artwork, the defining of the particular type of work and how it relates to other entities within the database. This is incorporated into the context of the OHRM interface. There is also technical documentation about formats and conversions carried out on the files. No metadata manual or dictionary has been formalized to date but there are standards that have been developed. The National Library has funded the conversion of data to its metadata system.

Data Access, Authentication, Authorisation and Security: This is a web based, open access collection freely available to the general public. Contributions to the collection are encouraged from all sound designers of public space in Australia.

PHASE TWO CONSULTATION

This project expressed interest in further consultation to identify and review sustainability issues around the long term preservation of the collection and their records. This activity has commenced and will continue as part of the services provided by Information Services personnel in Information Management.

Activities included:

- The Digital Repository Coordinator contacted the NLA to commence processing for regular archiving of the Australian Sound Design Project website: http://www.sounddesign.unimelb.edu.au/site/index1.html Publisher's copyright and disclaimer statement was provided and the site is now being archived by Pandora. It is located at: http://pandora.nla.gov.au/tep/58565 and will be updated at the end of August 2006.

3.1.11 The Kidneyome project

The Melbourne Kidneyome project forms part of the Physiome Project; a worldwide collaboration of loosely connected research teams in New Zealand, Australia, France, the US, the UK and Denmark. The Physiome Commission of the International Union of Physiological Sciences, IUPS, provides leadership to the Physiome Project through its satellite and central meetings and through the University of Auckland's IUPS Physiome Website[77].

The focus of this audit was the Melbourne role in the *"eResearch Grid Environment of Distributed Kidney Models and Resources"* project. This project[78] aims to establish an interactive web interface at the international level, to a collection of distributed legacy models at all levels of kidney physiology, with, for each curated model: documentation, physiological context, easily interpreted output, a statement of model limitations, interactive exploration, and user-customisation of selected parameter values. Interaction with the resource will be through a 3D-virtual-kidney graphical user interface (GUI).

Project team members consulted were:

- Professor Peter Harris, Faculty IT Unit, Faculty of Medicine, Dentistry and Health Sciences, Project leader.
- Dr Andrew Lonie, Department of Information Systems, Faculty of Science

Project Partners/Collaboration:

- Professor Peter Harris - Department of Physiology, University of Melbourne
- Dr Andrew Lonie - Department of Information Systems, University of Melbourne

[77] http://www.physiome.org.nz/
[78] Information taken from project proposal: D3 Outline of proposed initiative, provided by project team.

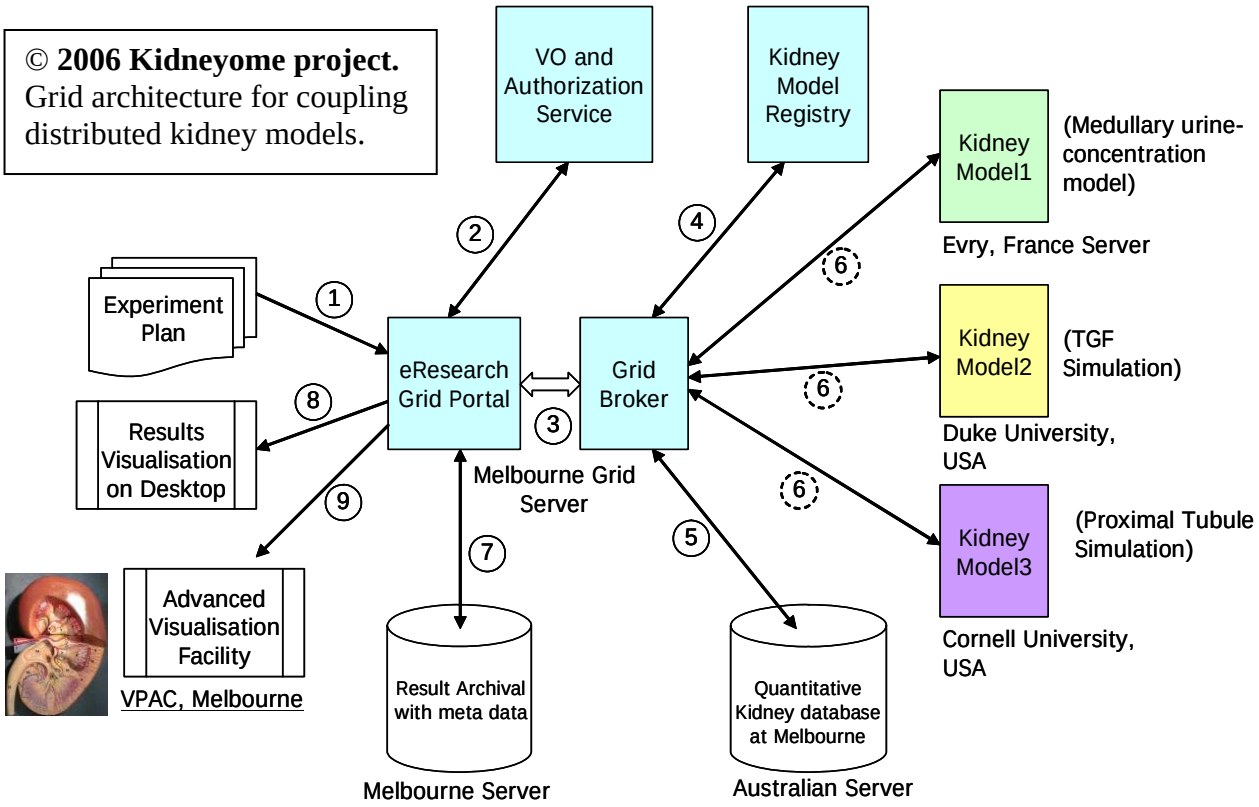
- Dr Raj Buyya, Department of Computer Science and Software Engineering, University of Melbourne
- Dr S. Randall Thomas, Informatiques, Biologie Intégrative et Systèmes Complexes, Universite d'Evry Val d'Essonne, France
- Department Mathematics, Duke University, North Carolina, US
- Department Physiology and Biophysics, SUNY Health Sciences Center, Stonybrook, New York, US
- Bioengineering Institute, Auckland University, New Zealand
- Cornell University Medical College, Department of Medicine, New York, US
- Institute of Medical Physiology, University of Copenhagen, Denmark
- VPAC and APAC

Funding sources are

- ARC grants
- Departmental/University funds
- External grants via the collaboration

Data Management Processes

Data Acquisition: The data are currently acquired from within the collaboration. Data generated by synchrotron experiments or CT scanners, are integrated with simulated models developed by the collaboration and re-used by research partners. Acquisition occurs via the Grid interface (using Globus). This grid infrastructure (illustrated[79] below) is being developed as part of the collaboration, and identifies how data will be stored, distributed and accessed.



IP/Copyright of Data and scholarly output: The ownership of the data and its IP remains a matter of discussion among the international collaboration, particularly as it relates to the larger Physiome consortium.

[79] The diagram below was taken from the Project proposal provided by the project team.

Local processing of data is Melbourne IP as such but the new models are made available to the collaboration. It is expected that this is acknowledged and that the IP remains with the creator but accessible to other researchers. Raw data are accessed across the collaboration.

Data Quantities: The amount of data currently held is variable across the collaboration.

The Melbourne project maintains a small amount of data that is mostly derived/post-processed anatomical data (<1MB – text files) but it is projected that this will increase as the project progresses over the next 2-3 years to around 100GB.

The raw data are what is important to keep long term and this is currently maintained by other members of the collaboration. This is data produced from 'one-off' never to be repeated experiments in the synchrotron. There is a lot of this data from a single experiment. It is estimated that these holdings will be in the vicinity of 500GB by project mid-late stages.

Data storage and Backup: This is a collection of national and international significance (for researchers, teaching and practitioners in clinical practice) and will continue to grow as the Kidneyome and Physiome projects continue to plot and model all body systems. The Quantitative Kidney Database (QKDB[80]) which will be mirrored in Australia (as per diagram above) is public access with approximately 1,000 entries in the database; this will grow very quickly now. This database also contains references and comments relating to scholarly works which are accessible via interrogation of the system[81]. The QKDB is hosted and maintained offshore and based at LaMI, Evry University, France[82]

Current needs for this project are small but these will grow rapidly as the data continues to grow. Current resources are not sufficient for the needs of the project and are in effect the researchers' departmental allocations. Data will need to be housed elsewhere and some of the data could be maintained off line. A data centre/storage facility either within the institution or off site would meet the projected needs of the group.

Locally produced data are maintained on the Faculty servers including a back up facility. Despite a fundamental confidence that the data are well managed by these departmental processes there are no specific records around these processes. General information provided included:

- Data are maintained on the Departmental server (Information Systems-Science). It is assumed that this server will be under a standard back up protocol but the researcher was not aware of the specifics of this process. There is also an expectation that data are backed up on tape but how readily accessible this data would be if needed is unclear.
- Researchers maintain a single CD backup which is created at the time when the file is originally created. There is no maintenance or checking of these CDs.
- No information available regarding offsite data backup of the derived data.
- Raw data are mostly generated elsewhere in the collaboration and therefore available via the Grid network. This would not be the case for raw data are generated and maintained by the Melbourne team.
- It is unclear what the impact of system breakdown would have on project workflow delays and turnaround in the case of disaster recovery.

Data formats: Open standard formats including locally produced tools and applications by the collaboration are used in this project. Most of the data are in simple binary format. MicroCT, CellML[83] (XML based) is used to store and exchange computer-based mathematical models. A variety of open source freeware software is used for analysis/post-processing of data, e.g. when rendering data into a 3D image for presentation.

[80] http://www.lami.univ-evry.fr/~srthomas/qkdb/index.php
[81] Search accessed at: http://www.lami.univ-evry.fr/~srthomas/qkdb/query/query_form.php
[82] Le Laboratoire de Méthodes Informatiques (LaMI) – Data Processing Research Centre: http://www.lami.univ-evry.fr/presentation
[83] http://www.cellml.org/

Metadata: The current project is looking at the development of standardised schemas for classifying information (essentially metadata schema) for this community. VPAC is working with the project team on the ontology generation and the taxonomy side of this. It is clear from the project's goals that the areas which will be the focus of the taxonomy around the kidney models themselves will include:

- Model documentation
- Physiological context
- Interpreted output
- A statement of model limitations
- Interactive exploration capabilities
- User-customisation capabilities for selected parameter values.

The global community has metadata schema around some of the formats used to store and transfer some data. CellML includes mathematics and metadata by leveraging existing languages, including MathML and RDF. FieldML (XML-based) is also used. The international Physiome collaboration is currently working on the ontology for the project looking at the different perspectives of the hierarchy including those of the National Library of Medicine, NIH.

Data Access, Authentication, Authorisation and Security: Distributed data are accessed via the Grid; a closed environment only available to researchers within the collaboration. Authentication is via Globus protocol.

The aim is to make the data, particularly the modelling data, widely available for research and practitioners in the clinical context

The QKDB is open access with any user capable of submitting a query[84] to interrogate the database. Submitting data into the QKDB requires that the user to be an acknowledged scientist working in kidney research or a related field. An authentication process occurs at the website: http://www.lami.univ-evry.fr/~srthomas/qkdb/login/create_account.php and is managed in France.

[84] Query form is located at: http://www.lami.univ-evry.fr/~srthomas/qkdb/query/query_form.php

3.2 Researcher Capabilities and Expertise

The project identified a number of capabilities across the projects audited.

- Structured, collaborative data acquisition processes requiring integration of diverse data – AUSTEHC, PARADISEC, MMIM, ICCR-Education
- Grid technologies – Experimental Particle Physics, Dr Raj Buyya (Department of Computer Science and Software Engineering)
- Large dataset management - Experimental Particle Physics, Astrophysics
- Digitisation, archiving and preservation of print and multimedia data - PARADISEC and AUSTEHC
- HDMS – Cultural collection management tool, locally developed and supported by AUSTEHC
- OHRM - Web publishing tool for cultural collections, locally developed and supported by AUSTEHC
- Distributed virtual database/repository framework – MMIM
- Database management – HILDA, MMIM
- Video Analysis Research – particularly with *StudioCode* software - ICCR-Education

It became apparent during the project that there is limited opportunity to access information about the research expertise that exists across the university. Much of this exchange of information occurs by accident; often at social gathering or via loose networks among colleagues.

3.3 Sustainability considerations

3.3.1 Technology Issues

- Each project has independently chosen its data and metadata formats and handling procedures. This has resulted in there being a variety of commercial and open source formats and software being used, and consequently, few options for sharing expertise at the technical level.
- The frequent loss, or threatened loss, of technical expertise due to project based funding model which does not include sustainability considerations.
- There was an unmet need for access to expertise, information and/or technology solutions raised by six of the eleven groups

3.3.2 Curation/Archiving Issues

IP/Copyright of the raw data ownership varied across the groups. The ownership of data infers the onus to maintain the data.

- Six groups stated it belonged to the researcher/contributor of the data/item.
- One group stated it belonged to the global collaboration.
- One group stated it belonged to the instrument facility (observatory).
- One group stated it belonged to the research project partners.
- One group stated it belonged jointly to the researcher and the participant/patient.
- One group stated it belonged to the Australian government.

Metadata:

- There are a variety of locally produced, national and international standards in use.
- The quality of metadata across groups was variable.
- Biomedical groups in particular, have underdeveloped metadata schema/ontology for their data – these are mostly under development in collaboration with international bodies.

Access:
Data are accessed in a number of ways across the groups audited. Most have or are developing distributed models of data presentation and storage to improve access. One group has data security requirements that prohibit electronic transfer of data.
Research data versions are not stored by all communities. The onus remains with researchers accessing data for scholarly work to maintain their own copy of the data version that has been used for their analyses.

Authentication and Authorisation:
Groups have varying methods for the authorisation of users to access data; ranging from nothing at all (public anonymous access) to having requirements undertaking a legally binding contract regarding data access, use and storage.

- Three groups have their data accessible via a publicly available website.
- Two groups provided some access to their data via a public website but required authentication to access the data itself.
- Six groups had closed collections available only to project partners or researchers on application.

3.3.3 Data Storage Issues

Storage needs varied across groups. The current and projected quantities of data varied widely and the projected storage requirements over the next ten years for these groups will be in the vicinity of 600+TB with the two physics groups requiring the bulk of this resource. All groups are currently doing some data management; however documentation of how the data is/should be managed throughout its life cycle tends to be poorly assembled. Eight groups identified the need for well managed data storage facilities, particularly for their off site back up needs and for long term preservation. Disaster recovery planning is mostly ad hoc suggesting a reliance on the faith that backed up data will be accessible and that project workflow will not be greatly affected should disaster strike.

- Three projects have their data managed offsite by an external store, two by APAC and one offshore;
- Five projects use Faculty servers, and
- Five projects manage their own server for storage (some in addition to using Faculty resources).

Ten of the eleven groups stated the desire to store and preserve some of their data indefinitely. Eight of these projects do not have specific strategies in place for this preservation.

3.3.4 Sustainability Risk Factors

Back up and disaster recovery protocols are not well documented. Four projects appear to be taking some levels of risk with their current practice for some aspect of the dataset management, e.g. no managed off site back up.

Four projects identified a concerning lack of financial sustainability with short term project funding cycles.

4. Discussion and recommendations

In addition to the specific findings for each group audited, the project findings also provide information about more general sustainability of data management practices. Meeting the needs of the researchers interviewed will take resources, and at present much is left to the academic department; often leading to either no or limited action or to 'reinventing the wheel' and resulting in a less than efficient institutional response to eResearcher needs.

These findings point to a number of issues that can help to inform an e-research strategy for the university. Eight recommendations have been formulated for consideration by key stakeholders.

4.1 The importance of an institution-wide strategy for eResearch.

The findings from this project reinforce the work of Professor Geoff Taylor, Ms Linda O'Brien and the eResearch Advisory Group, identifying the need for an institution-wide strategy to progress and manage eResearch engagement and support. In particular, the findings demonstrate that when it comes to digital data management, there is variable capability among our research communities to comply with the University's Policy on the Management of Research Data and Records[85] and the (consultation draft) Australian Code for the Responsible Conduct of Research.[86] Data management, including access, discovery and storage, must be a fundamental component of such an institution-wide strategy. A broad eResearch strategy can also position the University to meet the challenges of the Research Quality and Research Accessibility Frameworks[87].

Recommendations three to eight below provide some of the essentials for such strategic planning. The Research and Research Training Committee (R&RT) would provide the governance for enabling its implementation.

***Recommendation 1*: That the University develops a strategy that broadly addresses the policy, infrastructure, support and training needs of eResearch.**

***Recommendation 2*: That the University's R&RT Committee consider forming a subcommittee to provide governance for enabling eResearch at the university. This committee should have broad representation and include Information Services and eResearch leaders.**

4.2 A lack of information policies and guidelines

There is a lack of best practice guidelines and policy statements available to support researchers with their data management decision making processes. The lack of shared language and terminology around many aspects of data and its management suggests the importance for all policies and guidelines to include clear definitions of concepts and terms used.

Areas of need include:

- Implementation of research record keeping principles and requirements.
- Data management for short term sustainability and long term preservation.
- Metadata standards, principles and systems:
    - Across the discipline divide.

[85] http://www.unimelb.edu.au/records/research.html
[86] http://www.research.unimelb.edu.au/hot/current.html#draft
[87] http://www.research.unimelb.edu.au/hot/current.html#quality

- For raw and processed research data.
- For web presentations.
- For other scholarly works.

- Authentication and authorisation standards and systems for access and storage of scholarly IP.

Recommendation 3 – that Information Services initiate a consultative process for the development of appropriate guidelines and, where relevant, policy statements, to support researchers with the management of their research data and records.

4.3 Absence of a coordinated data management infrastructure for research.

The findings suggest a need for centrally supported flexible data management, authentication and access systems. Groups audited were found to be managing their own data and developing their own access and presentation systems. The need was also identified by several groups for managed data storage facilities. Groups are supporting a variety of software. Group needs around authentication and access differed; requiring a variety of public, local, national and international collaborator access. The need for data management capabilities that are internationally interoperable; allowing for local storage and collections to federate internationally was highlighted. There will also be a need to promote among the University research community our capacity for digital data management. This emphasis on developing and marketing 'platforms for collaboration' through ICT within and across institutions is a key aspect of the National Collaborative Research Infrastructure Strategy (see capability area 16).[88]

4.3.1 A case for centrally supported data management, authentication and access systems

A centrally managed data storage and access facility would provide a secure, backed up and sustainable repository for data. This would allow the groups to concentrate on research rather than technology and allow for data to be preserved beyond the life of the particular project. It would also encourage more standardization as groups would find it easier to choose similar software and standards to other groups (group wisdom) leading to a consolidation of expertise. A centrally supported system could also act as a base or starting point for those research groups with no existing data infrastructure, reducing the need for group level development efforts and leveraging central support and expertise. The institution (campus) has been identified as "a logical nexus for the development of cyberinfrastructure … and that it is worth considering a holistic view that would promote larger, sharable, campus systems"[89].

4.3.2 The need for flexible infrastructure

Centrally supported infrastructure must accommodate the realities of the global collaborations of many of our eResearch communities. An authentication and access capability must allow for public, local, national and international collaborator access. Ideally, data management capabilities would be internationally interoperable and allow local storage and collections to federate internationally. It is recognised that it may not be possible for all research groups to take advantage of a centrally supported system as research domains and collaborations may dictate standards and software usage.

Central services need to concentrate on providing the base technical support of systems and facilities and allow users freedom to use these in whatever way they want. Ideally, central service would act as utilities where the utility has no interest in what the user does with the service.

Recommendation 4 – To review ICT infrastructure for research, paying urgent attention to data management infrastructure.

[88] See http://www.dest.gov.au/sectors/research_sector/policies_issues_reviews/key_issues/ncris/platforms_for_collaboration.htm
[89] Workshop funded by the National Science Foundation (NSF-US) to consider effective approaches for campus research cyberinfrastructure. Workshop report: http://middleware.internet2.edu/crcc/docs/internet2-crcc-report-200607.html

4.4 Capabilities needed by eResearchers

The audit identified expertise used in the conduct of eResearch across a variety of disciplines. The findings show that an eResearch consultation service needs to include at a minimum, information and access to expertise in:

- Database management
- Middleware development, management and support
    - Data management systems
    - Grid and other distributed systems
    - Authentication and Authorisation management
- XML advice and expertise
- Metadata advice: metadata systems, schema and taxonomy development
- Curation and Preservation advice and support for raw data and scholarly output
    - Business case development advice and support
    - Discipline based advice and support around sustainable data format selection
    - Obsolescence planning – knowing what to keep and why and what to delete

***Recommendation 5** – To establish a structured consultation process for eResearch support*

4.5 Difficulty accessing information about eResearch activity and capability

This project has identified problems around access to information around eResearch activity, capability and support with much information exchange occurring fortuitously. It is recommended that an information exchange strategy be established to increase the dissemination of information about support for eResearchers. A springboard to this process could be the delivery of an E-Research Expo in December 2006 to showcase university-wide activity in eResearch.

***Recommendation 6** – To establish an Information Exchange Strategy around eResearch*

Part of the information exchange strategy is a registry of research capability across the university would facilitate the dissemination of this information. The feasibility of linking such a registry to the Themis Research Management System should also be established; minimizing the need for duplication of data entry by our researchers.

***Recommendation 7** – To establish a Registry of e-research expertise*

4.6 Implications for education and training

The skill set for researchers is evolving and some consideration should be made to identify which of those associated with eResearch might be considered part of the essential generic skill set mix for trainee researchers, which might be discipline specific, and which might remain in the domain of expert service support.

It is considered that much of the expertise listed in 4.4 are not fundamental to all research disciplines and are therefore inappropriate for broad, in-depth education and research training. However, as research practices are rapidly adopting information and communications technology (ICT), researchers should be made aware of the services and expertise available to them; locally, nationally and globally. An awareness and basic understanding of research data policies, responsibilities, collections, curation, preservation, copyright/IP, metadata and standards must be included in a researcher and postgraduate induction program and reinforced throughout their candidature. An essential part of such a training program would include information about the terminology and underlying principles for managing data

throughout its entire life cycle. Academic staff supervising postgraduate students may also welcome an opportunity for such training. If a centrally maintained and supported data management, authentication and access system were to be provided, regular introductory courses would be necessary. Researchers and postgraduate students also need to be made aware of central research computing support (high performance computing, Grid and visualization services) available to them and how these services may help make their research 'smarter'. Other technical expertise and knowledge could be transferred as part of the consultation process outlined in 4.4, as project context and specifics can greatly affect the technology implementation.

Project findings reinforce the view expressed by the Australian Government's e-Research Coordinating Committee that "The ultimate success of the implementation of a strategic e-Research framework will be dependent on people with attitudes, skills and an understanding of the benefits that the framework can deliver":

> Three groups of skills development are needed to hasten the adoption of e-Research methodologies. Firstly, researchers need easy and structured ways of acquiring basic e-Research skills. Secondly, researchers need a researcher/skilled IT interface, to provide them with day-to-day support. Thirdly, researchers need high level ICT and information management professional support.[90]

We need to look at how we can assist University researchers (staff and students) to acquire and develop skills in e-Research to facilitate their research and to ensure compliance with data management requirements (incl. University Policy on the Management of Research Data and Records). This would include an understanding of research data policies, responsibilities, collections, curation, preservation, copyright/IP, metadata and standards. We need then to ensure they know to access skilled support and high-level infrastructure.

<u>Recommendation 8</u> – To review the implications of project findings for researcher education and training

[90] An e-Research Strategic Framework: Interim Report of the e-Research Coordinating Committee, 30 September 2005 – see http://www.dest.gov.au/sectors/research_sector/policies_issues_reviews/key_issues/e_research_consult/

5. Appendices

5.1 Audit questionnaire

Survey Questionnaire

QUESTION	**ANSWERS/COMMENTS**
Q1 Date of interview(s):	
Q2 Names of researchers	Q2A Interviewers:
Q3 Phone:	
Q4 Position:	
Q5 What is the name of the data/project?	
Q6 Association/Partners:	
Q7 What is relationship with this data? ❑ Primary creator ❑ Other creator ❑ Repository provider ❑ Administrator ❑ Other	?
Q8 University:	University of Melbourne
Q9 Broad subject:? ❑ Humanities? ❑ Social Science? ❑ Medical Science? ❑ Science – other ❑ Multidisciplinary	?
Q10 Who has prime responsibility for it? ❑ Primary creator ❑ Other creator ❑ Institutional repository ❑ Administrator ❑ Other	
Q11 General description of the dataset	
Q12 How many objects/digital objects are in the collection? Q12A How much data is there?	
Q13 What do you define as a digital object/asset?	
Q14 Is it ongoing, or closed, collection? ❑ Ongoing ❑ Closed ❑ Open but not yet closed, when to be closed?	

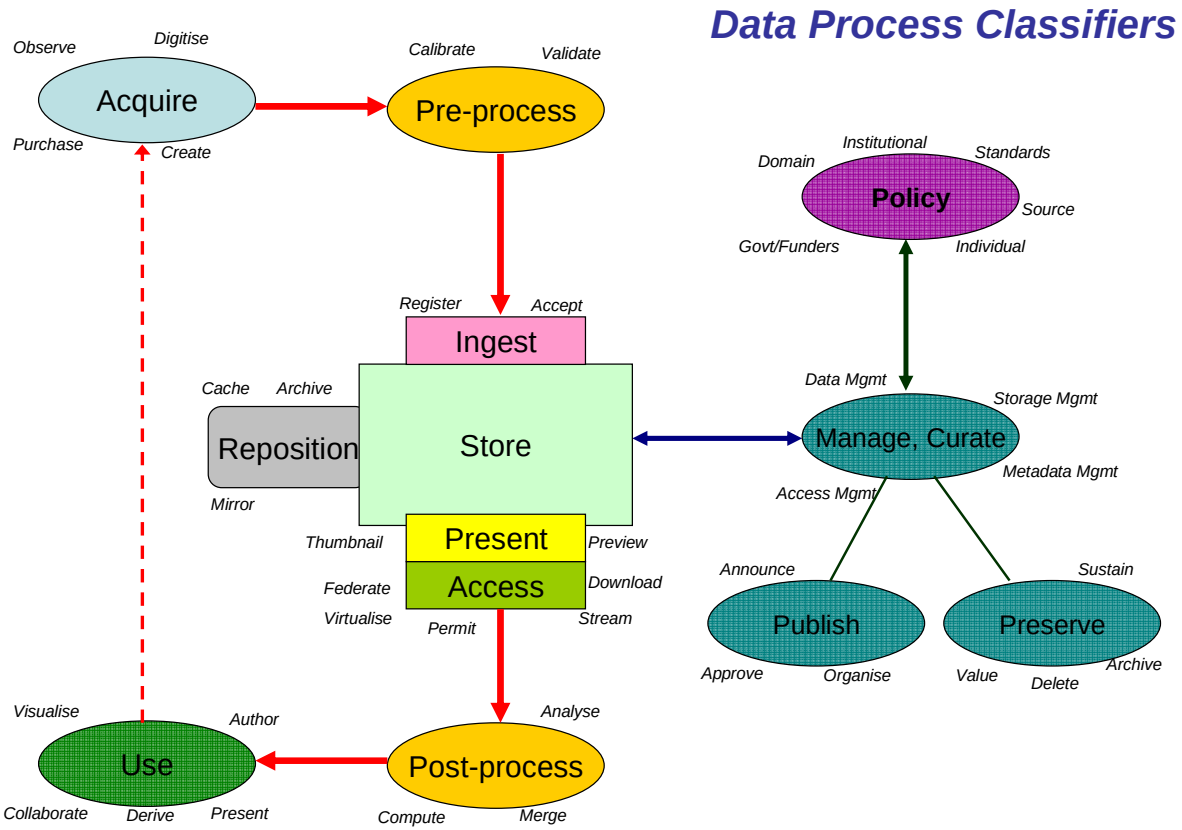
Q15 What is the source of the data? ❑ Created by researcher/s ❑ Published research output ❑ Re-use of purchased data ❑ Re-use of other data ❑ Other	
Q16 Who owns the copyright to this data? ❑ Creator ❑ Individual contributors ❑ Public domain ❑ Other, Who?	16a How does the current IP regime impact on the management of the data (does this vary for different datasets) What solutions do you see for these issues?
Q17 What formats was it created in? ❑ Analogue ❑ Commercial software, which? ❑ Open standard, which? ❑ Locally developed	?
Q18 Does it need to be, or has it been, converted into another format? ❑ Yes – some ❑ Yes - all ❑ No	
Q19 If yes, who will be/was responsible for doing this? ❑ Creator ❑ Local system administrator ❑ Local IT support ❑ External entity ❑ Other	
Q20 Is/was the process of conversion documented? ❑ Yes, How? ❑ No	
Q21 Is/was anything lost in conversion process? ❑ Yes, What? ❑ No	
Q22 Did it matter? ❑ Yes ❑ No	
Q23 What kind of quality control is/was used when converting the data? ❑ Visual scan ❑ Automatic checking ❑ Other	
Q24 What file formats are/will be used to access the data? ❑ Commercial ❑ Open standard ❑ Locally developed	Which? Comment,

Q25 What file formats are/will be used to store the data? ❑ Commercial ❑ Open standard ❑ Locally developed	Which? Comment,
Q26 What software is required to access the data? ❑ Commercial ❑ Open standard ❑ Locally developed	Which?
Q27 What are the main features of the software that you depend on?	Q27A Is it likely that storage of this software will be required at a future date?
Q28 Where does the meaning, or the importance of the data reside? (Is it in the accuracy of the colours or the layout, is the relationships or the functionality or what?).	
Q29 Is the data available in different versions? ❑ Yes ❑ No	
Q30 If yes, how do you define the different versions?	
Q31 What categories of metadata are/will be used to describe the data? ❑ Rights and permissions ❑ Provenance (documented history) ❑ Technical metadata ❑ Administrative/management ❑ Bibliographic/descriptive ❑ Structural ❑ Other	
Q32 Does this involve the use of any particular known scheme or standard? ❑ Yes ❑ No	If so, which?
Q33 Do you record metadata about different types of entity? ❑ Digital object ❑ Collection ❑ Non-digital source object ❑ File ❑ Metadata ❑ Other	

Q34 Who is responsible for the metadata creation? ❑ Primary creator ❑ Other project team member ❑ Individual contributors ❑ Research assistant ❑ Librarian ❑ Programmer ❑ Editor ❑ Other	
Q35 How is the metadata stored and updated? ❑ Relational database ❑ Bundled with related content files ❑ XML database ❑ Proprietary database or format ❑ Flat files ❑ Object-oriented database ❑ Other	
Q36 Where is the data held at present? ❑ Own/departmental server ❑ Institutional repository ❑ Remote server ❑ Own hard drive ❑ Floppies/CD/DVD/other media ❑ Analog ❑ Other	
Q37 Is there any backup procedure in place? If so, what is it? Where is the backup kept? ❑ Not backed up ❑ Backup held locally ❑ Backup located elsewhere	
Q38 How often is the data backed up? ❑ Once a week or more ❑ Once a month or more ❑ Other	
Q39 Who are the main users of this data? ❑ Self/research team ❑ Researchers in the same discipline ❑ Undergraduate students ❑ Postgraduate students ❑ Other	

Q40 How do the users access the data (directly? What medium?) ❑ No access allowed ❑ Internet – open access ❑ Internet - passworded ❑ Closed network ❑ Transfer on request via other electronic media	How do you ask for permission to get a password? (criteria? Email?) Who manages the network? If multi-institutional – how does it operate? Are there security issues for the data on transfer?
Q41 Who else might be interested in using it? ❑ Researchers in same discipline ❑ Researchers in related disciplines ❑ Undergraduate students ❑ Postgraduate students ❑ Government ❑ General public ❑ Other	
Q42 Do they currently have access? ❑ Yes ❑ No	
Q43 How might it be in the future? ❑ Open access likely ❑ Could have with permission ❑ No future access	Q43A What IP issues may limit these goals?
Q44 Who funded the creation of this data? ❑ ARC grant ❑ Departmental/University funds ❑ External grant	
Q45 Is there funding to sustain the data? ❑ Yes – built into existing funding ❑ Yes – separate (where from?) ❑ No	
Q46 Who do you think should have future responsibility for the long-term sustainability of your data? ❑ Institutional repository ❑ Local area ❑ Self ❑ Other	

5.2 Data Process Classifiers



This diagram seeks to set out a structure for the *classification* of data processes, i.e. a reasonably standard set of terms that can be used for comparisons and analysis of common requirements. These in turn could lead to the identification of appropriate standards for various processes, and the development of best-practice guidelines. It could be used by end-users to map out their processes, by service providers and educators to elicit information from data-related activities, and by service providers to identify coverage of support activities. The core terms are within the shapes; the additional terms around the shapes provide examples or clarifications