AUSTRALIAN NATIONAL DATA SERVICE (ANDS)
Australian Research Data Commons
Education Investment Fund (EIF)

Annual Report 2, 30 September 2011
# TABLE OF CONTENTS

1# PROJECT STATUS.................................................................................................................. 4#
1.1# Background .......................................................................................................................... 4#
1.2# Major Activities, Breakthroughs, Highlights, and Issues ..................................................... 6#

2# ACTIVITIES UNDERTAKEN.................................................................................................... 10#
2.1# Research Infrastructure ...................................................................................................... 10#
2.2# Data Capture ....................................................................................................................... 10#
2.3# Metadata Stores ................................................................................................................... 20#
2.4# Public Sector Data ............................................................................................................... 23#
2.5# ARDC Core Infrastructure .................................................................................................... 27#
2.6# ARDC Applications ............................................................................................................ 40#
2.7# Project Office ..................................................................................................................... 43#
2.8# Promotion .......................................................................................................................... 44#
2.9# Risk Management ............................................................................................................... 46#

3# PROGRESS AGAINST MILESTONES............................................................................... 48#
3.1# Data Capture Infrastructure ............................................................................................... 48#
3.2# Metadata Store Infrastructure ............................................................................................ 49#
3.3# Public Sector Data Infrastructure ....................................................................................... 50#
3.4# ARDC Core Infrastructure .................................................................................................. 50#
3.5# ARDC Application Infrastructure ....................................................................................... 52#

4# DEVIATIONS FROM THE PROJECT PLAN........................................................................ 53#
1 Project Status

1.1 Background

Research data is steadily becoming more voluminous, more complex, and more important. The nature of research is changing. It has become more investigative as it is possible to assemble significant data collections that enable much broader problems to be addressed. Thus it is very important that research data is managed, able to be assembled, connected to other data and used to address problems that may well be different to the reasons for gathering this data in the first place. The Australian Government recognised the significance of data as key infrastructure in supporting research excellence and research innovation and thus established the Australian National Data Service.

The Australian National Data Service has been in operation since January 2009 as part of the NCRIS initiative. Its aim of having more researchers re-using research data more often required establishing partnerships beyond ANDS and this is continuing to increase. In May 2010 the Australian Research Data Commons (ARDC) project was announced as an EIF funded Super Science Initiative and an agreed Project Plan was submitted in June 2009 and accepted in September 2009. Some activity in the NCRIS funded ANDS project was transferred to the ARDC project as a result. Most recently, ANDS has extended its operations beyond June 2011 to June 2013. Each of these changes has had a high impact on the activities of ANDS and the second change has substantially affected the ANDS project as agreed in the 2009-10 Business Plan. As each of these changes has occurred, ANDS has continued to manage the ANDS and ARDC projects together, as they are strongly co-dependent. There has been no significant variation subsequent to this extension, so this report describes activity taking place against the agreed 2010-11 Business Plan.

At the time of the submission of the 2009-10 Business Plan, ANDS had four programs of activity:

- Developing Frameworks – the frameworks that will enable research data producing institutions to capture, manage and share research data;
- Providing Utilities – services that reduce the cost of capture and ease the task of discovery;
- Seeding the Commons – improving local data capture and populating the data commons; and
- Building Capabilities - improving Australia’s capability to manage its research data.

As a result of the ARDC project, the NCRIS ANDS project consolidated to two programs of activity:

- Frameworks and Capabilities – the frameworks that will enable research data producing institutions to capture, manage and share research data; and improving Australia’s capability to manage its research data; and
- Seeding the Commons – improving local data capture and populating the data commons.

The associated ARDC project has five programs of activity:

1. Data Capture – an institutionally based program to automate the capture of data and metadata from instruments (broadly defined) in data intensive research

30 September 2011

ands.org.au
2. **Public Sector Data** – a program of making more public data collections visible and available through the ARDC

3. **Metadata Stores** – an institutionally based program that enables metadata to be stored coherently across an institution that supports data management, publishing, sharing and re-use

4. **ARDC Core Infrastructure** – an ANDS driven program that puts in place the national services that enable research data to be published and discovered (It is an expansion of the Providing Utilities program.)

5. **ARDC Applications** – a program that develops tools and services to support demonstrations of the value of exploiting data in the ARDC.

These programs are complemented and enveloped by the two NCRIS programs, *Seeding the Commons and Frameworks and Capabilities* as shown in Figure 1.

Figure 1: Relationship between Programs

Taken together, the intent of the two investments is:

- To create an “essential meeting place where the Australian path forward for research data management can evolve and where a vision can be achieved.” – *Towards the Australian Data Commons* (TADC), developed during 2007 by the ANDS Technical Working Group
- Enable the following capability: “Research data and research outputs from all sources can be discovered and reused across disciplines and over time through an integration of repositories and data centres supporting national and specialist discovery services.” – TADC
- Create and populate the Australian Research Data Commons which “will support the discovery of, and access to, research data held in Australian universities, publicly funded research agencies and government organisations for the use of research.” – ARDC Investment plan

The ARDC investment statement can be seen as an intensification of effort in support of the second TADC statement. This report describes progress against this intent.
1.2 Major Activities, Breakthroughs, Highlights, and Issues

As a result of the first few years of ANDS activities many Australian researchers in a wide range of disciplines across Australia’s research institutions are able to:

- capture data, with rich metadata,
- automatically store it,
- simultaneously publish through a discipline portal and Research Data Australia,
- integrate that data with other data discovered through a portal, and
- publish both the results and the data of their investigations.

Consequently, the Australian Research Data Commons (ARDC) has been established.

The ARDC is a combination of the set of shareable Australian research collections, the descriptions of those collections including the information required to support their re-use, the relationships between the various elements involved (the data, the researchers who produced it, the instruments that collected it and the institutions where they work), and the infrastructure needed to enable, populate and support the commons. ANDS does not hold the actual data, but points to the location where the data can be accessed. The ARDC can be envisaged below, where ANDS is contributing to the green pipes and boxes:

![Diagram of Australian Research Data Commons](image)

**Figure 2: Australian Research Data Commons**

In summary, all components of the ARDC exist, but not all components have been established and used at all relevant institutions, and by all relevant researchers. Substantial progress has been made. In Figure 3:
Australian Research Data Commons Progress, we will describe the composite effect of the activity of ANDS to date, including this year.

Secondly ANDS has provided a meeting place that enables research data management to be progressed for the whole of Australia. Beyond establishing this interchange, ANDS also enables a coherent Australian response to the rise of research data, and has enabled all relevant participants to engage. ANDS is engaged with all major research institutions, and importantly they are engaged with each other. ANDS is engaged with major research data collection agencies, such as GeoScience Australia and the Australian Bureau of Statistics. ANDS is engaged with the data intensive infrastructure partners such as the Integrated Marine Observing System, the Australian Telescope, AuScope, and the EBI mirror. ANDS is engaged with funding agencies, principally the ARC and the NHMRC, and ANDS is engaged internationally with national counterparts such as the UK’s JISC, the NSF in the USA and the Netherland’s SURF, as well as international initiatives that are developing between Europe and the USA.

The major activities for ANDS as a whole in the 2010-11 calendar year were:

- Enabling research data collections to be described and harvested by ANDS at 26 research institutions
- Enabling research data collections to be described and harvested by ANDS at 27 research data providers including other research infrastructure providers, public sector providers, cultural institutions and research consortia
- Making 26,746 research data collection pages from 21 collections providers discoverable through Research Data Australia, Google, and other search engines
- Establishing infrastructure to identify, register and publish collections descriptions through Research Data Australia, the ANDS portal into the Australian Research Data Commons
- Helping to establish a coherent approach to research data management at all major research institutions, along with the tools and technologies that enable them to participate in the Australian Research Data Commons
- Concluding the NeAT program of discipline enhancing tools for improved collaboration and exploitation of research data

There have been many highlights over the past year. They can be described as the establishment of national services, the establishment of coherent institutional research data infrastructure, using all of this infrastructure to enable the population of the commons, and enhance the ability of the research system to exploit this improved research data environment.
Establishing national services:

- Enhanced ANDS data collections registration service,
- Enhanced ANDS collection description publication service,
- Enhanced Research Data Australia – a data collections discovery service,
- A dataset identification service established, and
- Party Identifier Infrastructure service run by the NLA enabling richer collections descriptions in the commons

Helping to develop coherent institutional research data infrastructure:

- Tools have been deployed to automatically capture rich metadata along with the data for a wide range of instruments
- 9 institutions operate enterprise-supported metadata stores with 15 more under development
- 4 institutions have operational research data management plans and with 15 more under development
- 30 institutions have installed services that enable collections information to be harvested

Populating the Commons:

- 26,746 collections were described and available through Research Data Australia by June 30th, 2011
- Over thirty research data providing institutions were engaged through direct projects and our partnerships with AuScope and the Museum Metadata Exchange projects,
- 22 institutions will provide collections descriptions feeds to ANDS, both research institutions and public sector data holders, and
- 2 discipline oriented portals were cross connected to Research Data Australia.

Enhanced ability to exploit the research data environment:

- The majority of the tools developed in NeAT projects concluded an delivered benefit to researchers in the relevant domains
- Improved licensing regime supported through AusGoal enabling simpler use and data integration
- International initiatives to ensure compatible approaches are adopted to research data internationally

During this substantial set of activities ANDS confronted a number of issues that it had to deal with and learn from:

- The major challenge facing ANDS was to enable research institutions to develop and exploit their research data ambitions whilst meeting the goals of Seeding the Commons and Data Capture programs.
The pace of engagement was slower than planned, but appears to be delivering stronger outcomes as a result of deeper engagement.

ANDS has tested its national approach to collections that augments disciplinary approaches internationally. This approach has found to be internationally valuable but able to interact with other jurisdictions that take other approaches.

ANDS decided that in order to have researchers change their approach to research data and its management, the optimal path was through the research institutions. This approach has continued to be strongly supported by the institutions, and not resisted by researchers.

The ANDS partner agreement and steering committee was shown to be resilient to a significantly changing environment.

To now summarise the progress of the establishment of the ARDC, we show the many components of the ARDC that have been completed by 30 June 2011 or are currently under development at that time (the number in brackets). This emphasises just how much work is being conducted at and by our institutional partners and their e-research providers.
2 Activities Undertaken

2.1 Research Infrastructure

ANDS has continued to make progress towards its goals of providing greater support to enable researchers to work in the new world of data-intensive research, notably through the substantial increase of effort in describing data collections and making the descriptions automatically visible through Research Data Australia, the increased use of specific services (Identify My Data and Register My Data), roadshows and “boot camps” to improve the capability of institutions to manage and share their research data, engagement with specific institutions to better support their data management, and the provision of advice and documentation in various data management areas, via the ANDS website. More detailed reports on progress in this area are contained in section 10.2. The effort on infrastructure development can be seen in Figure 2 where the dark green pipes and green boxes show the infrastructure being created in the ARDC project.

Complementary infrastructure established as part of the NCRIS ANDS project is described in a separate progress report. Some of this infrastructure was to be established in the ANDS project, but as a result of the ARDC project plan and a modified ANDS Business Plan, the work was undertaken under the ARDC project.

The infrastructure is being constructed through 5 programs:

- Data Capture is constructing the pipes that connect data sources to the data stores and the metadata stores
- Public Sector Data is connecting data held in public sector agencies to the commons either from their data and metadata stores to the ANDS portal
- The Metadata Stores program is creating a set of metadata store solutions that can be deployed at research institutions
- The ARDC Core program is creating the infrastructure that enables collections to be identified, harvested and discovered through the ANDS portal
- The ARDC Applications program is designed to enable researchers to exploit the whole of the ARDC infrastructure to get new value from existing data.

The next section describes the specific research infrastructure created in the 2010-11 financial year.

2.2 Data Capture

2.2.1 Overview of program

The Data Capture program aims to simplify the process of researchers routinely capturing data and rich metadata as close as possible to the point of creation, and depositing these data and metadata into well-managed stores. Metadata will need to be held at both collection and object level in order to support re-use.
The Data Capture program will achieve this aim by augmenting and adapting existing data creation and capture infrastructure commonly used by Australian researchers and research institutions to ensure that the data creation and data capture phases of research are fully integrated so as to enable effective ingestion into the Research Data and Metadata Stores at the institution or elsewhere. This integration will make it easier for researchers to contribute data to the ARDC directly from the lab, instrument, fieldwork site, etc. It will also ensure that higher quality metadata (critical for re-use and discovery) is produced through automated and semi-automated systems. The approach taken will be to partner with leading research groups and Super Science initiatives to augment or adapt data creation and capture systems.

The resulting infrastructure components will include software to integrate tightly with the experimental environment of the researcher to take the data that is being captured/created, and augment this with metadata that describes the setting within which the data is being captured/created, as well as other relevant details (where available) about the research project, researcher, experiment, sample, analysis and instrument calibration details. ANDS will also adopt/adapt/develop software to facilitate automatic/semi-automatic deposit from instruments into data stores/repositories.

The Data Capture program was originally allocated $12M in the EIF ARDC Draft Project Plan. Following the process of public consultation around this Draft Project Plan, this amount was increased to $18.47M. The consultation process also validated the decision to take an institutional approach in allocating the bulk of the available funds. An analysis of research intensity for the major Australian research-producing institutions was undertaken in late 2009 based on the most recent publicly available data on research productivity, and $11.6M of Data Capture funds was allocated in bands of $1M, $500K, or $200K. In late 2009 institutions were each sent an individual invitation to take part in an Expression of Interest process.

### 2.2.2 Outline of projects

The intent of the Fast Start activities (further details for these can be found at Appendix 9.2.1) was twofold: to start expending the allocated funds (which at the time had to be expended by the end of 2010-11), thus smoothing somewhat the expenditure curve; and to quickly undertake a range of activities from which ANDS could learn and thereby fine-tune the process of expending the remainder of the Super Science funding. Given the short timeframe, ANDS sought and received approval from DIISR to select a small number of institutions where discussions were already underway and relationships had already been established.

For Data Capture, this involved the following institutions:

- ANSTO
- Australian Synchrotron
- CSIRO
- ANU
- Monash University
- University of Sydney
- University of NSW
- University of Melbourne
- University of Queensland
Details of these projects were all agreed by late 2010, and they are incorporated into the project table below.

The following NeAT projects (further details for these can be found at Appendix 10.2.2) were funded from the Data Capture program in 2010-11:

- Australian Node of the Human Variome Project
- Auscover Workflow
- Aus-e-Stage
- DataMiNX
- PODD
- Remote Tomography

An analysis of research intensity for the major Australian research-producing institutions was undertaken in late 2009 based on the most recent publicly available data on research productivity, and $11.6M of Data Capture funds were allocated in bands of $1M, $500K, or $200K. In late 2009 institutions were each sent an individual invitation to take part in an Expression of Interest (EOI) process.

At June 30, 2011, ANDS had either entered into contracts (or had substantially agreed on project descriptions) for at least one Data Capture project at all of the EOI institutions except the University of Western Sydney and University of Technology, Sydney, (with some additional projects at selected institutions still under discussion). A breakdown of the progress made in relation to this is provided below. As reported last year, this has been a considerably slower process than anticipated, and while improved processes within ANDS have improved the situation, circumstances beyond our control (staff changes and workloads at partners and legal issues) have meant that a small number of projects are still to be agreed, and that some started later than expected.

Of the twenty four institutions that ANDS engaged with through this process, by June 30th 2011, the majority had agreed a project description with ANDS, and work was underway. The next table indicates the status of the projects at this time. A description of the projects underway in the reporting period is in section 10.2.3.

<table>
<thead>
<tr>
<th>Institution</th>
<th>Project Title</th>
<th>Project agreed</th>
<th>Contracted</th>
<th>Underway</th>
<th>Completed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australian National University</td>
<td>Earth Sciences</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Optical Astronomy (Skymapper)</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Phenomics</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Humanities and allied disciplines</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANSTO</td>
<td>Metadata management for neutron beam instrument data - a joint project with Australian Synchrotron</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

30 September 2011

ands.org.au
<table>
<thead>
<tr>
<th>Institution</th>
<th>Description</th>
<th>X</th>
<th>X</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australian Synchrotron</td>
<td>Meta Data Capture and Storage for the Three Mature Beamlines at the Australian Synchrotron - a joint project with ANSTO.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSIRO</td>
<td>A series of institutional data capture activities not yet determined but under discussion</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Australian Synchrotron</td>
<td>ANDS-CSIRO-ATNF Pulsar Data Management Project</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Curtin University of Technology</td>
<td>Curtin deployment and configuration of Institutional Metadata Repository and Research Data Portal</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deakin University</td>
<td>Filtration Membrane Fouling Data Collection for Water Treatment Research</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Flinders University</td>
<td>Automated measurement of the responses of wildlife populations to climate change</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Griffith University</td>
<td>Smart Water</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Griffith University</td>
<td>Adult Stem Cell &amp; Neurobiological Microscopy Instrumentation and Research Data Management</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>James Cook University</td>
<td>Tropical Data Hub</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>La Trobe University</td>
<td>CMSS RLI Metadata Capture and Publication</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Macquarie University</td>
<td>Glycomics Repository</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Macquarie University</td>
<td>Papyri Data Capture</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monash University</td>
<td>Research Data Management of the Monash Weather &amp; Climate Program (Climate and Weather)</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Monash University</td>
<td>Biomedical Data Platform (Molecular Biology)</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Monash University</td>
<td>Tools for curating and publishing research data in the form of media collections (Multimedia Collections &amp; ARROW)</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>University of Technology</td>
<td>Greenhouse Gas Emissions from Australian Soils</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-----------------------------------------------</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>University of Adelaide</td>
<td>Genomics Data Capture</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Automated capture and publishing of data generated on high throughput plant phenomic platforms.</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>University of Melbourne</td>
<td>Melbourne Neuropsychiatry Centre (MNC) Bioinformatics Development Project</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Youth Research Centre’s Life Patterns Project: Longitudinal qualitative and quantitative survey data capture and reuse</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Video data in the Social Sciences. Optimising Metadata Capture, Data Sharing Procedures and Long-term Reuse</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Federated Neuroimaging Collections in the National Data Commons</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Humanities and Social Science Research Data at the University of Melbourne</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Capture of Complex Data to Support Clinical Research in Cardiovascular and Neurological Medicine</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Founders and Survivors Project</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Institution</td>
<td>Project Description</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>-------------</td>
<td>---------------------</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>University of New South Wales</td>
<td>Enhanced Metadata Capture for Sustainable Management, Sharing and Re-use of APN Histopathology Research Data</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>ARDC Linked International Glycomics Repository &amp; Instrument Data Capture</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>An international antibiotic-resistance gene cassette database</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>ANZNN Neonatal Data Capture Portal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Data capture and integration across multiple platforms</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Managing and Sharing Genomic Data</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>University of Newcastle</td>
<td>Data Capture for the Data Commons</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>University of Queensland</td>
<td>Spatially Integrated Social Science</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Microscopy/Microanalysis Image and Data Repository</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DIMER Diffraction Image Repository</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Aquatic Species Tracking Repository</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>3D Anthropological and Archeological Collection Repository</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>The Health-e-Reef Project</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Linking the EMBL Australia EBI Mirror with the Australian Research Data Commons</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>University of South Australia</td>
<td>Development And Testing Of A Data Capture Tool For Social Datasets Being Used For Record Linkage</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>University of Sydney</td>
<td>SKAMP Data Capture: astronomy</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NSW TARDIS Node</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>AgDataCapt: Capturing Agricultural Data</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AMMRF Live Cell Microscope Data Capture</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Metadata Store/Aggregator</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FieldHelper: a workflow and tools for improving fieldwork data collection and submission to institutional repositories</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
### 2.2.3 Activity/Deliverables for 2010-11

The following projects were completed during the reporting period. Software and other deliverables are made available from the ANDS projects registry: [http://projects.ands.org.au/](http://projects.ands.org.au/). Where no institution is named the project was completed as part of the NeAT program. As noted in the table in section 2.2.2 there were 50 projects underway or completed during the reporting period, with an additional 13 agreed and expected to begin during the next reporting period.

<table>
<thead>
<tr>
<th>Project</th>
<th>Project Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DTS: Data Transfer Service</td>
<td>This project worked with ARCS Data Services team to develop a general-purpose data transfer service with initial deployments at NCRIS Characterisation facilities.</td>
</tr>
<tr>
<td>Human Variome: Software and Data Support for the Australian Node of the Human Variome Project</td>
<td>The Human Variome project is creating a national data repository called the Australian Human Variome Database (AHVD). The database will hold and provide access to information on genetic variations associated with human disease that have been characterised by Australian laboratories and clinics. The project will develop services to enable submission of laboratory and clinic data to the AHVD using existing organisational workflows.</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>BioFlows: Bioinformatics Workflows</td>
<td>The Bioflows project is providing a simple Web-based workflow tool that enables life sciences researchers to specify genomics and proteomics workflows that can be executed on the ARCS Compute Cloud and interface with the ARCS Data Fabric. The system is deployable as an “appliance,” with required software, middleware and server hardware able to be installed at a site and managed remotely, if required. The appliance can interface with local high performance computing systems and/or submit compute jobs to the ARCS Compute Cloud. The appliance concept is being tested with trial deployments at the Bioinformatics Facility at Murdoch University, the Queensland Facility for Advanced Bioinformatics and the Life Sciences Computation Centre in Victoria.</td>
</tr>
<tr>
<td>Aus-e-Stage: Collective Intelligence and Collaborative Visualisation for Creative eResearch</td>
<td>The Aus-e-Stage project is developing two new visually interactive services for exploring information in the AusStage database. It is also creating the capability to generate a new data set of immediate, on-location responses from spectators of Australian performing arts.</td>
</tr>
<tr>
<td>Biosecurity: Biosecurity Collaboration Platform</td>
<td>This project has implemented a collaboration platform at the CSIRO’s Australia Animal Health Laboratory (AAHL) facility that comprises two nodes – one on each side of the bio-containment barrier. This will greatly assist in the flow of complex information across the containment barrier from a variety of data sources including pathology and microscopy systems, live in-vivo animal experimental data (e.g. heart rates) and data from simulation models and historical information in both visual and written form. This platform is expected to have broader applicability within the NCRIS Australian Biosecurity Information Network.</td>
</tr>
<tr>
<td>PODD: Phenomics Ontology Driven Data Management</td>
<td>The Integrated Biological Sciences component of NCRIS contains two major Phenomics initiatives: the Australian Plant Phenomics Facility and the Australian Phenomics Network. These facilities have common requirements to gather and annotate data from both high and low throughput phenotyping devices. The PODD project is delivering a data management service that can handle multiple phenotyping platforms and data formats (text, image, video).</td>
</tr>
<tr>
<td>Remote CT: Remote Computed Tomography Reconstruction, Simulation and Visualisation</td>
<td>The Remote CT project is developing a three-part service for 3D reconstruction and visualisation of Computed Tomography (CT) images. The service will be deployed at the Imaging and Medical Beamline at the Australian Synchrotron and the ANU micro-CT facility.</td>
</tr>
<tr>
<td>AusCover Workflow: Workflow Services to Enable a Large-Scale Temporal-Spatial Ecosystem Digital Information Service</td>
<td>AusCover is a component of the National Collaborative Research Infrastructure Strategy (NCRIS) Terrestrial Ecosystems Research Network capability. It focuses on organising remote sensing data sources and products for terrestrial ecosystems research. AusCover will enable – for the first time in Australia – the online storage of data sets in a form that makes them directly accessible to the user community. The AusCover Workflow project is providing easy-to-use workflow tools and services that enable researchers to process AusCover data sets using the ARCS Cloud Computing infrastructure. The same workflow tools also will allow AusCover data providers to more easily process raw satellite data to generate derived data products in the standard formats users require.</td>
</tr>
<tr>
<td>ATNF Pulsar Data Management Project (CSIRO)</td>
<td>The ANDS-CSIRO-ATNF Pulsar Data Management Project enables the discovery of, assessment of and access to Pulsar data observed at the Parkes Telescope.</td>
</tr>
<tr>
<td>Glycomics Repository (University of NSW)</td>
<td>This project advanced the state of data capture and management within the glycomics discipline by installing new repositories, create an automated data and meta-data capture system at APAF, with an OAI-PMH feed will be created to allow Research Data Australia (RDA) to harvest details of the collections stored within the GlycoSuiteDB repository. This is a joint project with Macquarie University, with the Macquarie section still to be completed.</td>
</tr>
<tr>
<td>Enhanced Metadata Capture for Sustainable Management, Sharing and Re-use of APN Histopathology Research Data (University of Melbourne)</td>
<td>This project helped establish a national database of mouse pathology to enhance the utilisation of mouse models of disease by Australian researchers. It enhanced metadata capture facilities for the Histopathology and Organ Pathology Service based at the Department of Anatomy and Cell Biology, The University of Melbourne as part of facilitating the sharing and re-use of mouse pathology research data both now and into the future. The project addressed current metadata scalability and sustainability issues associated with the service in order for the Melbourne Histopathology Service to participate in and contribute to emerging research data networks like PODD and ANDS.</td>
</tr>
<tr>
<td>An international antibiotic-resistance gene cassette database (University of NSW)</td>
<td>The project made technologies that allow us to archive relevant elements and to identify them in bacterial DNA sequences. It has also built a knowledge repository of antibiotic resistance gene cassettes available to the wider research community and to allow the community to contribute new entries to the repository as these are found. Providing this application should enhance the team’s global reputation as leading researchers of antibiotics resistance and pioneers in the analysis of larger-than-gene structures. The application makes this research accessible to many research teams that don’t have the necessary computer skills to otherwise use it, thus increasing our impact on the scientific community.</td>
</tr>
</tbody>
</table>
Managing and Sharing Genomic Data (University of NSW) | A new generation of DNA sequencers has recently been installed at UNSW, Southern Cross University and the Australian National University. These instruments can generate DNA sequence data 1000x faster than old technology, and can sequence the genomes of small organisms in a week. This project established databases for this DNA sequence information, so that users of the DNA sequencers can access their information in an efficient way. This will centralise DNA sequence data from these facilities, enable collaborative projects and facilitate data sharing. On publishing, research data and associated metadata can be made available for public use, contributing to the Australian Research Data Commons (ARDC).

The Health-e-Reef Project (University of Queensland) | This project is developing the data capture and sharing services for coral-reef related data being generated by researchers at the University of Qld Centre for Marine Studies, together with their collaborators in community/volunteer groups (CoralWatch, ReefCheck) and government organizations (EPA, DERM). Together these researchers are monitoring and studying the impact of climate change and human activities on coral reef ecosystems. There will be particular focus on automatically capturing the metadata necessary to support discovery, decision, and reuse. The discovery metadata will be made available through RDA.

Clarke eHealth (Early Activity): Capture, management, re-use and discovery of breast cancer microscopy virtual images (University of Sydney) | This project constructed software to allow wide use and re-use of microscopy images in breast cancer research, generated at the Westmead Institute for Cancer Research (WICR), in analysis and study by approved researchers across Australia, with collection descriptions available through RDA.

### 2.2.4 Program Highlights, Issues and Breakthroughs

Particular highlights of the Data Capture program so far have been the level of enthusiasm by many institutions to improve their data capture infrastructure, as demonstrated by the number of institutions that were ready to engage, and the number of proposals submitted by some institutions, as well as the interest by a range of players in putting forward proposals.

The completed ATNF and Antibiotic Gene Cassette projects noted above have both attracted the interest of overseas researchers, who have engaged with the data now made available to further their own research.

As far as the NeAT projects that are now part of the Data Capture program are concerned, particular standouts are:

- Phenomics Ontology Driven Database (PODD): this is progressing particularly well under strong project leadership and also interacting well with other ANDS-funded related activity at the University of Melbourne.
- Australian Node of the Human Variome Project (HVP): this is enthusiastically supported by a range of players in the research and diagnostic community and appears set to demonstrate the value of an integrated approach across genomics, phenomics and diagnostics.

### 2.2.5 Program Learnings

It had been expected that the metadata requirements for adding records to Research Data Australia were relatively straightforward and would not pose many delays, especially if metadata experts such as librarians were involved. However it has become clear that despite ANDS having offered some training and provided extensive documentation, many of our partners were struggling with the concept. ANDS also had limited in-house expertise to advise on and assess records. This was further complicated by changes to our standards as we learnt more, and by the introduction of new services, such as the People Identifier services.

**Learning:** ANDS needed to provide more advisors and assessors, which was done in the first half of 2011, and offer more training, which is being done as a combination of improved documentation and more sessions. Communication around changes needed to be improved as well, which has occurred.

Overall, everything took longer than expected: discussions on projects took time to converge on a set of activities that met the institution’s desires and ANDS’ requirements, negotiations between institutional legal staff often took time, even with standardised contracts, and the finance processes of the lead agent introduced delays of between one and two months before funds flowed to the institutions. Once funds were in place recruitment was often required at partners, which was often difficult, and in some cases the recruitment was too focused on software needs, rather than the requirements of describing data collections. Assessment of deliverables by ANDS was too slow for many partners, and impacted on payments.

**Learning:** Despite many improvements to internal processes many projects could only be agreed through extensive face to face contact to ensure that the partner gave it their full attention. More resources were added to ANDS for metadata assessment, and the assessment process in general was reviewed and improved. The addition of a dedicated contracts manager to the team greatly improved the contracting and payment processes.

### 2.3 Metadata Stores

#### 2.3.1 Overview of program

Information that can be held about data can be grouped into four categories. The first is *information for discovery*, and is primarily held at the level of a collection. This consists of the range of pieces of information that will assist in the discovery of the collection. The second is *information for determination of value* (also primarily at collection level). This includes information such as the name of the researcher, institution or funding program that might help a potential user to decide whether they want to access the data. The third is *information for access* that might be a direct link to the data objects (stored elsewhere, such as on national and institutional data stores), both at collection and possibly object level, or contact details for
where to source the data. The fourth kind of information is *information for re-use*, and will include things like reading scales, field names, variables, or calibration settings that are needed in order to effectively re-use the data. This will mostly be at object level.

The metadata store infrastructure that will be created through this program needs to manage these kinds of information about data collections and data objects, as well as information about associated entities. These include parties (both people and organisations), activities (that produce the data) and services (associated with the capture of, and access to, the data). These associated entities serve as part of the rich context for the data collections, and also contribute to the information for discovery and information for determination of value.

The software will need to support a range of functions. The first is the creation and management of these kinds of information, or their harvesting from other sources (research management systems, human resources systems, finance systems). In addition, the software will need to manage the relationships between the information about data collections/objects and the data collections/objects themselves. The software will also need to support queries over the data by users within the institution. Finally, the software will need to be harvestable by ANDS services, as well as by other organisations.

This program is therefore developing, and will assist to configure and make available, this metadata infrastructure at research producing institutions. Institutions with existing suitable software can also have this configured or supplemented to the required standard if they choose to; this program will assist with this.

A common pattern that ANDS is seeing is a single institution metadata store for collection information (often as part of a Seeding the Commons project) and a number of metadata stores for object information, each associated with a particular instrument (often as part of a Data Capture project).

### 2.3.2 Outline of projects

All of the agreed 2010-11 activities were originally funded through the Early Activity component of the EIF ARDC Project Plan. These early activities deliberately targeted a range of approaches, dealing with different subsets of information about the entire range of ISO 2146 entities (Collections, Parties, Activities, Services). The first approach was to investigate three different candidate solutions for standalone metadata stores: one based on the “myTARDIS” codebase to integrate with an existing large research data store (Monash University), another developed as an adjunct to an institutional repository (USQ/ADFI), and a third building on the RDF triple store-based VITRO code developed at Cornell University (University of Melbourne). To complement these, ANDS also funded work to implement feeds of parties and activities information to ANDS from corporate systems (Monash University), and a combined Research Metadata Stores Hub deployable across two institutions with very different infrastructure (joint activity between QUT and Griffith).

In addition to these projects that were focussed specifically on metadata management, the joint ANSTO/Australian Synchrotron project also had a strong focus on managing the metadata being produced by a range of instruments attached to various sources of imaging energy. As flagged in the business plan, no new metadata stores projects were commenced in 2010-11.
2.3.3 Activity/Deliverables for 2010-11

The Ingec/ReDBoX solution was on track to complete within time and budget when the Australian Digital Futures Institute decided to shift its focus. As a result, all of the development team had their contracts terminated. In response to this, ANDS contracted QCIF to support the ReDBoX solution for the next year. This met a number of goals: it ensured successful completion of the project, it meant that there were developers available to fix bugs arising from the Newcastle deployment, and it provided a support infrastructure for the software to encourage adoption. Without this, it is unlikely that any universities other than Newcastle would have considered adopting ReDBoX. The intention is that by mid 2012 enough institutions will have deployed ReDBoX and taken out a maintenance contract with QCIF to sustainably fund its ongoing support. The original ReDBoX project has now completed.

The QUT/Griffith Research Metadata Hub project(s), which selected VITRO, have now completed successfully, and both are in production. In addition, Griffith has decided to invest in further development of their solution.

The solution being developed at Monash University to provide a feed of activities and collections is about to go into production, after a period of testing. The project is essentially complete.

The solution being developed at Monash University to provide a metadata store associated with a large data store is scheduled to complete in September 2011.

The ANSTO/Australian Synchrotron project has not yet completed. A combination of personal issues for the project manager and a number of resignations from the development team have caused delays. The project is now scheduled to complete by the end of 2011. Working demo software is already installed at both sites, and the organisations have repeated their commitment to adopting the solution.

It should be noted that in addition to the above projects, there has also been a significant amount of metadata store activity that has been funded/cross-subsidised by projects under the Seeding the Commons and Data Capture programs. This is because those projects needed a solution before the Metadata Stores projects were ready.

2.3.4 Program Highlights, Issues and Breakthroughs

One of the highlights of the program was the completion (or near completion) of all of the funded projects, and the use of some of them in production. In addition, there has been an increasing awareness across all the institutions with whom ANDS has been engaging of the importance of metadata about research data outputs, and the need to make an institutional commitment

The major issue in this program has been determining the next steps to take after the initial round of funding, and the best way to implement the decisions made. After a series of iterations of internal strategy documents, it has been decided to undertake a conversation with senior university stakeholders from the Library, IT, Research Office and e-Research (if this exists) to identify research metadata ambitions, and how best to meet these. Any decision about ANDS-funded development or deployment will then be as a consequence of this prior discussion about institutional priorities, rather than the need for a decision driving the discussion.
2.3.5 Program Learning

The installation of a metadata store that meets ANDS’ requirements to provide feeds to the ANDS Collections Registry, and the integration of this store into a range of institutional information sources, is a non-trivial process. There are challenges in having this work occur and also in having the solution continue once the ANDS funding is finished.

**Learning:** In order for an institutional metadata store (of the form described above) to be sustainable, it must be seen as a core part of the institution’s infrastructure, and must deliver value to the institution in its own right. Any decisions about further funding of development or deployment of possible solutions must reflect this reality.

2.4 Public Sector Data

2.4.1 Overview of program

Many areas of research are heavily dependent on government data – from cadastral data to economic data to government organised surveys – or could increase their use of such data if it were more widely discoverable and accessible. The responsibilities inherent in data custody are a shared challenge and include the need to address preservation, access and description. As such there is a very close potential relationship between ANDS’ concerns and those government agencies that are custodians of data or that are influential in data policy.

The Public Sector Data program will provide the infrastructure necessary to ensure that feeds of data collection descriptions are made available from a range of public sector agencies. Identified agencies include producers of research data, such as the Bureau of Meteorology (BOM), the Australian Bureau of Statistics (ABS), GeoScience Australia (GA), CSIRO and Departments of Primary Industry (DPI). Owners of data gathering activities and collections, such as the museum and library sectors, which might be possible inputs to other research activities are also in scope. ANDS also needs to maintain and develop stronger relationships with other organisations with significant data holdings or interest in these areas such as the National Archives Australia (NAA) and the Australian Government Information Management Office (AGIMO), for example. Finally, ANDS will explore ways to incorporate public data collected by citizens, through exemplar projects.

The key deliverable from this program is to make existing public sector data resources more discoverable to the research community and to work with federal, state and territory government agencies to improve access to data. Activities will vary across agencies according to their existing infrastructure and the types of data being made available. In all cases there will be a strong preference to have data services exposed using relevant international standards.

The Public Sector Data program was originally allocated a $10m budget in the ARDC Draft Project Plan. During the review phases for ANDS mid 2010 this budget was reduced to $6.45m. This was as a result of discussions with key government agencies in the first quarter of 2010 where they indentified that their
desire was for capability assisted infrastructure development from ANDS in preference to the provision of funding to undertake the infrastructure development themselves.

2.4.2 Outline of projects

ANDS has either entered into contracts for, or in negotiation for the following Public Sector Data projects and engagements:

<table>
<thead>
<tr>
<th>Agency or Institution or Project</th>
<th>Project Status and Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EIF003 – Auscope SISS Deployment</td>
<td>Extension granted – continuing relationship with VeRSI and BoM. AuScope Discovery Portal AuScope in RDA</td>
</tr>
<tr>
<td>EIF021 – CSIRO Water Resources Observation Network</td>
<td>Complete – 5 major Murray Darling Basin Sustainable Yields (MDBSY) collections exposed and software deliverables for build of CSIRO Data Access Portal. MDBSY in RDA</td>
</tr>
<tr>
<td>EIF023 – AODN Data from National Research Vessels</td>
<td>Complete – over 7000 collections exposed via RDA AODN portal Underway in RDA</td>
</tr>
<tr>
<td>EIF024 – Australian Legal Information Institute (AustLII)</td>
<td>Extension granted – 400 collections exposed. Extension to garner further collections AustLII portal AustLII in RDA</td>
</tr>
<tr>
<td>EIF041 - PowerHouse Museum (CAMD/MA)</td>
<td>In final stages – over 1000 collections from 18 Australian museums in test environment. Museum Metadata Exchange</td>
</tr>
<tr>
<td>EIF042 - National Archives of Australia</td>
<td>Abandoned – NAA have not indicated an interest in pursuing this further.</td>
</tr>
<tr>
<td>National Criminal Justice Research Data Network</td>
<td>ASSDA contracted to host data and ingest is currently underway NCJRDN portal</td>
</tr>
<tr>
<td>GeoSciences Australia</td>
<td>Engagement in final stages. On site component complete.</td>
</tr>
<tr>
<td>Australian Institute of Health and Welfare</td>
<td>Transitioned from project proposal to engagement and at the stage of developing a statement of work</td>
</tr>
<tr>
<td>Australian Bureau of Statistics</td>
<td>Engagement – statement of work presented to ABS and currently under consideration.</td>
</tr>
<tr>
<td>Bureau of Meteorology</td>
<td>Engagement here is largely through the EIF003 project, by their wishes.</td>
</tr>
<tr>
<td>Department of Environment Water Heritage and the Arts</td>
<td>Engagement postponed – agency not ready</td>
</tr>
<tr>
<td>Australian Sports Commission</td>
<td>Engagement postponed - agency not ready</td>
</tr>
</tbody>
</table>

2.4.3 Activity/Deliverables for 2010-11

The majority of planned deployments under the ASRDC project are finalising, in line with the project plan. SISS technology has been installed at the National Computational Infrastructure (NCI) to provide interoperable access to GeoScience Australia’s (GA) satellite imagery data. Through the NCI, a Virtual
Geophysics Laboratory has also been set up and “Cloud-enabled”. This virtual laboratory allows researchers to submit processing jobs using large volume data. Other GA deployments of SISS are complete – Geochronology Raging Spot; Borehole Support; Geotranssects. The AuScope catalogue has also been harvested by Research Data Australia. The ASRDC project has established a relationship with the Victorian eResearch Strategic Initiative (VeRSI), who will become a “SISS Node”. VeRSI will focus on the Groundwater community in Victoria which includes – DPI Vic, DSE Vic, and University of Ballarat. Delays were experienced in obtaining a suitable unencumbered water dataset from Bureau of Meteorology (BOM). This delay has now been overcome and a Water information model developed. The ASRDC project will be extended to June 2012 to focus mainly on the BOM engagement, in particular Geofabric datasets and BOM’s national register of products and services.

CSIRO completed a project to capture data from the Water Resources Observation Network and Murray-Darling Basin Sustainable Yields data was released into the ARDC as an outcome. The project has a number of IT deliverables and was the first of two ANDS funded projects that incrementally commenced the build of the CSIRO Data Access Portal. This is an enterprise tool to facilitate the capture, management and release of CSIRO research data for re-use.

The Underway data project was also completed during this period. It delivered data from the research vessels: Southern Surveyor and Aurora Australis, via the AODN portal. The infrastructure established also allowed AODN to leverage and deliver all of its metadata records made available via its portal, exposing data from University of Tasmania, Australian Institute of Marine Science, Australian Antarctic Division, CSIRO, BoM and Royal Australian Navy. The aggregated nature of this data raises some issues associated with data presentation and connections that present a use case for ANDS to develop best practice around the presentation of data in Research Data Australia.

AustLII completed the requirements of their contract in June, delivering over 400 records into RDA and establishing an automated feed. They have requested a contract extension to December 2011 to pursue further datasets under the aegis of an ANDS funded project.

The team at Powerhouse Museum is in the final stages of the Museum Metadata Exchange project. Over 1000 records from 18 museums around Australia are poised for release into the ARDC via RDA. This project will allow a significant volume of museums data to be exposed electronically for the first time. It also delivers a discipline based portal: Museum Metadata Exchange for humanities and social science researchers. Other deliverables include the development of a vocabulary to enhance the discoverability of the data and the development and implementation of Collection Level Descriptions across the sector. This latter is the first time that the sector has taken this approach to describing its data and a module being developed for the collection management systems will aid the cultural change being effected. Plans are currently underway to investigate options for extending this initiative in the GLAM (Galleries, Libraries, Archives & Museums) sector.

Of the two activities that were under negotiation, the National Archives of Australia approach has not resulted in a project. There may be other opportunities to expose this data through other means. The negotiations with GeoScience Australia were completed and the assisted engagement was undertaken. Two ANDS staff spent time on site establishing the infrastructure to support a feed into RDA. Currently there are
7000 records in the final stages of quality assurance prior to being transferred. This engagement was able to leverage off the work the AuScope project (EIF003) was undertaking in its engagement with GA. The next stage of this work is for GA to identify the marine related data for harvesting into AODN, thereby increasing the connections and discoverability.

The GA engagement has been the first assisted engagement and has established a base process for service delivery in this regard. What it has also done has been to bring about the need for an ANDS service package to support the initial negotiations ANDS staff have with agencies.

Further engagements have been initiated with mixed success. As expected, this is influenced by an agency’s resourcing capability and the level of awareness and development of data management within the organisation. PSD has commenced two major engagements. The first is with ABS, and there is currently a ‘statement of work’ which outline the proposed activities, under negotiation. The second is with Australian Institute of Health and Welfare and a statement of work is currently under development.

PSD has also commenced an engagement with Atlas of Living Australia to establish a feed of their collections into RDA.

The development of CIRDI will take the progress of PSD engagements from opportunistic to a more strategic approach and the experience to date will inform development. Work that has commenced in PSD to develop a service package, a process for negotiating an engagement and an environmental scan will now be leveraged to a wider ANDS approach.

Of the NeAT projects in this program, The National Criminal Justice Research Data Network (NCJRDN) has contracted ASSDA to host its data and is in the process of adding data to this repository. See: http://ada.edu.au/crime-and-justice/home.

### 2.4.4 Program Highlights, Issues and Breakthroughs

This period has been marked by the substantial and successful progress on contracted projects. With one exception all are either complete or in their final stages and delivering some significant outcomes and flow on effects.

A highlight has been the outcomes from the CSIRO Water Resources Data Management (WRDM) project. It has set the foundation for a data management, publication and access tool for the enterprise. It is particularly significant given the volume of data generated by the organisation. The software tools are also open source enabling other agencies to deploy them as well.

The Museum Metadata Exchange has established a metadata store and discovery portal that allows some museum data to be visible in a way it has never been. The 18 agencies contributing to the exchange are a cross section of the sector and range from the major museums: Powerhouse, Australian Museum and state museums, to smaller regional museums such as Apollo Bay Museum, Historic Houses Trust and Allport Library and Museum of Fine Arts. It is unlikely that the smaller museums would have had the resources to expose their collections and this project has enabled that. The project has also introduced new practices in the sector around the definition and description of collections. It has also had the incidental benefit of assisting resource sharing within the sector by exposing collections in an aggregated manner.
AODN has broken ground in bringing forward issues associated with establishing a feed of aggregated data. An outcome of this has been the commencement of the development of best practice recommendations around the definition and presentation of collections, activities and portal services. This work will extend beyond AODN to other organisations and could influence work commenced with ALA.

The assisted engagements approach has proved very productive and informed a broader ANDS approach. The first engagement with GA established an approach which, with some fine tuning, can be scaled for use across ANDS. It also raised the need for a service package: an understanding of what ANDS could offer and effectively deliver.

### 2.4.5 Program Learnings

A challenge for engagements will be each agency’s ability to resource data management and the general receptiveness for so doing. On completion of the proposed environmental scan it may be worth investigating some of the synergies that may exist in combined approaches.

Increasingly activities that are developed and conducted from other ANDS programs are relevant across all and there is considerable effectiveness that can be gained from a coherent effort. An ANDS enterprise approach will certainly enhance PSD’s ability to engage.

### 2.5 ARDC Core Infrastructure

#### 2.5.1 Overview of program

Researchers and research organisations are increasingly expected to make their data public. This allows verification of research claims and the building of innovation upon previous work. Similarly, public funders of research are increasingly requiring public access to the inputs and outputs of research. Research assessment frameworks are also moving towards ways of acknowledging the publication of data as a research output.

The ARDC Core program is gradually building out the national infrastructure services that enable research data to be published and discovered through a program of development of software utilities, both in-house and with partners, which enable organisations and individuals to publicise and share the collections they hold.

#### 2.5.2 Outline of projects

**Register My Data**

ANDS Register My Data enables the harvesting of description information about collections and the activities, parties and services (for example, RSS feeds) which relate to collections. This information is exposed on the World Wide Web through Research Data Australia, a mesh of highly findable web pages. Research Data Australia is publicly available online at: [http://services.ands.org.au/pages](http://services.ands.org.au/pages). Register My Data is underpinned by the ANDS Collections Registry.
The ANDS Collections Registry is a software application that holds and manages collection description information and associated context. Generally, ANDS participants set up an automated harvest of collection description information from their repository or data store to the ANDS Collections Registry. This enables the description information to be kept current via updates from the originating source. The ANDS Collections Registry supports a number of dynamic exchange and harvesting protocols to automate communication with managed data environments, including OAI-PMH.

The ANDS Collections Registry accepts automated feeds of collection descriptions in an XML format called Registry Interchange Format – Collections and Services (RIF-CS). RIF-CS is based on the ISO 2146:2010 (Registry Services for Libraries and Related Organisations) standard. ISO 2146 is an international standard intended to operate as a framework for building registry services for libraries and related organisations.

Publish My Data
Publish My Data allows individuals to manually register a collection description and to obtain a persistent identifier for their collection. Collection descriptions are stored in the ANDS Collections Registry and are publicly discoverable through Research Data Australia. To be registered, collections must be accessible online.

Publish My Data is an entry level product which requires a minimum of researcher effort to use. Collection descriptions must include a title, URL and a short description of the collection. Researchers may also add additional information about other contributors, subject keywords, spatial coverage and access rights.

ANDS prefers to harvest collection description information automatically, at the institutional level, as this allows for the responsibility of ongoing maintenance of collection description information to rest with the institution. However, this is not always possible. Publish My Data is intended for use by researchers at organisations where there is no formal data archiving service and where ANDS has no distributed services in place. Effectively an institutional Publish My Data is available to research organisations that use the machine to machine interfaces of both Register My Data and Identify My Data.

Research Data Australia
Research Data Australia is a set of web pages (see Figure 4 for a sample) describing collections produced by or relevant to Australian researchers. It is designed to promote visibility of research data collections in search discovery engines such as Google and Yahoo, and to encourage their re-use. It is the face of the ANDS Register My Data service. Research Data Australia includes discovery tools and spatial coverage display. ANDS will continue to enhance Research Data Australia during 2011.
Identify My Data

Identify My Data provides cost free creation (called minting), resolution and management support for persistent identifiers for the Australian research and cultural collections sectors.

A persistent identifier (PID) is a number or a code that is allocated to an object or a resource as a long lived identifier. Persistent identifiers provide a globally unique identification. Persistent identifiers can be used to create hyperlinks on a web page. When clicked on, the hyperlink will take the web-user to the URL associated with the identifier.

Persistent identifiers are useful because when an object or resource changes location on the internet, the persistent identifier does not change. The owner of the resource is able to communicate with the persistent
identifier database and change the object location associated with the persistent identifier. All links which use the persistent identifier will then point to the new location without the resource owner needing to update them.

Identify My Data can be from either from a web form, or via machine-to-machine transactions. Identify My Data is useful for:

- software developers needing to build software which assigns persistent identifiers to objects within their application where there is a requirement to assign, update and resolve identifiers.
- individuals wishing to provide persistence for citing materials held on their web site who may want to assign identifiers to individual pages, papers, or a web site
- institutions acting as an authoritative source of information about people, organisations or community vocabularies where globally unique identifiers need to be assigned to individual records
- institutional repositories not wishing to manage their own identifier service who may wish to assign persistent identifiers to their holdings.

Identify My Data uses the international Handle System developed by the Corporation for National Research Initiatives (CNRI). The Handle System provides identifier resolution services for use on the Internet. It includes a set of protocols that enable a distributed computer system to store identifiers – known as handles – and to resolve those handles into the information necessary to locate, access, contact, authenticate, or otherwise make use of the associated resources. This allows the handle of an item to persist despite changes of location and other changes. The Handle system is very robust and is widely used internationally among repositories.

**DOI Project**

DOIs form the basis of publication identification in current academic publishing practice. ANDS' aim is to treat ‘published’ datasets in a way analogous to journals and monographs. Citation indexes already use DOI as the basis of counting citations, and discussions with ThomsonReuters and Scopus have confirmed their willingness to deal with DOIs allocated to datasets. DOIs are known and used by researchers, publishers and citation indexes.

Research communities understand DOIs because of their use in journal publications. The experience here and overseas is that data creators and managers are more comfortable using the same persistent identifier infrastructure for published data sets as they have for journal articles. ALA, APN, CSIRO Land and Water, AANRO, ANUSF, and ASSDA have all confirmed this in informal consultation.

The ANDS implementation of DOI will complement the existing ANDS Handle-based persistent identifier services.

ANDS has joined the Datacite international consortium. The goal of this consortium is to establish a not-for-profit agency that enables organisations to register research datasets and assign persistent identifiers to them, so that research datasets can be handled as independent, citable, unique scientific objects. DataCite is a coalition of nationally-focused organisations dedicated to registering and allocating identifiers to scientific datasets.
ANDS has developed prototype minting software for Datacite, and contributed to the development of a minimum metadata set. During 2011 has released a pilot service Cite My Data, which allows Australian eresearch organisations and data centres to allocate citation optimised identifiers (DOI). This is one of the enabling services which will allow data collections to be cited and tracked as first-class objects in the scholarly communications cycle. ANDS will develop this infrastructure and tailor it to research organisations’ needs over the next two years.

**Vocabulary Service Project**

To improve the accuracy and precision of research reporting, most research communities use fixed terminologies, concepts, units of measure and so on. Some, although not all, of these items take the form of controlled vocabularies. Given the importance of controlled vocabularies to science, a national Vocabulary service would benefit the research community by providing a managed infrastructure and support service that would not only make it easier to access and use existing vocabularies and but also provide new opportunities for other communities to formalise their terminologies and bring it under a common framework for their sustainable management and governance.

Controlled vocabularies are widely used in library and information sciences to better organise and describe knowledge. Controlled vocabularies help standardise the use of language in bibliographic and metadata descriptions and enhance the precision of retrieval against known terms and concepts. ANDS has been working to consult the research community on the appropriate development of a Classify My Data service which would provide a set of web and online services to support the creation, management, and publication of human and machine-readable controlled vocabularies for use by the Australian research and higher education sector.

Among many other benefits, the service will provide the research community with a managed infrastructure and support service that will not only make it easier to access and use existing vocabularies, but also provide new opportunities for other communities to formalise their terminologies and bring it under a common framework for their sustainable management and governance.

The first phase of the project will establish infrastructure allowing communities to publish and consume existing vocabularies. The functions offered in this phase include:

- Upload vocabularies (GUI and web services)
- Retire vocabularies
- Update vocabularies (GUI and web services)
- Retrieve vocabularies (via web services)
- Download vocabularies
- Browse vocabularies
- Search vocabularies

The functionality to be provided in subsequent phases is still being investigated.

ANDS has convened a vocabulary services technical working group with input from NCRIS facilities and commonwealth agencies.
**Party Infrastructure Project**

The objective of this project is to improve the discovery of research data and research publications by linking them through common researchers and research groups. ANDS has partnered with the National Library of Australia to provide infrastructure to allow Australian researchers and research organisations to more easily use an independent party identifier when publishing their information about data collections.

![Diagram showing the Party Infrastructure workflows](image)

**Figure 5: Party Infrastructure workflows**

Having a persistent identifier assigned to researchers enables different institutional systems about researchers to interoperate. The institutional repository, research management systems and other systems providing information about researchers can exchange information using a common identifier. That identifier can also be linked to other information external to the institution, such as former and concurrent affiliations and publications as well as cross-institutional collaborations.

Figure 5 shows the workflows for content providers using the NLA Party Infrastructure.

**Location Infrastructure Project**

An important goal of the Australian Research Data Commons is to enable cross-disciplinary discovery of related research data, and spatial location is a vital linkage mechanism in this process. The value of the data commons will be increased if the dataset descriptions include spatial coverage data encoded as geographical points or polygons rather than just text. ANDS' vision for a data commons would see non-GIS-experts from
arts, humanities, and science able to enrich their discipline specific data with standardised spatial information.

Achieving this goal requires the establishment of a robust national infrastructure that would allow place names to be validated by both individuals and software systems against an Australian gazetteer service in an efficient manner. There will need to be distributed sources of gazetteer data, depending on jurisdiction, feature types, temporal coverage and language. A comprehensive national service will need to provide interoperable query services across these sources, irrespective of their differing construction.

This infrastructure is intended to increase the amount and quality of spatially-marked-up research data. This will enable new kinds of research and innovation based on new data linkage and data merging opportunities. The infrastructure aims to unlock significant innovation and productivity. It will bring benefits well beyond the research and innovations sector.

ANDS is partnering with Geoscience Australia (GA) to develop and run an online Gazetteer location service. This project seeks to provide a publicly available interface (both web and machine readable) to an authoritative national gazetteer of place names and other useful spatial information. GA is the national geoscience research and spatial information agency. The Office of Spatial Data Management (OSDM) is encompassed within GA and coordinates the implementation of Australian Government policy on spatial data access.

The national value of this project lies in its ability to enable more spatially enabled data to be produced by any researcher, group, or data repository. GA will develop, deliver and run the Australia Gazetteer service as a national service into the future. GA will also undertake all software development activities leading to the deployment of human usable web services together with machine to machine services.

The specific deliverables will include:

- Development of a gazetteer data schema, database, query and web service
- Development of a user interface with search and display mapping functionality
- Commissioning systems and services

Data from this service will be publicly available at no cost (previously these data were subject to charge).

**Activity Infrastructure Project**

The project will leverage the significant information holdings of the Australian Research Council (ARC) and the National Health and Medical Research Council (NH&MRC) with regard to research activity underway in Australia.

Proposals for projects with the ARC and NH&MRC are currently under development for them to maintain persistent URI identifiers for all of their awarded research grants which will resolve to structured data which can be accessed by other systems as well as by humans. However, this is a longer-term goal which is not likely to reach final implementation until later in the ANDS project.
In the interim, these records have been ingested as Activity records into the ANDS Registry, and be available for linking to collection records by ANDS participants, which will increase and add value to the mesh of navigable research records discoverable through Research Data Australia.

2.5.3 Activity/Deliverables for 2010-11

The initial ANDS utility services were launched into production during the 2009-10:

- Research Data Australia
- Register My Data
- Identify My Data
- Publish My Data

In 2010-11 three iterative updates to these services were released as part of a commitment to continual improvement balanced with operational reliability. This included cycles of usability and unit testing.

In early 2011 the systems began to provide significant support for manual record creation in support of the ANDS Seeding the Commons and Data Capture projects. New automated linking of records is supported as well as phase one of support for bibliographic tools such as Zotero and Endnote in citing collections in Research Data Australia. Initial arrangements for linking data collections to researchers, research groups, and research projects have been established.

Planning and design for a totally new look and feel and back end to the Research Data Australia site is complete and this will be released early in 2011-12.

ANDS continues to contribute to the international consortium DataCite, and leads the DataCite Services working group. In June 2011 ANDS initiated a pilot of the new Cite My Data service which provides citation optimised identifiers (Digital Object Identifiers) for use by Australian e-research facilities and data centres.

The Party Infrastructure project is practically complete and the infrastructure is transitioning to operational status at the National Library of Australia. Linking of datasets to people and organisations is now supported using a public globally unique identifier. A phase of promotion and benefits realisation is planned over the next twelve months through activities in the ANDS NCRIS project.

The Location Infrastructure phase one project with GA and OSP (formerly OSDM) is in testing and acceptance phase and should come through to operation in early 2011-12.

The Vocabulary Support Infrastructure project has been re-scheduled for 2011-12. This has allowed ANDS to consult more widely with the research and public sector players in this field. ANDS has jointly convened with the Office of Spatial Policy (Commonwealth Dept of Resources, Energy, and Tourism) a Vocabulary Services Technical Working Group with membership from NCRIS facilities and Commonwealth govt agencies. The working group is to inform the design of the planned ANDS service. The previous concept, proposal and target business model documentation is still in place for Vocabulary support infrastructure. The Australian Bureau of Statistics has given in principle approval to a joint project to publish the ANZSRC codes through the ANDS vocabulary service.
A service desk, change management system, and reporting functionality have been implemented during 2010-11 as part of the establishment of robust national services and infrastructure.

The RIF-CS Advisory Board has been convened. This board, constituted and independently chaired by partners and members of the community, gives advice to ANDS on changes to the registry interchange format.

On the following pages are the service usage reports for 2010-11.

### ANDS Annual Management Report

#### PUBLIC SYSTEM:

<table>
<thead>
<tr>
<th>Service</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>Handles Minted</td>
<td>4,279</td>
<td>6,323</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trusted SW Agreement</td>
<td>10</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Registry:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Source Admins</td>
<td>12</td>
<td>21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Provider Org</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Publish my Data</td>
<td>51</td>
<td>63</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Records</td>
<td>2,906</td>
<td>26,746</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Collections</td>
<td>1,319</td>
<td>23,899</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parties</td>
<td>894</td>
<td>1,960</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Services</td>
<td>3</td>
<td>30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activities</td>
<td>690</td>
<td>857</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Research Data Australia:

<table>
<thead>
<tr>
<th>Service</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>Page Views</td>
<td>192,212</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Page Hits</td>
<td>-</td>
<td>56,659</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Filtered Page Views</td>
<td>-</td>
<td>37,527</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### TRIAL SYSTEM:

<table>
<thead>
<tr>
<th>Service</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>Handles minted</td>
<td>-</td>
<td>15,399</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trusted SW Agreement</td>
<td>-</td>
<td>30</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Sandbox:

<table>
<thead>
<tr>
<th>Service</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Source Admins</td>
<td>91</td>
<td>128</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Provider Org</td>
<td>60</td>
<td>105</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Records</td>
<td>33,258</td>
<td>85,881</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Collections</td>
<td>3,753</td>
<td>32,416</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Services</td>
<td>29</td>
<td>44</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parties</td>
<td>2,392</td>
<td>26,270</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activities</td>
<td>27,084</td>
<td>27,151</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**NOTES:**

1. “Page Views” are all Research Data Australia page views including hits from crawlers and robots (this is the raw data gathered from the web server logs). This method was used in 2009-10.
2. The Google Analytics tool was used to gather page view reports from July 2010 to June 2011.
3. “Page Hits” are Research Data Australia page views cleansed by the Google Analytics tool (thus excluding robots, crawlers etc)
4. “Filtered Page Views” are Research Data Australia unique page views from Google Analytics tool (filtering out repeat viewings by the same individual in the same session)
Note: “Trusted SW Agreements” are agreements for e-research organisations to access ANDS identifier services using the machine to machine interface.
Note: Data Source Admins are representatives from organisations who manage feeds of information from their organisation to the ANDS registry.
ANDS Production Registry Report 2011
Yearly Summary

<table>
<thead>
<tr>
<th>Total</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collections</td>
<td>1,319</td>
<td>24,658</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parties</td>
<td>894</td>
<td>2,221</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Services</td>
<td>3</td>
<td>41</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activities</td>
<td>690</td>
<td>857</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ANDS Sandbox Registry Report 2011
Yearly Summary

<table>
<thead>
<tr>
<th>Total</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collections</td>
<td>3,753</td>
<td>27,480</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parties</td>
<td>2,392</td>
<td>3,353</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Services</td>
<td>29</td>
<td>77</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activities</td>
<td>27,084</td>
<td>21,238</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2.5.4 Program Highlights, Issues and Breakthroughs

The commissioning and establishment of the initial ‘central’ services of the Australian Research Data Commons (portal, registry, and identifier services) was a highlight of the first year of operation of ANDS. These services have provided some fundamental ‘glue’ for the construction of the ARDC.

The 2010-11 year has seen those ANDS services adopted by ANDS Partners through the formal contracting process of the Seeding the Commons, Data Capture, Metadata Store, and Public Sector Data projects. Usage statistics (above) reflect this uptake and the projections for the coming year are for a similar continued growth.

From a systems point of view this has been an opportunity to test system design assumptions with real world data and work practices. The learnings from this experience are manifest in the 2011 iterative updates and the Project Belvedere re-think of the Research Data Australia.

2010-11 has seen the introduction of a significant service provision capability at ANDS. This has been a breakthrough step in being able to establish reliable research infrastructure and national services. This has included the introduction of policies, procedures and systems based broadly on the ITIL framework and various ISO standards for service provision, service quality, IT security, and risk management. Under this framework a change management system and service desk facility have been established during the 2010-11 year. The implementation of this framework will continue into 2012. ANDS aims to establish solid infrastructure and services that can be sustained after the life of the project.

In early 2010 ANDS took the significant decision to join DataCite as a foundation member. The strategic value of this initiative lies in the potential for the acknowledgement of publishing of data by enabling and promoting data citation, citation indexing, and reward. In 2010-11 ANDS has established a pilot service “Cite My Data” to allow Australian research organisations and data centres to allocate “citation-optimised” Digital Object Identifiers. This is the first step in establishing an infrastructure to track and acknowledge data re-use. The timeline for the bedding down of this infrastructure from a technical, policy, academic, workflow, organisational and individual point of view seems to be in the order of five years.

A continued highlight has been the successful partnership of ANDS with the National Library of Australia and Geoscience Australia (Office of Spatial Data Management) for the establishment of critical informatics infrastructure for two of the fundamental features of datasets: location and parties (places, people and organisations). This “partnering” approach has been very effective and has resulted in ARDC core infrastructure being operated by recurrently funded agencies. It is hoped that similar relationships can be forged with the ARC, NHMRC, and the ABS.

A breakthrough has been achieved in the formation of the RIF-CS Advisory Board. This board is constituted and independently chaired by partners and members of the community. This is a nice example of the community stepping up to take ownership of the standards and protocols of the Australian Research Data Commons.

An issue for this program is the unclear role of ANDS as a project vs. service provider organisation which remains as flagged last year.
2.5.5 Program Learning

The project-driven uptake of ANDS Online Services (mentioned above) has enabled ANDS to learn from the experience of real-world data and work practices. Improved user interfaces and support for metadata crosswalks and automated cross linking are all part of that feedback driven improvement process. In some senses the first iteration of ANDS services was exploratory and anticipatory of demand.

The iterations during 2011 have been informed by provider feedback and designed to facilitate research organisations publishing information about their data assets. A challenging direction for the next years of ANDS will be to similarly optimise services for consumers of information about research data collections.

Another challenge for the next 12 months will be the transition from uptake driven by funded projects to uptake driven by self-interest. From the systems perspective of the ARDC Core program that means optimising the value-add features of publishing data using ANDS national services.

2.6 ARDC Applications

2.6.1 Overview of program

The overall goal of the Applications program is to produce compelling demonstrations of the value of having data available for re-use. These demonstrations of value should:

- result in data being transformed or integrated across multiple sources to produce new forms of information that enable innovative, high-quality research outcomes;
- be relevant to a range of government portfolios; and
- engage with national research capabilities.

As described in the 2010-11 Business Plan, the intention for this program was to achieve this by pursuing three co-ordinated approaches.

The first was to provide *demonstrators of the value* inherent in the use, re-use and exploitation of the newly created pool of data assets in the Data Commons. This would be achieved by identifying big problems that require combined data resources and working closely with leading researchers, researcher groups and Super Science initiatives to commission and deploy specific data sets, tools, and services that meet their data needs and enable them to undertake new kinds of research.

The second approach was to leverage the existing data capture investments by targeting high profile *research champions* in high profile disciplines who have a requirement to link disparate data holdings, and then working with them to identify their data requirements and establish the services and tools they need. These requirements would then be used to create a program of work that will include the development of identified application services as articulated by the identified key senior researchers and informed by the services road map.

As a third approach, ANDS also planned to offer a modest and targeted amount of funding to support community developers in a similar manner to the GOV2.0 ‘hackfest’ events. It was anticipated that these
development activities will produce software tools that could then be re-engineered as robust infrastructure to support the use and re-use of data holdings from the ARDC.

After further consideration by the ANDS Directorate and discussion within the ANDS Steering Committee, it was decided that a better model would be to combine approaches 1 (demonstrators of value) and 2 (research champions) and not to proceed with approach 3 (hackfest).

The resulting program will leverage the outputs from the other ANDS programs, which have been designed to:

- provide underpinning infrastructure to support discovery and citation (ARDC Core)
- enable rich metadata about data to be managed and accessible (Metadata Stores)
- make new data and associated metadata available from a range of instruments (Data Capture)
- make a selection of existing data and associated metadata available from the bulk of Australia’s research-producing universities (Seeding the Commons)
- make data and associated metadata available from government departments (Public Sector Data)
- provide the overall policy and practice frameworks to support better data management and re-use (Frameworks and Capabilities)

### 2.6.2 Outline of projects

Much of the work in this program during the period 2010-11 was carried out through a range of National eResearch Architecture Taskforce (NeAT) projects (refer section 9.2.2. for details of these):

- SISS – Spatial Information Services Stack
- DIAS-B – Data Integration and Annotation Services in Biodiversity
- Aus-e-Lit – collaborative integration and annotation services for Australian literature
- ASSESS – data curation, ingestion and analysis for social sciences data

ANDS also initiated an activity with the University of Queensland to write a report on options for a Who/What/When/Where data combination service that would enable new analysis opportunities on top of this marked-up data.

### 2.6.3 Activity/Deliverables for 2010-2011

All of the NeAT projects were essentially feature-complete by the end of this funding period, with most using a small extension of time to complete their final reports. For details, see the separate NeAT report that is being produced.

The main non-NeAT activity in the Applications program for this funding period was the initiation of a number of discussions about possible projects that might be funded. These discussions followed a process that can be described in general terms as the following:

1. The Applications team and the ANDS Executive Director identify a possible portfolio of activities.
2. The Applications team explore the possibilities within each area - this might involve preliminary discussions with stakeholders (e.g. NCRIS capabilities) or Senior Researchers to see if there is a possible activity that delivers value to those involved.

3. Once the team believes that an area is worth pursuing, they arrange for the ANDS Executive Director to meet with the relevant DVC(R) or equivalent.

4. At the meeting, possible activities are discussed, informed by the preliminary discussions under step 2. Note that a DVC(R) may well indicate a preference for some other area of activity.

5. Once candidate researchers have been identified, Applications staff then meet with them to further refine the engagement and help document it as a fundable activity or set of activities.

6. The activities are contracted and work commences.

At the end of the funding period, ANDS had proceeded to contract for one project (the JCU Tropical Data Hub) and was in serious discussion around a number of other possibilities. The first sets of demonstrations of value will be based around Climate Change Adaptation. ANDS is proceeding with a set of coordinated discussions to enable climate change adaptations and impacts researchers to derive user-specified down-scaled outputs (so-called Geographer Friendly Products) from global climate models. This will require co-ordination between NCCARF, NCI, NeCTAR, RDSI, Griffith, and probably the British Atmospheric Data Centre (BADC) and we are in active discussions with all of these players. ANDS is also looking to commence projects in the Climate Change Adaptation area dealing with refugia simulations, marine-ecosystem responses to climate change (including range extension and species change derived from video), and planning impacts in the built environment.

After a somewhat drawn out maturation period, a number of prospects for other possible Applications projects have been identified. These include imaging/characterisation in the bio space, image analysis, bioinformatics, SMART infrastructure, ecoinformatics, and cultures & communities. The intention is to target each of the national research priorities. Discussions are underway for all of the above areas.

### 2.6.4 Program Highlights, Issues and Breakthroughs

The highlight for this reporting period has been the identification of a number of exciting prospects for demonstrating the value of data aggregation. At present there is more potential activity than can be funded from the program. This is appropriate, as not all of these will come to fruition.

The challenge for Applications will be to converge over time on a coherent set of projects that make up a balanced portfolio. These engagements are the most complex ANDS has yet undertaken, and will require a greater degree of ANDS involvement than most of the existing Data Capture and Seeding the Commons projects. The risk of failure at the initiation and development stages is also higher.

### 2.6.5 Program Learning

The delay in execution of the Applications Program was due to refinement of the program, but also and importantly, the ability of the sector to engage in this program simultaneously with the Data Capture and Seeding the Commons Programs. However there was another important factor not considered in the
Business Plan. That is that the ability to partner with other data intensive infrastructure investments is far greater now than it was 12 months ago. ANDS has engaged in a very rich set of discussions in the first quarter of 2011-12 that was not possible in the same period a year ago.

The other learning for the current reporting period is that getting from early discussions to a fundable contract has the potential to take a significant amount of time (up to six months), perhaps even more than it was for Data Capture and Seeding the Commons projects. This is because of the numbers of conversations that need to be had and the complexity of the desired outcome. ANDS is working actively to accelerate this process.

ANDS view is that the current set of projects under planning is substantially stronger than it would have been if they had had to occur as described in the business plan.

### 2.7 Project Office

#### 2.7.1 Overview of program

This program is designed to ensure the effective and efficient delivery of the ANDS-funded projects with proper reporting of outcomes. It has also adapted to support the delivery of all ANDS outcomes with the goal of delivering operational excellence.

#### 2.7.2 Outline of projects

The function of this program has evolved as ANDS has matured. Its first major task was to run the consultation process on the ARDC jointly with DIISR. This involved visits to each state to run consultation with all key stakeholders to ensure that the ARDC project met the needs of the stakeholders and had endorsement of the stakeholders.

The next major task was to build a process for the development of ARDC infrastructure by Research Institutions.

The other important activities are to manage contracts – a very large number given the outcomes of the EIF process – and provide appropriate reporting on the ARDC project to DIISR, the Steering Committee, Monash, ANU, CSIRO and public reporting through the web site.

The program has oversight of contract management, business management and communications for ANDS. As the projects progress into varying stages of completion the activity is transitioning from contract establishment to contract monitoring. This will ensure that deliverables are accounted for, milestone payments are made in accordance with the contractual obligations and outcomes are reported.

An aspect of the monitoring will be communicating the outcomes of projects to share the knowledge and deliverables derived from them. This activity will become increasingly important as the number of projects reaching completion increases.
2.7.3 **Activity/Deliverables for 2010-11**

The majority of contracts were agreed during this period and have now moved into execution stage which requires significant monitoring. A workflow has been designed and implemented in JIRA software to manage this activity.

As the projects reach completion and outcomes are delivered, there is a need to communicate what has been learned and delivered. The program has developed and is now implementing relevant communication strategies to maximise the impact of these investments.

Key reports and business plans have all been produced on time.

2.8 **Promotion**

ANDS has undertaken a large number of promotional activities during the period July 2010 to the end of June 2011. These include the following.

2.8.1 **Presentations/attendance at Conferences**

ANDS staff have presented at and/or attended a range of international and local conferences to promote the service and to establish relations with other parties. These include:

- Australian Digital Forum, Melbourne
- International Linked Open Data in Libraries Archives and Museums Summit, San Francisco, CA, USA
- Participating in a Metadata Community of Interest, which has been established to influence outcomes for the broader information community and discuss whole of government approaches to metadata issues to enhance the visibility, accessibility and usability of information. This has been an opportunity for ANDS to network with public sector data publishers including ABS, GeoScience Australia, the Office of Spatial Data Management, National Archives of Australia, and AGIMO, the publishers of data.gov.au
- Meta2011 in Canberra, hosted by Institute of Metadata Management
- The IEEE e-Science 2010 Conference, Brisbane
- CAIRSS Community Day 2010, Melbourne
- eResearch Australasia, Gold Coast
- NCCARF Queensland Roadshow, Brisbane
- Metadata Australia Conference 2010, Canberra
- General meeting of CAUL (Council of Australian University Librarians), Queensland
- 15th Australasian Remote Sensing and Photogrammetry Conference (ARSPC), Alice Springs
- JISC (Joint Information Systems Committee) Managing Research Data (MRD) international workshop, Birmingham, UK
- ANDS Roadshow, Townsville
- ANDS Community Events in Brisbane, Adelaide, Melbourne, Perth and Sydney
- ANDS Data Capture Briefing, Melbourne
- ANDS Data Commons Boot Camp, Canberra
- Invitational Research Data Infrastructure Workshop, Prato, Italy (an ANDS-facilitated event)
- Australia-EU Research Infrastructure Workshop, Brussels, Belgium
- Second International Colloquium on Data Management for eScience, Melbourne
- Go8 Data Management Forum

2.8.2 Forums

ANDS has hosted or presented a wide range of forums over the reporting period. The highest profile of these was a series of ANDS Community Events which were run as part of ANDS’ ongoing community building efforts. These were held in Brisbane (QLD), Adelaide (SA), Melbourne (for VIC and TAS), Perth (WA) and Sydney (for NSW and ACT) between February and May 2011, with a total of 193 partner staff in attendance. These partners presented on 92 different ANDS-funded projects.

Feedback from partners indicated that they found hearing about other projects, making connections with others, and learning from other partners and ANDS staff valuable. Having various experts in the room also meant that partners were able to seek clarification on issues that were specifically relevant to their project. During each event, partners broke into discussion groups to either problem-solve or share issues. These discussions have been captured by ANDS to help prioritise future support.

At the end of each Community Event, partners were encouraged to join the ands-partners Google Group and Community Bulletin Board, to help further share information.

In April 2011, ANDS facilitated a three-day invitational Research Data Infrastructure Workshop at the Monash University Prato Centre. The reason for choosing this venue was to facilitate the greatest possible participation from Northern hemisphere invitees. The objective was to bring together research data infrastructure providers and stakeholder representatives in order to explore the potential for much closer collaboration and co-ordinated activity. ANDS will seek to facilitate actions arising from this workshop that are of particular strategic relevance to our next Business Plan, and will encourage action on the other outcomes. ANDS’ staff are already having follow-up meetings to keep the momentum going after the workshop, and starting to deliver on the outcomes. It was agreed by all participants that it had been a successful workshop, with many of them expressing their appreciation to ANDS for bringing this group together.

Other forums included the ANDS Data Commons Boot Camp in August 2010 (for staff working on ANDS-funded Seeding the Commons projects) and the ANDS Roadshow in Townsville in September which provided information on ANDS services and Research Data and the Code for the Responsible Conduct of Research, and was open to all interested parties. In September 2010, ANDS also organized a Data Capture briefing at the Monash Conference Centre for staff of the University of Melbourne, RMIT, Monash University, La Trobe University and the Australian Synchrotron who were engaged in ANDS-funded Data Capture projects. The briefing was designed to provide an introduction to ANDS and its services. A number of participants also provided descriptions of their institutional projects.

ANDS staff, via eResearch SA, have run a series of events during the course of the 2010-11 financial year, called Bright Ideas over Breakfast, at which they have presentations from researchers. To date, all the featured researchers have been involved in ANDS-funded work. Some of the showcased projects have been
the University of Adelaide’s genomics data capture project, a number of projects being managed by CSIRO and the Flinders University data capture project centred on Sleepy Lizards.

2.8.3 Consultation meetings

ANDS staff have consulted extensively with potential and current partners and stakeholders to discuss the services that ANDS offers and how they might be of interest to them.

2.8.4 Newsletter

The ANDS quarterly newsletter, Share, continues to create awareness of ANDS and its activities amongst the research community and stakeholders by providing updates on ANDS-funded projects, highlighting achievements and promoting ANDS events and objectives. The first 3 issues in the 2010-11 financial year focussed on the new uses of research data. At the start of the 2011 calendar year, the decision was made to have 4 themed issues of share, each focusing on one of the four transformations that ANDS seeks to enable. Thus Issue 8 – the last issue in 2010-11 – released in April 2011, focused on the management of research data, showcasing ANDS-funded projects that are enabling this transformation.

2.8.5 Other activities

ANDS has established a mailing list (ands-general) to promote news about the service, as well as participating in other mailing lists as appropriate. A list for people working on ANDS-funded projects (ands-partners) has also been established, as has a bulletin board - http://community.ands.org.au – to encourage the exchange of information about the creation of the Australian Research Data Commons.

ANDS also hosted a visit from a delegation of Danish librarians interested in ANDS’ data management strategies in October 2010.

2.9 Risk Management

ANDS maintains a Risk Register. The risk assessment methodology, adapted from the Australian Risk Management Standard AS/NZS 4360:2004, involves identifying and analysing each risk in terms of how likely it is to happen (Likelihood) and the possible impacts (Consequence). The risk score for each risk is calculated by combining Consequence score with the Likelihood score. This will give a risk score of between 2 and 10, which can then be mapped onto a Risk Scoring Matrix to give a risk rating of HIGH (8-10), SIGNIFICANT (7), MEDIUM (6) or LOW (2-5). Where there is more than one risk measurement area for scoring consequence, the highest combination of scores is taken as the final risk score.

The list of risks is provided in section 10.4.

The Risk Register is updated and evaluated once a quarter. In June 2011, ANDS assessed the residual risk level of all 11 risks in the register, taking into account the effect of the risk mitigation strategies that have been put in place.
It was found that 7 of the 11 risks had been reduced from a rating of high/medium to low. These were the risks around political and governance issues as well as relationships with stakeholders and partners. The risk around recruiting and retaining high quality staff has also been reduced from high to low.

There continues to be medium risk around data providers/federators making their data available and the fact that re-users of research data may not use ANDS services to discover, access and exploit data. However, we have taken measures to mitigate this risk.

Two risks have been lowered from a high to medium rating. There is still some confusion about the role of ANDS versus other related service providers in the eResearch sector, but ANDS has been developing and implementing communications strategies to mitigate this risk, including actively engaging with NeCTAR, RDSI and other eResearch organisations to ensure clarity of the different roles. Another risk that has been reduced to a medium rating is that the standards and technologies that ANDS adopts are not adopted more widely. This has been done by participating in review meetings with the providers of these technologies, such as DataCite, ISO2146 and RIF-CS. ANDS has also been conducting boot camps, community events and workshops to educate a larger audience on how these technologies are developing in order to increase their readiness to adopt them.
## 3 Progress against milestones

### 3.1 Data Capture Infrastructure

<table>
<thead>
<tr>
<th>Milestone Date</th>
<th>Milestone</th>
<th>Progress</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>10Q3</strong></td>
<td>Software testing commences for Early Activity (EA) projects</td>
<td>Achieved: Work on and testing of software for the EA projects underway on 10 projects. Work continued on specifying requirements for EoI projects, with 11 agreed in this period. Records received from 3 projects. Slippage: Fewer records received than expected due to complexity of the task, issues with resources to check records and assist partners, and delays in beginning projects.</td>
</tr>
<tr>
<td></td>
<td>Work has commenced on specifying requirements for Expression of Interest (EOI) projects</td>
<td><strong>10Q4</strong></td>
</tr>
<tr>
<td></td>
<td>Most manual metadata entries lodged through ARDC</td>
<td><strong>11Q1</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>First redeployments of EA software</td>
</tr>
</tbody>
</table>
commence

ATNF project software used in the build of the CSIRO Data Access Portal.

11Q2

All EA software lodged in open source repositories
All EOI software completes testing
First redeployments of EOI software commence

Achieved: 5 projects have made open source software available.
Slippage: Not all EA projects completed, so not all software is available.
Not all EoI projects underway, so not all have commenced testing.
No EoI software has been redeployed.

3.2 Metadata Store Infrastructure

<table>
<thead>
<tr>
<th>Milestone Date</th>
<th>Milestone</th>
<th>Progress</th>
</tr>
</thead>
<tbody>
<tr>
<td>10Q3</td>
<td>Work has commenced on developing a range of EA metadata store solutions</td>
<td>Achieved: Work had commenced on VITRO, ReDBoX, MeCat (ANSTO/ASF). Achieved: Initial assessment complete.</td>
</tr>
<tr>
<td></td>
<td>Assessment of institutional needs and intentions for metadata management is complete</td>
<td></td>
</tr>
<tr>
<td>10Q4</td>
<td>Some EA metadata store solutions enter testing</td>
<td>Achieved: Testing during development cycle commenced for VITRO, ReDBoX, MeCat. Changed: Assessment of needs for further development deferred until production deployment of all existing solutions under development.</td>
</tr>
<tr>
<td></td>
<td>Determination of what needs to be built (either on top of, or in addition to) the existing EA activities to satisfy those requirements is complete</td>
<td></td>
</tr>
<tr>
<td>11Q1</td>
<td>Some EA metadata store solutions in production</td>
<td>Achieved: VITRO in production at Griffith. Changed: Decision on how to best support further development deferred until production deployment of all existing solutions under development.</td>
</tr>
<tr>
<td></td>
<td>Decision made as how to best support deployment of metadata store solutions where desired</td>
<td></td>
</tr>
<tr>
<td>11Q2</td>
<td>Initial round of deployment support provided</td>
<td>Changed: Deployment support decision deferred until production deployment of all existing solutions under</td>
</tr>
</tbody>
</table>
3.3 Public Sector Data Infrastructure

<table>
<thead>
<tr>
<th>Milestone Date</th>
<th>Milestone</th>
<th>Progress</th>
</tr>
</thead>
<tbody>
<tr>
<td>10Q3</td>
<td>Thematic based workshops and forums facilitated</td>
<td>Changed: ANDS development of collections based on themes was put on hold and these workshops were postponed pending a review of the approach. Engagement with GA nearing completion. Engagements with AIHW and ABS under negotiation.</td>
</tr>
<tr>
<td></td>
<td>Engagement with contracted agencies to commence</td>
<td></td>
</tr>
<tr>
<td>10Q4</td>
<td>Initial deposits of manually created collection level metadata information provided to ARDC by key gov’t agencies</td>
<td>Achieved: ABS, AIHW currently in preparation for establishing automated feeds. Other automated feeds listed below have been through this process.</td>
</tr>
<tr>
<td>11Q1</td>
<td>Automated feeds of collection level data by key government agencies demonstrated</td>
<td>Achieved: 40 collections to date from ALA, 1000 from 18 museums via Museums Metadata Exchange.</td>
</tr>
<tr>
<td>11Q2</td>
<td>Automated feeds of data into the ARDC has been provided by key government agencies</td>
<td>Achieved: Over 400 collections from AustLII; AuScope and AODN have feeds contributing nearly 9000 collections from a range of government agencies including Australian Antarctic Division, CSIRO, BoM, AIMS, state agencies such as primary industries and sustainability.</td>
</tr>
</tbody>
</table>

3.4 ARDC Core Infrastructure

<table>
<thead>
<tr>
<th>Milestone Date</th>
<th>2010-11 Milestones</th>
<th>Progress</th>
</tr>
</thead>
<tbody>
<tr>
<td>Period</td>
<td>Description</td>
<td>Achieved: The core infrastructure projects were all complete by the beginning of 2010-11:</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td>-------------------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>
| 10Q3   | Community consultation, specification and procurement of core infrastructure projects complete | - [Research Data Australia](http://www.research.data.au)
- [Register My Data](http://www.registermydata.com)
- [Identify My Data](http://www.identifymydata.com)
- [Publish My Data](http://www.publishmydata.com)
RIF-CS Advisory Board was convened from representatives from the community and partners to advise on interchange standards. |
| 10Q4   | Iterative build of core infrastructure projects initialised and prototypes available | Achieved: During 2010-11, three iterative software builds were released to add functionality to the above infrastructure in a phased rollout (July, Nov, March). |
| 11Q1   | Early release and preliminary testing of core infrastructure | Achieved: Formal third-party unit and functionality testing as well as extensive usability testing was performed on all three software releases. |
| 11Q1   | Phase 2 projects specified | Achieved: The following new infrastructure elements were specified in 2010-11:
- The RIF-CS 1.2 upgrade – Q3 2010
- The new manual record creation interface - Q4 2010
- The ANDS DOI service - Q1 2011
- The ANDS Belvedere project to re-design the Research Data Australia portal was specified - Q2 2011 |
| 11Q2   | Core infrastructure projects commissioned into production | Achieved: The following elements were commissioned in 2010-11:
- The RIF-CS 1.2 upgrade
- The new manual record creation interface |
<table>
<thead>
<tr>
<th>Milestone Date</th>
<th>Milestone</th>
<th>Progress</th>
</tr>
</thead>
</table>
| 10Q3           | First demonstrator of value problem selected, and relevant researchers selected  
                   First candidate institutions for research data champions selected | Delayed as a result of delayed choice – see 2.6.5 for discussion.  
                   Delayed.                                                                 |
| 10Q4           | Development of proposals for first round of research data champions completed  
                   Work for first demonstrator of value scoped  
                   Hackfest date decided                                                   | Changed: Initial discussions held with researchers to identify possible opportunities.  
                   Delayed.                                                                 |
|                |                                                                           | Changed: No longer prioritised.                                          |
| 11Q1           | Work for first demonstrator of value commenced  
                   Research data champions institution selected and funded, work commenced  
                   Hackfest organised and held                                             | Delayed.  
                   Delayed.                                                                 |
|                |                                                                           | Changed: No longer prioritised.                                          |
| 11Q2           | Hackfest organised and held                                               | Changed: No longer prioritised.                                          |

Service management policies were adopted formally in Q3 2010.
Service desk and change management systems were established in 2011.

**3.5 ARDC Application Infrastructure**

<table>
<thead>
<tr>
<th>Milestone Date</th>
<th>Milestone</th>
<th>Progress</th>
</tr>
</thead>
</table>
|                |                                                                           | **Achieved:**  
                   - Belvedere phase 1  
                   - DOI service pilot  
                   - Belvedere phase 2  
                   - DOI Service – full public release  
                   - Terminology support service  
                   Changed: Annotation service not pursued. |
4 Deviations from the Project Plan

There have been three significant variations from the Annual Business Plan 2010/11 (ABP10/11) but all are consistent with the overall project plan. They are that the proposed approach of developing consortia of research institutions to develop and install metadata store solutions at an organisational level, as envisaged in ABP10/11 was not deemed to be attractive after advice from stakeholders, that the ARDC Applications projects took longer to identify than was envisaged, and that expenditure overall was at a lower rate than was planned for. This section describes all deviations for both projects, as they are jointly planned.

The ARDC Metadata Stores identified the value of having institutionally supported services for both object level metadata and collections level metadata. However ANDS could not fully fund different solutions at every institution, so the approach taken was to fund some exemplar solutions and develop a process to enable institutions to form clusters around solutions that best fit their needs. However the desire to support institutional metadata stores was quite varied and substantial numbers of institutions were not yet in a position to determine their needs, or their preferred solution. At the same time, a number of institutions were progressing their own approaches as part of their ANDS funded Seeding the Commons and Data Capture activities. As a result the proposed process was not adopted and new approaches to metadata stores were devised for execution in the 2011/12 Annual Business Plan.

The ARDC Applications program is intended to demonstrate the value of improved use of research data, and demonstrate this in the light of Data Capture projects whenever possible. Due to the delay in commencement and execution of these projects, there was a consequential delay in the commencement of the Application projects. This delay has had the negative effect that the demonstrations are occurring later than planned, but the positive effect that it is allowing projects to commence that are very often exploiting infrastructure developed by both ANDS and our research data intensive infrastructure partners – notably EMBL, IMOS, ALA, TERN, BPA and others.

Actual expenditure was lower than budgeted. Two of the reasons have just been discussed – slower execution of the Metadata Stores program and the Application Program. The other two reasons were lower expenditure on staff than was budgeted – this was simply the time taken to get good recruits, and the more significant reduction was the time taken to move from a commitment – a letter of offer to an institution, to an agreed project, to an agreed contract to receiving invoices. From a budgeting perspective, we had not modelled this time, assuming that shortly after the letter of offer, there would be a first invoice. This has not been the case, and has actually been extremely beneficial. The pie charts (Fig. 8) show the breakdown of expenditure, commitments, and uncommitted funds showing steady progress in moving commitments to contract. This is further discussed on page 55.

This three year project plan shows an un-even level of expenditure (as ANDS had already made substantial commitments) but does balance the need to engage with the sector over a longer period of time, and to demonstrate value early. The business plan for 2010-11 described the proposed activity over the first year of a three year plan concluding in June 2013, and the following chart shows the intended expenditure pattern for the various programs (showing both NCRIS ANDS and EIF ARDC).
Figure 6: Intended expenditure pattern as at November 2010

Figure 7: Intended expenditure pattern at September 2011
5 Co-Investment

5.1 Access and Pricing
All ANDS services are available free of charge to all Australian researchers at publicly funded research institutions and discovery is available to all. As such no cash has been received for the provision of ANDS services, which is in line with expectations.

5.2 Project Co-Investment
As a result of the intended timeframe for the project of 2 years, it was agreed that it would be inappropriate to require co-investment in ANDS projects. The 2009-10 and 2010-11 Business Plans identified that ANDS would place a wholly-funded ANDS staff member within an institution to achieve the aims of that institution and of ANDS. This has been changed as part of the EoI process to a more flexible process of contracting with partners to allow for a mix of staffing needs. In addition, many research organisations have contributed effort to the ANDS projects beyond the ANDS investment. CSIRO, Monash University, Queensland University of Technology, and Griffith University are examples of institutions that have contributed effort.

It is pleasing to note that this additional effort is being continued beyond the life of the projects, showing the importance being placed on research data.
6 Performance Indicators

6.1 KPI Report

The following are the KPIs agreed in the 2010-11 Business Plan, with actual results and commentary included (some of these results have been achieved through EIF ARDC project activity, but cannot be usefully separated):

1. The number and coverage of data repositories providing metadata feeds to the national registry compared to the number of data repositories. ANDS intends to build at least 20 automatic plus 80 manual metadata feeds. This will cover at least 30 out of the approximately 50 research data-holding institutions that we know about.

   **Result:** 21 research institutions fed collections descriptions to RDA, along with 63 individual collections. From these institutions, 38 individual data source feeds have been set up (13 automatic feeds and 25 manual feeds)

2. The number and coverage of institutions and number of research groups with which ANDS has engaged: ANDS will continue to engage with all Australian universities, PFROs, and 4 major Government data providers this year, and through them at least 50 research groups.

   **Result:** ANDS is currently engaging with the following:

   - 35 universities
   - Publicly Funded Research Organisations: ANSTO, CSIRO and AIMS
   - Government data providers: ANDS has engaged with over 25 government agencies apart from the PFROs. These include GeoScience Australia (GA), Australian Institute of Health & Welfare (AIHW) and Australian Bureau of Statistics (ABS) directly; Australian Antarctic Division (AAD) and Royal Australian Navy through the engagement with AODN; and 18 museums through the Museum Metadata Exchange project including Powerhouse Museum, Australian Museum and state museums. Through the AustLII project we have exposed public data from Attorney General's Department and various courts around Australia including the High Court. Through the project with AuScope, engagement has been with Bureau of Meteorology and various state Depts of Primary Industry and Sustainability & Environment
   - National facilities: Australian Animal Health Laboratory (AAHL), Australia Telescope National Facility (ATNF), Australian Synchotron and research vessels: Southern Surveyor and Aurora Australis

30 September 2011

[ands.org.au](http://ands.org.au)
3. The number of institutions with research data management policies and practices consistent with ANDS recommendations: 5

**Result:** 4 – Monash University, University of Melbourne, Queensland University of Technology, Griffith University.

4. The number of times a search is initiated with an ANDS discovery service: 0 – In this year ANDS will concentrate on supporting Google discovery of ANDS data pages, and “see also” services.

**Result:** This indicator is not yet considered important to ANDS current goals, and has not been measured. From September 2011, ANDS will begin to log and measure searches in Research Data Australia and will report these in 2011-12.

5. The number of times an ANDS data page (defined below) is accessed: 100,000 – this is based on a ramp-up as there were few pages that could be discovered at the start of 2009-10, but ANDS will concentrate on directing traffic that way as the pages increase in number.

**Result:** 56,659 filtered page views and 37,527 unique views or visits (from Google Analytics tool).

Note: Previously ANDS measured this indicator with raw page hits from the web server; from this year ANDS will report the more standard measure of filtered page view and unique page visits using the standard filters of the Google Analytics tool. Using the old raw web server metrics the result for 2010-11 would have been approximately 201,054 hits (extrapolation from 2011 logs).

6. The satisfaction of researchers and partners (see below) with ANDS services as measured by an annual survey - no number can be given here, but a report will be provided.

**Result:** Prior to commissioning the survey, ANDS Directors met with senior staff and agreed on the aims and objectives of the survey. The target audience included researchers, research support staff and data custodians.

The aim was to capture understand the community attitude to research data, in particular the areas of focus of ANDS, and then to determine whether ANDS had been successful to date responding to the needs identified by the community survey. This aim recognised that the survey was being undertaken at an early stage in the ANDS program and that not all of ANDS target audience would have had time to comprehend and utilise the ANDS process and support materials.

Insync Surveys was appointed to conduct the Attitudes Survey. Respondents were presented with multiple-choice questions related to managing research data and ANDS. They were also asked to rate their statements in two ways: firstly, to measure the importance of each of the statements to them, and secondly, to measure their impression of ANDS’ performance on each statement.
The survey was completed by 154 respondents who were invited to participate through the various ANDS communication channels. These respondents came from a large pool identified by ANDS staff to cover all identified audiences at sufficient depth to provide significant results.

Insync Surveys reported the following:

Researchers and Research Support Services clients indicated that the most important issues were related to ANDS’ role in improving the management of research data, and helping to publish that data. Other areas included effective communication about ANDS’ online services, and wide consultation with stakeholders.

Among researchers and research support services clients, a number of issues emerged.

They recognized ANDS’ role in improving Australia’s capability to manage its research data, but they were less satisfied with ANDS’ communication of its services and support material.

The survey results suggest ANDS did not appear to have consulted across the research sector, and in particular with researchers. The respondents indicated that ANDS could improve on the current performance; this result has a positive aspect in that while ANDS has not met the needs of the stakeholders, they are sufficiently interested to want to hear more and to be engaged in the process.

Clients also indicated that universities and other publicly funded research organizations receive the most attention in relation to ANDS’ promotion of the re-use of data, the development of data management ambitions, the identification of policy issues and the development of data management frameworks.

The survey indicated very substantial differences in the visibility of ANDS amongst different audiences – where highest visibility is with librarians and the lowest with researchers. This accords with ANDS approach of supporting institutions in their engagement with researchers on research data, rather than direct engagement.

7. The number of data access and sharing agreements with stakeholders – principally research institutions, government data agencies, government research agencies: ANDS aims to strike at least 10 agreements to make data available.

**Result:** 49 organisations with whom ANDS has agreements and who are creating records in either the public or draft systems.

There are two measures that ANDS will not have full control over, but that are important and will measure our success in influencing others’ behaviour:

8. The number of research data sets in harvestable repositories: 5000

**Result:** 26,746 collections as at June 30, 2011. 51% of these collections were from Queensland Facility for Advanced Bioinformatics (QFAB) and 31% came from Australian Ocean Data Network (AODN). The number of ANDS research data sets in harvestable repositories increased almost 23 times from last year.

9. The number of research data sets with persistent identifiers: 7000
Result: 6323 persistently identified datasets, 48% increase from last year’s 4279 datasets.

There is a final measure that ANDS aspires to – it will be measured but is unlikely to be a useful short-term KPI.

10. The number of times a data set is reused and referenced – the ultimate long term measure. At present ANDS is unable to report on this. However, the soon to be launched data citation service will encourage publication of data collections with persistent identifiers in a citable form. Consequently we will be able to use international citation services to measure this.

Notes:

An ANDS data page is a page generated from the ANDS collections registry that describes a data set, a collection, a research group, a research project, or an institution.

ANDS will focus on monitoring Institutions that are research data producing organisations, such as the Bureau of Meteorology, Landsat, the Australian Synchrotron, the Cultural Collections sector, and the research data using organisations, such as the Universities, the PFROs, and affiliates. Many organisations have both roles.

Researchers have many partners in carrying out research and ANDS needs to satisfy there needs as well – this includes funders, assessors, institutional representatives, such as DVC-Rs, eResearch Directors, Information providers such as libraries, IT providers such as University ITS Departments, partner service providers, such as ARCS and NCI, as well as umbrella organisations such as disciplinary bodies such as the Academies, international research bodies, etc.

The qualitative measures are intended to capture not only usage figures, but also attitudinal attributes – ANDS only succeeds with cultural change, so this will be measured as well. The first survey will again set benchmarks, but also help inform future surveys.

6.2 Progress over the Life of ANDS Projects

Having been in progress for 30 months it is possible now to indicate life over the project, along with expectations based on current activity, and when appropriate an indication of what might be possible. Based on activity taking this year, it is possible to determine an estimate for 2011-12, and when appropriate give a figure indicating total coverage. The measures that are described augment the KPI information with additional measures that help understand the Australian Research Data Commons.
6.3 Overall Progress

In Section 1.2, ANDS reported 2 major achievements:

1. Establishing the ARDC, and
2. Providing a meeting place that enables research data management for the whole of Australia to be progressed.

As well as substantial progress in meeting another key objective:

Note 1. Research Institutions denotes all 39 Universities and 4 Publicly Funded Research Organisations

Note 2. Research Data Provider participation is measured by those organisations that are not Universities and Publicly Funded Research Organisations that have an agreement with ANDS to publish research data collections descriptions – this may be indirect through Infrastructure provider partners

Note 3. Research Data Infrastructure partners refer to those NCRIS and EIF infrastructure providers that could exchange research data collections descriptions – this measures how effectively ANDS is partnering with other problem specific data investments

Note 4. Research Fields of Research based on ANZSRC FOR codes – all bar DIVISION 22 PHILOSOPHY AND RELIGIOUS STUDIES are currently covered in RDA

Note 5. Research Data Australia had not been launched as at 30th June 2010

Note 6. This denotes the total number of Research Institutions: all 39 Universities and 4 Publicly Funded Research Organisations
3. Populating the ARDC.

There is a strong relationship between our KPIs, established at the start of ANDS, and achieving the objectives of ANDS. We have shown in the previous section that there are other measures that can also be added to help understand progress.

**Establishing the ARDC:** This requires that all pieces of infrastructure have been installed to enable a researcher to capture research data, the associated metadata determined by a research data management plan, to store both the data and metadata, to connect the data with its context, to publish that data, and to discover and use research data from others. The measures showing the data collections with a persistent identifier available through Research Data Australia indicate that the ARDC has been established, although the measures showing research data management planning and associated institutional research data infrastructure such as metadata stores and automated systems for the capture of metadata and connections tools shown in Figure 3: Australian Research Data Commons Progress provides this information.

**Providing a meeting place for research data management:** Measures that indicate research data management plans, and the participation in the ARDC by research institutions, research data providers, and research infrastructure partners all provide an indication that ANDS has achieved this objective. However there are other important indication provided in this report that help demonstrate success against this objective. ANDS has reported increased level of international engagement, and important elements have been triggered by international requests. ANDS staff have been invited to provide international keynote addresses, participation in international forums, membership of DataCite, and notably participation in an EU/US summit on data infrastructure at EU invitation.

**Populating the ARDC:** ANDS has refined its objectives in this regard: ANDS wishes to populate the commons with collections that are managed, connected, discoverable, and re-useable. Current KPIs particularly focus on registered and discoverable collections, but provide less insight into just how well collections are managed, connected, and re-usable. By way of example, the work done by CSIRO as part of our collaboration with AuScope has lead to collections of professionally managed data from the Geological Surveys that are described, made available, and easily integrated with other similar data using a rich set of web services. Many collections are described, with contact details provided that enable the start of a discussion on possible access to the data. Each approach might be highly appropriate, but they are all counted as a collection that is registered and made discoverable through Research Data Australia. Naturally, as reuse increases, which will be able to be tracked through our data citation services, we will be able to determine the value of easily integrated data via web services, compared to mediated access. Consequently it is of value to track each of the properties that we are considering: that data is collected, managed, connected, discovered and used.
7 Appendices

7.1 Confidential Information
There is no confidential information.

7.2 Project Description Detail

7.2.1 Data Capture Fast Start Project Summaries

ANSTO/Australian Synchrotron

These were originally funded as two separate projects. It became clear to ANDS during the early planning stages that there were significant synergies between the two national facilities, as well as an overlap between the researchers they were servicing. Accordingly, ANDS brokered a meeting between the relevant principals of both institutions at the eResearch Australia 2009 Conference to discuss the possibility of a converged project. This meeting was successful, and the project was constituted with a single project committee and project manager.

The Australian Synchrotron operates 9 beamlines producing up to 2TB of experimental data per day across a wide variety of disciplines from protein crystallography, medical (cancer research) through to the conservation and restoration of cultural objects and works of art. Last year (2009) over 500 groups conducted research at the Australian Synchrotron.

The Bragg Institute is the strongest neutron and X-ray scattering group in Australia. The Institute is named as a tribute to the father-and-son team of William and Lawrence Bragg who were jointly awarded the Nobel Prize for Physics in 1915 for pioneering the analysis of crystal structures by means of X-rays. The facility operates a number of neutron beam instruments, including diffractometers, reflectometers and spectrometers.

The objective of this project is to provide services to researchers to manage their experimental data and to provide data search and access to the broader research community. These services will provide better use and reuse of the data. The ultimate aim is to combine these services into a collaboration environment to allow project teams to interact with the instruments and the data. This project will also improve data and meta data capture for the three most mature beamlines at the Australian Synchrotron, the two High-throughput Protein Macro Crystallography (PX1 and PX2) beamlines and the Infrared (IR) beamline. It will develop appropriate meta data schemata, combine meta data from various sources into a richer set of meta data, and set up a meta data store for the meta data from these beamlines that can be harvested by ANDS for publication of the meta data in the ARDC. This meta data store will be designed so that it can later be utilized for other beamlines at the Synchrotron. Furthermore, the project will extend existing facilities for raw data to be transferred into more permanent storage, including preparation for services to transfer data...
from raw storage at the Synchrotron to the ARCS data fabric once this becomes available. For the IR beamline, the project will also develop a reference spectral database of different materials (e.g. paints, proteins, inks) that enables reuse of these data by users of the IR beamline. This database will be hosted at the Synchrotron on a permanent basis.

CSIRO

The scope of the ANDS‐CSIRO‐ATNF Pulsar Data Management Project is to enable the discovery of, assessment of and access to Pulsar data observed at the Parkes Telescope. The pulsar data observed at the Parkes telescope is publicly available after an embargo period. This project proposes to develop tools required to automate the capture of these data collections, and develop metadata for harvesting into the Australian Research Data commons (ARDC) to increase their discoverability and potential for re‐use. It will include:

- Development of data translation toolsets and implementation of metadata standards, to ensure that valuable science data assets are more easily discoverable and more readily reused.
- Establishing infrastructure to allow direct access to datasets allowing researchers to either download subsets of data direct to their desktop or copy complete datasets utilising ARCS data fabric infrastructure to appropriate advanced scientific compute facilities.
- Establishing middleware systems that track and inform researchers about embargoes for the datasets ATNF holds. Where embargoes prevent direct access to data the system will inform the enquirer who owns the data and their contact details.
- The population of the CSIRO repository with metadata describing the pulsar data sets and the subsequent harvest and population of metadata into the Australian Research Data Commons (ARDC) and the Virtual Observatory.

Monash University

Comprehensive Data Management for Microscopy Research Datasets - This solution allows the Faculty of Medicine, Nursing and Health Sciences to generate and capture experimental information and raw data from optical microscopy instruments in a centralised location, rather than the current practise of capturing this on researchers’ labnotes and personal disks. This approach will establish a concept for the mandatory annotation of digital imaging data to maximise the use, re-use and distribution of experimental information within the scientific community.

University of Sydney

This project will construct software to allow wide use and re-use of microscopy images in breast cancer research, generated at the Westmead Institute for Cancer Research (WICR). In principle the audience could include the entire international breast cancer research community. Australia has significant strength in cancer research, and the Australian research population in breast cancer numbers in the thousands, with many more internationally. There is also potential for other cancer research communities to partner in this project, making it a highly significant international resource. There will be several software outcomes from
the project:
- Tool to segment and convert segments of extremely large images, make them navigable and downloadable in a size and format suitable for analysis, distribution and reuse
- Web portal for search and discovery of images and sections of images from a repository to
- Tool for harvesting of appropriate image metadata to ANDS collection registry

Taken together, these components will provide a path for images from the scanner, to the immediate Westmead group for analysis, and beyond them in de-identified form for use by the international research community.

The objectives of the project—format conversion, distribution portal, discovery, and identity security—generalise beyond breast cancer to a wide spread of other medical research and large image management. Techniques developed and experience gained in this project will be of broad value for future eResearch development, and particularly useful in the medical area.

University of NSW
A new generation of DNA sequencers has recently been installed at UNSW, Southern Cross University and the Australian National University. These instruments can generate DNA sequence data 1000x faster than old technology, and can sequence the genomes of small organisms in a week. This project will establish databases for this DNA sequence information, so that users of the DNA sequencers can access their information in an efficient way. This will centralise DNA sequence data from these facilities, enable collaborative projects and facilitate data sharing. On publishing, research data and associated metadata will be made available for public use, contributing to the Australian Research Data Commons (ARDC).

University of Melbourne
This project will be a key part of establishing a national database of mouse pathology to enhance the utilisation of mouse models of disease by Australian researchers. It will investigate enhanced metadata capture facilities for the Histopathology and Organ Pathology Service based at the Department of Anatomy and Cell Biology, The University of Melbourne as part of facilitating the sharing and re-use of mouse pathology research data both now and into the future. The project will address current metadata scalability and sustainability issues associated with the service in order for the Melbourne Histopathology Service to participate in and contribute to emerging research data networks like PODD and ANDS.

University of Queensland
This project will develop the data capture and sharing services for coral-reef related data being generated by researchers at the University of Qld Centre for Marine Studies, together with their collaborators in community/volunteer groups (CoralWatch, ReefCheck) and government organizations (EPA, DERM). Together these researchers are monitoring and studying the impact of climate change and human activities on coral reef ecosystems. This project will focus on data associated with reefs located between the Southern Great Barrier Reef and the QLD/NSW border. The data will include the following types of observational data: coral cover, coral genetics, algal species distributions, biodiversity data (seagrass cover, fish, stingrays,
dugongs, sharks, dolphins, turtles etc), physical and chemical data (turbidity, salinity, sea surface temperature, pH), nutrient data (Nitrogen, Phosphorus, Chlorophyll a, Sewage Plume Mapping), satellite imagery, 3D benthic data, graphs and publications. In the first instance, we will define common metadata standards and associated services for describing, publishing and discovering these highly heterogeneous datasets. There will be particular focus on automatically capturing the metadata necessary to support discovery, decision, and reuse. The discovery metadata will be made available to the ARDC. Future work will involve the development of more sophisticated data integration and access control mechanisms that support increasing levels of access to the metadata and data over time through - keyword search, ontological and spatio-temporal/mapping interfaces and RSS/Atom feeds.

7.2.2 Neat Projects

These projects were jointly funded and managed by ARCS and ANDS and are described in Section 10.5.

7.2.3 Data Capture EoI Project Descriptions

Deakin University

*Enhancing Filtration Membrane Fouling Data Collection for Water Treatment Research:* Membrane filtration widely used by water/wastewater treatment and desalination industries. However, membrane fouling is one of the most severe performance limiting problems that has negative economic impact. The mechanisms governing membrane fouling is still not fully understood due to the complex nature of foulants and their interaction with membranes, especially in a mixed species environment. This project aims to develop image processing/analysis codes which will create new membrane fouling data from 3 dimensional images obtained by confocal laser scanning microscopy (CLSM). In addition, we will also develop tools to capture the instrument, processing and sample metadata. It is anticipated that the software developed and the collection of membrane fouling data and the accompanying metadata will significantly enhance the collaborations between researchers in membrane materials and water treatment at Deakin and other organizations to tackle the complex membrane fouling issues in water/wastewater and desalination and secure long---term quality water supply to Australian households and industries. The deliverables derived from this project will potentially be used and reused by many membrane materials scientists, water researchers and environmentalists around the world.

Flinders University

*Automated measurement of the responses of wildlife populations to climate change:* The aim of this project is to make the lizard data packages more broadly visible through the Australian Research Data Commons, and also publicly available through a Flinders University institutional research data repository. We intend to develop a system to convert csv files to an appropriate format, and to generate and store RIF-CS collection, service, activity and party metadata. The RIF-CS metadata will reside within a metadata store that features an OAI-PMH harvest point and that will interface with other ANDS webservices (Identify My Data - Pid;
Register My Data; and the Digital Object Identifier webservice currently in development by ANDS). The data itself will reside in the data repository in an appropriate format.

**Griffith University**

*Smart Water*: Water end use study, data capture from instruments, reporting on domestic water use. To build a software system to collect data from remote smart water meters and integrate it into related data sets obtained from other sources and use metadata to describe the data used and the cross-relationships between data. The system will be used to build and integrate data sets to explore patterns of water use, the data sets, descriptions and mega-processes developed.

*Adult Stem Cell & Neurobiological MicroscopyInstrumentation and Research Data Management*: To develop a software system to centralise the management of a large volume of microscopy image and related experimental metadata, allowing researchers within the National Centre for Adult Stem Cell Research ("NCASCR") to more effectively organise and analyse their biological imaging experiments.

Objectives of the project are to capture metadata from images generated from microscopy instruments, Import images in a standard format, allow web based browsing and management of image collections capabilities, share imaging collections, allow annotation of images and image collections, searching of metadata and it will generate and send "published" image collection metadata to Griffith's Research Activity Hub.

**LaTrobe University**

*CMSS RLI Metadata Capture and Publication*: The aim of this project is to leverage the existing ANDS services to support Australian researcher’s use of instrumentation at the La Trobe Centre for Materials and Surface Science facility (CMSS). This project will aid researchers in tracking their datasets and associated provenance metadata from creation at the instrument, in transfer back to institutional repositories, and right through to registration into the data commons. The capability that this project develops will also support collaborative data sharing through the use of persistent identifiers (IDs) that can track data location over time and across institutional boundaries. As a part of the NCRIS Australian National Fabrication Facility capability, La Trobe University’s CMSS will also act as an exemplar facility in eResearch development and the use of ANDS services.

**Macquarie University**

*Glycomics Repository*: This is a two part project: Part I. The development of a world-class glycan reference repository (to be known as UniCarbDB). Part II. Metadata capture enabled from mass spectrometers being used to generate data to support glycan structure study. For more information about this work: UniCarb-DB: A Database Resource for Glycomic Discovery [http://www.ncbi.nlm.nih.gov/pubmed/21398669](http://www.ncbi.nlm.nih.gov/pubmed/21398669)

**Monash University**

*Research Data Management of the Monash Weather & Climate Program (Climate and Weather)*: Monash researchers and research students create significant datasets, mainly through numerical climate model
simulations. The proposed infrastructure will help to revisit and reuse the work carried out by research students and postdoctoral researchers who have left MW&C (Monash Weather & Climate) after finishing their degree or when their contract has ended. Particularly large amounts of data will be created in the coming years when the newly deployed Australian ACCESS climate model is evaluated by a team of postdoctoral researchers. These simulations will be done on the NCI computer platforms in collaboration with other Australian universities, the Bureau of Meteorology and CSIRO. It will be very beneficial for the MW&C researchers to be able to manage this data efficiently and publish it so that other participants will be able to process the data further. Funding of more than $2,000,000 has been secured for this work. In 2010-12 MW&C will carry out a suite of computer simulations to study the urban meteorology in a changing climate. Simulations will assess rainfall patterns in localised urban environments to inform the development, adoption, and operation of stormwater harvesting solutions in collaboration with the Institute for Sustainable Water Researchers, Monash University.

Biomedical Data Platform (Molecular Biology): MyTardis seamlessly facilitates the capture and annotation of protein crystallography data from the Australian Synchrotron MX beamlines, the deposition of the data to the researcher’s institutional data store, and the metadata publication across multiple research data repositories, including TARDIS (Monash-run), Protein Data Bank (worldwide) and the ARDC (nationwide). It has been integrated with Monash University’s High Performance Computing grid to process the raw data and store the output. Visit www.tardis.edu.au to view the public solution.

Tools for curating and publishing research data in the form of media collections (Multimedia Collections & ARROW): This solution, built for the Faculty of Arts, provides a workflow that enables members of the public to submit their life stories about adoption in text and sound files, subsequently able to be reviewed and curated before being published on the History of Adoption website and on the ARDC. The previous manual workflow, involving many systems, has been simplified significantly through the integration work of the project. http://arts.monash.edu.au/historyofadoption/

Capture and publication of Australian ecosystem data from a network of measurement sites (Ecosystem Measurements): This project seeks to directly address the lack of coordination in data collection, archiving and quality control across the active groups in Australia and to promote access to the integrated research data by both Australian and international researchers. These goals are closely aligned to the ANDS principles specified in the Final ARDC EIF Project Plan.

The infrastructure being proposed consists of the implementation of several software systems and associated documentation and training designed to:

- standardise and automate the collection, archiving and quality control of measurements from a network of instrumented towers;
- integrate complementary data streams from different sources into a single data and metadata repository;
- provide documentation and some training in the newly developed system; and
- facilitate the linking the data into a common research data space, through the Australian Research Data Commons.
Capture and publication of data on the history of adoption (History of Adoption): The History of Adoption Project is an ARC-funded project undertaken by researchers from Monash and the Australian Catholic Universities. A central feature of the project is the collection of life stories as self-created sound files directly deposited by contributors, and the online publication of edited versions of these. The whole body of public data will be searchable, and collection records will be made available to ANDS for the Australian Research Data Commons. The project therefore requires infrastructure for data capture, enrichment and dissemination.

Data and metadata will be directly deposited by contributors: data as sound files both pre-recorded and directly streamed via Skype and Google Voice; metadata in the form of information identifying the source of the deposit and the historical context of the adoption described. Before deposits are accepted, depositors must sign off on ethical protocols determining their relations with the project and the project’s responsibilities to the depositors. Before data is made accessible to the public it will be edited for legal and ethical purposes, with additional metadata added at this stage. Different levels of access will be needed; in the first instance so that it is accessible only to its author, for review and confirmation; secondly, with full public access. At this point readers will be able to add comments to the data.

Data Publication to Interferome (MIMR/Interferome): Previously Ecosystem research data was difficult to access and hampered the actual research and analysis of the data.

The solution provided by the Ecosystem project has enhanced the research process and provided new research opportunities.

Overall, the solution cuts down the time significantly before the researcher can begin their research work and facilitates much simpler sharing of data. Due to these benefits the Ecosystem solution has been very well received by the OzFlux community of researchers. This adoption is a testament that the solution funded by ANDS is of significant value to the researchers in the community. The next key challenge is to enhance the Ecosystem solution to facilitate Ecosystem research on a larger scale at the national level.

This project realised a number of groundbreaking technological advances in the ecosystem distributed data system (Eddy). In particular, the project saw the development of an improved accessibility to the climate data and an auditable, searchable, distributed data system of research data. The Eddy system is fully integrated with the Research Data Australia Service so that the metadata of the research data is discoverable by the broader national and international research community. The Eddy system supports a range of scientific data formats, like NetCDF (Network Common Data Form), GRIB2 (GRidded Binary) which makes the creation, access and sharing of scientific data easy.

In addition, the Eddy system is a multiple-tier web-centric application and follows a proven classic design pattern, Model-View-Controller. Each component of the Eddy system is independent, so it is flexible for future development and can be easily maintained. The Eddy system also uses the Monash Large Research Data Storage (LaRDS) service, which ensures that the research data reliably backed-up and secure. The solution can be viewed at: http://ozflux.its.monash.edu/ecosystem/

30 September 2011

ands.org.au
Queensland University of Technology

*Greenhouse Gas Emissions from Australian Soils*: Data and metadata capture from automated N2O gas sampling systems of emissions from Australian agricultural soils. Software written in eclipse will take the metadata as a "data package" and then sent to Metacat, a web-based server application with a database.

*Biodiversity*: Study the effects that humans have on the environment. Adopt acoustic sensing, database and web service technologies to capture audio data. Generate audio data and tags. Software to facilitate the mapping and transformation of data contained in the relational database, into a form that will be able to be ingested into QUT's Metadata Hub.

*B150 BigJam*: A roster of 900 musicians improvising across a number of genres, performing and recording non-stop music for 150 hours. Software written using eclipse to facilitate the management of metadata relating to multimedia data captured. Management of metadata -- annotation, mapping and querying of multimedia data.

RMIT University

*Data Capture from High Performance Computing Multi-User Environments*: RMIT is currently a very large user of state and national high-performance computing (HPC) facilities. This project will develop and deploy software tools and applications which will be deployed for the NCI Supercomputer National Facility, the VPAC Supercomputer Facility and the RMIT HPC Facility.

The following simulation packages will be targeted:

- VASP (a Density Functional Theory code for simulation of material properties using plane waves)
- CRYSTAL (a Density Functional Theory code for simulation of material properties using Gaussian basis sets)
- SIESTA (a Linear Scaling (a Density Functional Theory code for simulation of material properties)
- GULP (Empirical potential code for simulation/modeling of material properties)
- DFTB (Density Functional-based Tight Binding)

University of Adelaide

*Automated capture and publishing of data generated on high throughput plant phenomic platforms*: Extract and curate data generated by the Plant Accelerator's LemnaTec Smarthouse platforms. Publish the data to a navigable public repository, and metadata to Research Data Australia.

University of Melbourne

*Melbourne Neuropsychiatry Centre (MNC) Bioinformatics Development Project*: The MNC has one of the largest databases of brain scans and associated neuropsychiatric research data in the world. It has National and International collaborators using and contributing to the database. Broadly ANDS-related activities will involve:

- Building a workflow for automatic documentation of dataset segments used in individual studies and publications. This will include researchers, datasets, associated projects and publications.
- Building a workflow for automating creation of citable persistent identifiers for unique studies and linking with publications.
- Building a software to automate capture of public facing metadata to University of Melbourne Registry which will deliver collections metadata to the ARDC.
- MNC has 270+ publications resulting from datasets stored in the MNC database.

Youth Research Centre's Life Patterns Project: Longitudinal qualitative and quantitative survey data capture and reuse: "The Youth Research Centre's Life Patterns Research Program maintains an extensive qualitative and quantitative data base on a cohort of 2000 young Australians who left secondary school in 1991 and of a second cohort of 3000 who left school in 2005. With ARC funding through to 2014 for annual qualitative and quantitative data capture for the second cohort (Gen Y) and biannual data capture for the first cohort (Gen X), this activity aims to enable wider access and use of the data by developing the infrastructure to make sets of the existing data available for re-use, streamline capture of new data so that it is more readily available for re-use, and build the capacity to efficiently respond to future requests for derived data sets.

Appropriate structures for the capture of relevant metadata (compliant with DDI2, DDI3 and RIF-CS schemas) and tools to extract this metadata from workflows will be developed.

Video data in the Social Sciences. Optimising Metadata Capture, Data Sharing Procedures and Long-term Reuse: The University of Melbourne has an especially rich humanities and social science research community that utilises video as its primary form of data capture. The increasing use of video as a research tool poses particular challenges for aggregated data storage initiatives. This project will integrate metadata capture facilities at selected sites within the University of Melbourne as part of facilitating sharing and re-use. The project will address current metadata issues associated with large-scale audio-visual repositories and workflows to enable efficient generation of metadata, ensuring that stored video data is accessible and searchable through the ARDC. The project will:

- Develop software to automate the capture of metadata from existing mature video storage systems developed by the ICCR (International Centre for Classroom Research),
- Develop and where possible - utilise existing infrastructure to identify generic workflow tools that will enable rich knowledge of data sets, access services and parties to the research to be systematically (RIF-CS) captured from the researchers,
- Develop standards compliant video data and metadata deposit services.

These are generic goals which are broadly applicable to activities elsewhere within the university, for example in the Faculty of Architecture, Building and Planning and the Faculty of the VCA and Music.

Federated Neuroimaging Collections in the National Data Commons: DaRIS is a raw data management system based on the Mediaflux digital asset management platform and has been in operation for the last 3 years at the Neuroimaging Computational and Data Management Facility (CDMF). There it has been used to routinely receive MR images from researchers and organise them into a subject-centric data model, ready for access by project members. It hosts over 70 mouse and human projects, each with many tens of subjects and some with time-dependent data.

- Map DaRIS project-metadata to the ANDS schema
- Write a DaRIS service to populate ANDS-compliant metadata,
- Develop an adapter to harvest the ANDS-compliant metadata from DaRIS
- Connect identifiers within DaRIS to ANDS persistent identifiers (PIDs).

**Humanities and Social Science Research Data at the University of Melbourne:** The University of Melbourne has one of the most rich and diverse humanities and social science (HASS) research communities in Australia and is well ranked internationally. HASS researchers at Melbourne generate and hold valuable data sets and associated materials that are currently not easily discoverable, accessible or configured for further research purposes. This project will build infrastructure (tools and services) to connect this diverse community with the UoM Registry (Vitro) which will in turn communicate the relevant metadata to the ARDC. The project will:

- Develop and utilize existing (OHRM-based) infrastructure to identify generic workflow tools that will enable rich knowledge of data sets and related materials, access services and parties to the research to be systematically (RIF-CS) captured from the researchers.
- Development of a generic web services-based data capture tool to be used both by researcher staff, data librarians or other staff in the data management fabric. This will be based on the 'pre-register' work done for the Australian Women's Register in 2009
- Develop standards compliant 'access service' descriptions
- Ensure project, data, party and service descriptions concord with Data Documentation Initiative (v2&3) requirements.
- It will inform the development and utilisation of digital and analogue archival preservation, curation and access systems for the University

**Capture of Complex Data to Support Clinical Research in Cardiovascular and Neurological Medicine:** Complex physiological data is routinely collected on patients as part of clinical care (echocardiography, intravascular ultrasound, x-ray angiography, optical computerised tomography, patient clinical data, etc.). However, this rich multi-model data is not usually subjected to subsequent analysis nor is it made available to researchers from other disciplines for novel analysis. Making this multi-model data available along with patient outcomes such as morbidities will provide the opportunity for collaborative groups to employ novel strategies to developed assessments and models based on this data. This project will form necessary base of making multi-model data collections available, enabling the establishment of new links between biomedical research groups in engineering, physics and bioinformatics. This project will occur in collaboration with BioGrid Australia where it will use the access, de-identification and privacy protection protocols already established there.

**Founders and Survivors Project:** "The Founders and Survivors Project (http://www.foundersandsurvivors.org/) has brought together a number of research data sets created from records relating to the 73,000 convicts transported to Tasmania in the 19th century and their descendent's to create a population database of national and international significance for historical, demographic and population health researchers.

This project will:
- Develop a toolkit based around the projects XML/TEI workflow for further relevant records sets to be systematically ingested into the population database,
Build the infrastructure to enable persistent identification and descriptions of derived data sets produced on request from the population database to be made available to the ARDC

University of New South Wales

ARDC Linked International Glycomics Repository & Instrument Data Capture: Data and metadata capture from mass spectrometers, hosting of world-class glycan repository.

An international antibiotic-resistance gene cassette database: Software to allow researchers to contribute and find data on grammatically-defined bacterial gene cassettes

Data capture and integration across multiple platforms: Manage and integrate data from the Analytical Centre, the Membrane Library, and the Electronic Lab Notebook

University of Newcastle

Data Capture for the Data Commons: The Health Behaviour Research Group (HBRG) is a key member of the Priority Research Centre for Health Behaviour at the University of Newcastle. HBRG enjoys a strong international reputation for quality research in behavioural and public health approaches to health promotion, health service evaluation and cancer control.

Collection and management of survey data is a fundamental activity for HBRG. On commencement of a research project, HBRG researchers determine characteristics of the proposed survey, for example characteristics of the (subject and control) participants who will take part in the survey; the questions to be asked; and the statistical analysis to be performed. The survey instrument, once designed, is then administered using paper-based questionnaires, or more recently using purpose-built computer software. Once datasets have been collected, the group’s researchers analyse it, leading to suggestions for change to the processes and procedures performed within the health care system.

University of Queensland

Spatially Integrated Social Science: Capture of derived spatially-integrated data from interrogation of data contained in the social science data archive that includes both legacy databases (e.g., Australian Bureau of Statistics (ABS) and Australian Electoral Commission (AEC) data) as well as derived datasets and newly generated data.

Aquatic Species Tracking Repository (Oz-Track): It will create a platform for capturing and managing acoustic array and satellite data which includes: animal location data (lat, long, depth) as well as biological data (body temp) and environmental data (water temperatures, river heights/tidal flows and salinities)

It will generate software to enable the capture of animal movement data from receiver arrays and its search, browse, retrieval and correlation with environmental, hydrological and oceanographic data

3D Anthropological and Archaeological Data capture of 3D digital models and deposit of metadata to Collection Repository: Project will streamline the capture and generation of 3D digital models of
anthropological, archaeological and museum artefacts from a Konica Minolta 3D laser scanner and enable their upload (and metadata ingest) into an institutional repository.

**Linking the EMBL Australia EBI Mirror with the Australian Research Data Commons: The Aims and Objectives of this project are:**

Component A: to populate RDA with collection descriptions of data held in the EBI databanks.

This will specify and implement automated systems to identify bio-molecular data collections already held in the EBI databanks that are ""Australian-associated"", extract metadata from the EBI databases for these data, generate RIF-CS collection-level descriptions, make these collections discoverable through RDA, and enable navigation from RDA to the corresponding primary data entries or sections in the EBI Mirror. ""Australian-associated"" data in this context is relatively broad and may be defined as: Data submitted from Australian-based researchers; Data associated with sets (and subsets thereof) of Australian species (e.g. native, introduced, agricultural); datasets of relevance to Australian-based researchers (e.g. related to common diseases facing the Australian population). The overall approach will enable researchers in a wide range of fields that are not necessarily bio-molecular domain experts (e.g. marine science, ecology, climate studies, agricultural science, health science etc) to use RDA to easily locate and use relevant molecular data in their studies.

Component B: to enable submission of descriptions to RDA for data associated with secondary analyses performed using the Australian EBI mirror.

This will specify and implement automated systems to extract metadata from secondary analysis of data (either EBI data or user-uploaded data) performed by researchers using the NCI-SF in Bioinformatics instance of Bioflow, to generate appropriate RIF-CS collection and service descriptions for datasets generated through the secondary analysis, make these metadata discoverable within the RDA, and enable navigation from RDA to the corresponding Bioflow workflow and EBI data in the EBI mirror. This will enable complex bioinformatics workflows that can be applied in the re-analysis of molecular data, to be found by a wide audience and re-used.

Component C: to deliver collection descriptions for EBI data aligned with BioPlatforms Australia (BPA) themes to RDA.

This will engage BPA and other relevant R&D communities to help conceptualise and specify views of Australian EBI Mirror data within the context of BPA theme projects, and organise data within the Mirror to constitute these views and make them discoverable through RDA. This will enable the data produced through the NCRIS/Super Science BioPlatforms Australia investments to be presented through the EBI mirror and RDA.

**University of South Australia**

*Development and testing of a data capture tool for instruments at the Ian Wark Research Institute:* Instrument (NanoTOF and Mastersizer) interface development. Data description, storage and access.
University of Sydney

**Metadata Store/Aggregator:** An institution-wide metadata store will be implemented in collaboration with University of Sydney ICT. This store will aggregate information from suitable University enterprise systems, have applicability across many areas of data capture and complement the Sydney “Seeding the Commons” project.

**FieldHelper:** The project will develop a cross-platform user-friendly and flexible software tool with standards compliant workflows for managing the flow of research data from fieldwork-based research projects through repositories into the Australian Research Data Commons. Discoverability and reusability of such data will be improved through facilitating standards compliance at the time of data capture, and interoperating with other ANDS services and projects. The tool will be designed to be adaptable to diverse areas of study in the humanities and social science domains.

**AgDataCapt: Capturing Agricultural Data:** Standards-based body of data and metadata will be coordinated and include data from sensors on soil, water and rain, greenhouse gas and carbon data, and weather. Generic and extensible tools will be developed to integrate appropriate standards across the areas of data capture. The areas of data capture are: Soil moisture, soil and air temperature, radiation, 3D wind speed and directions, CO2, water vapour, weather and tree water use data, greenhouse gas emissions from soils; GPS-referenced data on wheat and barley crop inputs and performance; Soil data from sensing system for monitoring of agro-ecosystems for sustainable landscape; Soil data and spatial prediction functions for soil variables; and Data from farm-based private rain gauges.

University of Western Australia

**Deployment and Configuration of Institutional Metadata Repository:** The overall aim of the UWA Data Capture projects is to ensure that the University of Western Australia has a metadata store for capturing metadata about research data collections, and that the University is in a position to contribute metadata in the RIF-CS format to the ARDC. The projects will also implement data capture tools and systems for three strategic research areas: rock art research, video data for marine ecology research, and integrated data capture for characterization and analysis in bio-imaging.

The UWA1 project specifically aims to ensure that the University of Western Australia has an institutional metadata store capable of capturing metadata about research collections, and that metadata from this store can be contributed in the RIF-CS format to the ARDC.

**UWA Rock Art Studies Data Management:** The desired result of the Rock Art project is to improve the ability of UWA archaeologists to maintain their data for personal use, and to share it with collaborating researchers and with the wider research community when appropriate. Making the data available to indigenous communities is also a goal.

The scope of the project is currently to provide a solution that works specifically to meet the needs of the Rock Art group within the UWA Archaeology department. There is however a strong desire to develop the system so that it may be used for other similar groups within Australia and for other archaeology groups.
The Rock Art Data System (RADS) solution will be connected to the UWA VIVO metadata hub, which in turn will be connected to the Research Data Australia metadata hub.

*Organisation of Australian Underwater Video and Still Imagery:* As technology for capturing information about marine environments improves and becomes more widespread, the volume of data collected has been increasing by several orders of magnitude.

Marine researchers in Australia and overseas are using ever more sophisticated imaging technology which is producing valuable data.

However the cost of creating this data can be significant and thus it is imperative that its value is fully extracted by sharing it with the wider research community to prevent overlap and find new results that aren’t necessarily obvious from smaller data sets.

The need is to create a system for managing the raw images, data that describes the various marine objects and properties (e.g. fish length, fish species, water depth, water temperature, salinity). It should also maintain a record of metadata that describes the research study being conducted, provides information about the relevant grants and organisations and importantly the researchers involved in carrying out the study.

*Integrated data capture for characterization and analysis:* The desired result from the CMCA project is to implement software and hardware infrastructure that allows the microscopy data (raw image and spectral data) to be stored, and searched based on appropriate metadata. This metadata includes machine settings and other directly available data. It also includes data that cannot be directly captured from the instrument such as information about the researcher, research group, grants funding the research and sample information (e.g. mouse brain). This latter type of information currently requires much manual effort to collate and thus automating its collation with the raw data is of significant benefit to CMCA management.

**University of Wollongong**

*Biomechanics Data Capture Project System:* Data and metadata capture from multiple instruments in biomechanics lab. In more detail: Kinematic data; OptoTRAK 3020 system, Force data; Kistler Multichannel system, Electromyography (EMG) data; Noraxon Telemyo system, Ligament laxity; Dynamic Cruciate Tester, Pressure data; Novel pressure measurement system, Survey dat; Survey Monkey

### 7.2.4 Public Sector Data Project Descriptions

**AuScope**

This project builds on the work undertaken in the SISS Project by:

- deploying the SISS offering at Government data providers with data holdings important to national research priorities and CSIRO flagships
- developing some of the component services and functional capabilities needed to realise a spatial information data commons within Australia supporting linkages with academia, research, public and private sector data and service providers and users
- leveraging the Open Geospatial Consortium (OGC) and ISO standards that have been adopted by a number of capability areas and institutions in Australia and worldwide for which various reference implementations already exist for information services and portal interfaces
- combining and building on existing expertise in this foundation area in order to assist multiple capability areas and institutions that have interest in geospatial web service middleware and clients.

CSIRO WRDM

The CSIRO Water Resource Data Management (drawing from the Water for a Healthy Country Flagship Sustainable Yields) project is developing tools required to automate the capture of data collections, develop metadata for harvesting into the Australian Research Data Commons (ARDC) to increase their discoverability and potential for re-use. It will include:

- Development of data translation toolsets and implementation of metadata standards.
- Put in place infrastructure to allow direct access to datasets allowing researchers to either upload subsets of data direct to their desktop or copy complete datasets to appropriate advanced scientific compute facilities.
- Establish a Water Research community data registry/repository building on the capabilities developed in both CSIRO Marine & Atmospheric Research/TPAC and AuScope initiatives.
- Establish middleware systems that track and inform researchers about the licensing arrangements for each of the datasets CSIRO holds.
- The population of the CSIRO repository with metadata and population of metadata in to Research Data Australia and the Water Resources Observation Network.

It is anticipated that other datasets will be added on completion of the project and establishment of the system.

AODN

Core data captured routinely from research vessels managed by CSIRO and AAD will be published in near real-time in a coordinated way that enables ready access to, and combination of the datasets via the Australian Ocean Data Network (AODN). These vessels are major national marine science facilities and are the source of a significant proportion of Australia’s in-situ blue water ocean research data. The AODN, when adequately populated, will be an online network of marine and coastal data resources, which will include data from the six AODC Joint Facility (AODC JF) partner agencies and other data providers, supported by standards-based metadata, and will serve data to support Australia’s science, education, environmental management and policy needs: Australia’s digital ocean commons. Publication via the AODN will provide the public with a simple access point for these core vessel-sourced data and ultimately other vessel datasets not currently routinely published. As well as being valuable data in their own right, the work associated with standardising/automating the publication of the routinely captured instrument data, could act as an exemplar for real-time data publishing in other domains. Vessel managers will also work with ANDS to automate the provision of researcher and project information relating to use of the ships to RDA. Collection level metadata would flow to the RDA along with the deep metadata, via the AODN. The AODN’s data turbine middleware will be harnessed to visualise the core instrument-sourced data.
AustLII

This project develops software to include information about the 300+ free access Australasian legal databases located on the Australasian Legal Information Institute (AustLII) system in the ANDS discovery services. It also develops software and automation procedures which will be tested and used, during the project timeframe, to make at least 40 additional ‘public data collections’ that are important for legal research available (via AustLII) and more accessible to researchers, so as to ensure greater use and re-use of these existing data resources. Outcomes will be greater awareness of accessible legal data, particularly in disciplines outside law, and improved research in a broad range of different research communities in Australia and overseas.

Powerhouse Museum

The primary aim of this project is to establish and populate the Cultural and Historical Collections Metadata Exchange with Collection Level Descriptions conforming to the RIF-CS schema for delivery into the Australian Research Data Commons (ARDC). This includes the establishment of data capture processes, data standards and infrastructure amongst museum data partners in the project and the establishment of protocols to enable automated transfer of the collection descriptions into the ARDC. The Exchange will be hosted by the Powerhouse Museum, drawing on established infrastructure and in-house expertise in terminology development. A number of additional services and tools are also proposed to extend and enhance the data set for researchers and to establish automated harvesting arrangements. The project will work closely with HASS researcher communities, primarily through its research partners including Monash University, Flinders University, RMIT University, University of Sydney and through the Australian Academy of the Humanities, and the Council for the Humanities, Arts and Social Sciences.

The principal contributors to the metadata repository will be the major museums which are affiliated with the Council of Australasian Museum Directors. These museums hold some of the largest and most significant museum collections of interest to HASS researchers. Museums Australia will assist in gaining participation of regional and local museums.

GeoScience Australia/OSDM

Geoscience Australia and the Office of Spatial Data Management have three catalogues which hold metadata of significant value to researchers. ANDS will analyse the catalogues and endeavour to map to RIF-CS.

Activities include:

- Analyse ASDD, Geocat and Geomet catalogues and if feasible, map all or part of catalogue records to ISO19115;
- Analyse any residual records which did not map to ISO19115 and map to RIF-CS where feasible;
- Crosswalk ISO19115 records to RIF-CS format;
- Install OAI-PMH harvest point at GA;
- To harvest the resulting RIF-CS feeds from Geoscience Australia to Research Data Australia.
# 7.3 Progress against activities

## 7.3.1 Data Capture Infrastructure

<table>
<thead>
<tr>
<th>Proposed Activity</th>
<th>Progress</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANDS has either entered into contracts (or has substantially agreed on project descriptions) for Data Capture projects at 24 agreed institutions. As some of the projects are still being refined, it is not possible to list all of the activities that will be undertaken. They appear likely to run across the entire disciplinary range with an understandable bias towards the instrument-centric sciences. The projects will all commence by providing a small number of hand-crafted feeds of Collections with associated Activities and Services. These feeds will be used to inform the process of specifying the software that is required to automate the creation of this metadata, as well as to allow feedback on the metadata to be produced. Over the remainder of the project, the software will be specified, built, tested and documented. The expected runtime for most of the projects is 12 months from the date of contract signing. A small number of universities have taken advantage of the time extension for ANDS to also extend the duration of their data capture activities into 2011/2012.</td>
<td>At June 30, 2011, ANDS had either entered into contracts (or had substantially agreed on project descriptions) for at least one Data Capture project at all of the EOI institutions except the University of Western Sydney and University of Technology, Sydney, (with some additional projects at selected institutions still under discussion). A breakdown of the progress made in relation to this is provided in section 2.2.2.</td>
</tr>
</tbody>
</table>

The following NeAT projects assigned to the Data Capture program will complete their final year of activity, as they have all been funded through to the end of the 2010-11 funding period. | All NeAT projects were completed or being wrapped up at the end of the reporting period. All relevant payments have been made. |

In addition to the institutional and NeAT investments, the following projects have been | These are now all agreed. The majority of these ‘fast start’ activities are |
agreed or contracts have been signed as a result of fast start activities completed or nearly complete.

7.3.2 Research Metadata Store Infrastructure

<table>
<thead>
<tr>
<th>Proposed Activity</th>
<th>Progress</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANDS will partner with the Australian Digital Futures Institute at USQ (subject to a suitable agreement being reached) to develop a solution that meets the requirements listed above. This solution would be integrated with other ANDS services and the ARCS Data Fabric.</td>
<td>This project is progressing well, and is being delivered in conjunction with separately-funded work at the University of Newcastle. This will ensure that the solution is fit for purpose.</td>
</tr>
<tr>
<td>ANDS will also commission software for early installation at Monash University and the Australian Synchrotron to improve metadata management for data objects.</td>
<td>Monash University has not yet commenced its work on metadata store infrastructure. The development of metadata store infrastructure at the Synchrotron is being undertaken in combination with the funded Data Capture project.</td>
</tr>
</tbody>
</table>

7.3.3 Public Sector Data Access Infrastructure

<table>
<thead>
<tr>
<th>Proposed Activity</th>
<th>Progress</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANDS will commission a fast initial rollout of the SISS offering to an initial group of institutions (that includes several Departments of Primary Industry with whom AuScope is already engaged), who hold well-managed spatial data and are willing to share it.</td>
<td>Engagements with Bureau of Meteorology and GeoScience Australia in progress. Extension of time granted to allow the project to further engagement with the BoM. Relationship with VeRSI established and training and transfer of knowledge has taken place. VeRSI is working with the groundwater community in Victoria – DPI, Vic; DSE, Vic and University of Ballarat.</td>
</tr>
<tr>
<td>ANDS will work with the CSIRO to identify and expose for discovery an initial set of data collections.</td>
<td>Project complete. Murray Darling Basin Sustainable Yields data exposed through an institutional portal and RDA. Software deliverables shared through</td>
</tr>
<tr>
<td>SourceForge and used as the basis for building a CSIRO enterprise data management and publication system.</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td></td>
</tr>
<tr>
<td><strong>ANDS will also fund the development of the National Criminal Justice Data Network to aggregate and provide increased access to Criminal Justice data from a range of state-based data holders in this area, including NSW Bureau of Crime Statistics and Research, NSW Dept of the Attorney-General, the SA Office of Crime Statistics and Research, SA Dept of the Attorney-General, the NT Department of Justice, the Australian Institute of Criminology, and the ABS National Centre for Crime and Justice Statistics.</strong></td>
<td></td>
</tr>
<tr>
<td><strong>The selection of further partnerships to make research data discoverable through the ARDC will be based on the public sector dataholders’ interest in making data available, their commitment to ongoing data services beyond the life of the project, and the importance of the data to Australian researchers. Consequently selection will occur through discussion with both the data holders and representatives of research disciplines and capabilities.</strong></td>
<td></td>
</tr>
</tbody>
</table>
| AODN – project complete. Records from project Underway made available via AODN portal and Research Data Australia. Additional data from AODN portal also harvested.  
AustLII – Over 400 datasets now available via RDA as well as directly from AustLII. Extension granted to seek further datasets.  
Powerhouse Museum – Over 1000 collections from 18 museums from around Australia are in the final stages of quality assurance before exposure via RDA.  
NAA – Proposal to identify collections and feed into ARDC has received no response and other opportunities will be sought to expose this data.  
GeoScience Australia/OSDM – Collaborative work to map key data catalogues for exposure of an initial 7000 records into ARDC is complete. This was an ‘assisted engagement’ where ANDS staff go on site and work with GA staff. It is awaiting some IT |
development within GA to allow the feed to be activated.

Aus-e-Lit – Annotation service and compound authoring tool for the Australian literature community. This project is in the final stages.

DIAS-B – Provision of annotation service for Atlas of Living Australia. This project is also in the final stages.

SISS – development of spatial data infrastructure for AuScope to build new spatial data services out of existing ones. Commencing with GA and BoM data using NCI.

### 7.3.4 Australian Research Data Commons Core Infrastructure

<table>
<thead>
<tr>
<th>Service</th>
<th>Progress</th>
</tr>
</thead>
</table>
| Dataset Identifier Service using Digital Object Identifiers (DOI) | ▪ Software developed and implemented; Cite My Data pilot service in operation  
  ▪ Contributions made to the international DataCite metadata kernel and prototype software contributed to the DataCite central services  
  ▪ Adrian Burton chairs the DataCite Services Working Group and Margaret Henty has contributed to the design and content of the DataCite website  
  ▪ ANDS is a full member of the DataCite consortium and has attended two general assemblies |
| Terminology Support Service *(Classify My Data)* - Develop and implement software | ▪ A Vocabulary Services Technical Working Group has been established |
| to enable the creation, management, and publication of human and machine readable ‘terminologies’ (also known as controlled vocabularies) for use by the Australian research and higher education sector. | with key commonwealth departments and NCRIS facilities
- Software development is in design phase and has been scheduled provisionally for release in early 2012 |
| --- | --- |
| Research Data Australia:  
- Update information and presentation design  
- Implement digital Finding Aids against the ANDS research themes  
- Facilitate linkages between  
- Develop faceted search results capability  
- Enable on-screen and export of citation for data collections  
- Allow map search by polygon  
- Allow structured searching | Belvedere project completed phase 1 with phase 2 and 3 in the development pipeline
- Finding aids requirements analysis under way
- Automated linkages (bi-directional links) implemented
- Faceted search designed and on schedule for Q3 2011
- Polygon search implemented
- Boolean search designed and on schedule for Q3 2011 |
| ANDS Collections Registry Ingest mechanisms: Develop alternate authoring, editing, and ingest mechanisms for information about collections, parties, activities and services to allow for a range of technical and platform capabilities | New manual input application released in Q1 2011
- New sandbox environment designed and on schedule for Q3 2011 |
| Research Party infrastructure: In collaboration with the National Library of Australia develop extensions to *People Australia* to enable:  
- The capture of authority data for participants in the Australian research and innovation community;  
- People identifier services to be available from software supporting repositories, data storage systems, research management systems, cataloguing, metadata | The NLA Party Infrastructure has been built to enable the contribution of records from the research sector either via Research Data Australia or directly and support for the ingest of party records in RIF-CS 1.2 has been provided in the NLA Harvester. This transformation is based on the ANDS Contributors Guide and designed to be easy for the Trove business area to manage.
- Support for the provision of party records in RIF-CS 1.2 has been provided via the SRU and OAI |
management, bibliographic management and online publishing tools and systems.

- ‘Name equivalence’ or ‘same-as’ type (disambiguation) services covering People Australia identifiers and other existing identifier services within the research sector, such as the Australian Access Federation, ARC, NHMRC, scholarly publishers etc.

- Interfaces and the bi-directional interchange of records with ANDS using OAI and SRU has been established.

- Nine early implementers have been harvested into production and one into test.

- A greatly improved interface to the Trove Identities Manager (TIM) (formerly named the ‘Party Administration Tool’) has been implemented along with reporting features.

- Rules for the auto-creation of records have been implemented along with refinements to the matching rules for people records. This will reduce the number of records needed to be matched by hand in the Trove Identities Manager.

- Documentation for contributors has been created and published on the ARDCPIP wiki. Internal documentation has been created to assist with operation of the Infrastructure.

- Three rounds of training have been provided to ANDS Liaison staff using the agreed train-the-trainer model.

- The project team have provided recommendations for the improvements of RIF-CS which have been tabled with the RIF-CS Advisory Group.

- The upgrade to the EAC-CPF standard was achieved along with the migration of all Party Infrastructure records to this format. The key benefit here is that the EAC-CPF standard provides for a better mapping from RIF-CS than its predecessor, EAC2004.
| Research Activity Description Infrastructure: In collaboration with the Australian research funding agencies (ARC, NHMRC, DIISR, CSIRO, UA), create software components and services that will improve the accessibility and quality of information about research activities undertaken within Australia. | • Project proposal draft submitted to ARC and has received in principle approval  
• Project expected to commence early 2012 |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Research Location Infrastructure: In partnership with responsible government agencies (GA) establish a robust national infrastructure that will allow place names to be validated by both individuals and software systems against an Australian Gazetteer (a directory that lists names of geographical place and features and includes spatial co-ordinates).</td>
<td>• The Location Infrastructure phase one project with GA and OSP (formerly OSDM) is in testing and acceptance phase and should come through to operation in early 2011-12.</td>
</tr>
</tbody>
</table>
| Infrastructure Establishment Coordination:  
• provide technical assistance and design support to those organisations building the distributed infrastructure of the ARDC  
• provide specialist technical backup in the uptake of ANDS infrastructure  
• finalise the establishment of ANDS infrastructure  
• continue to work to enable integration of ARDC Core Infrastructure with international infrastructure networks | • service management, quality, and security policies implemented  
• ANDS service desk implemented  
• change management process and system implemented  
• specialist support provided to ANDS partners  
• initial infrastructure establishment finalised  
• meetings held with JISC and NSF and several EU counterparts to promote integration |

### 7.3.5 Australian Research Data Commons Applications Infrastructure

<table>
<thead>
<tr>
<th>Proposed Activity</th>
<th>Progress</th>
</tr>
</thead>
<tbody>
<tr>
<td>In the area of data integration/fusion/merging, the ARDC project will fund AuScope through CSIRO</td>
<td>AuScope: ANDS is funding a NeAT project to develop the Spatial</td>
</tr>
</tbody>
</table>
to build new spatial data services out of existing services. The project will provide the Atlas of Living Australia with the ability to use annotation to improve data quality; and to integrate a variety of literary data sources for improved discovery. Lastly, the project will integrate a large number of sources of data associated with the coast from many disciplines to enable exploration of the data through common metadata and discovery tools.

Information Services Stack (SISS)
Atlas of Living Australia: ANDS is funding the DIAS-B NeAT project.
Coast: ANDS took part in a series of discussions and workshops to identify a possible activity related to coastal data. As yet, this has not led to anything.

In the area of data visualisation, the ARDC project will commission a tool that enables the visualisation of the connections between projects, researchers, and their research data to show the relationships between them.

This activity has not yet taken place.

In the area of data analysis, the ARDC project will initiate an activity with the University of Queensland as candidate service developer and a service provider (to be identified) to mark up existing data with a Who/What/When/Where data combination service and demonstrate new analysis opportunities on top of this marked-up data.

A report on the options for providing such a service is about to be delivered.

<table>
<thead>
<tr>
<th>7.4 Risk Register</th>
</tr>
</thead>
</table>

The key risks for ANDS in executing the Projects and the risk management strategies to be employed can be grouped into four major categories.

7.4.1 Political and Governance

Risk 1 – That there are persistent negative perceptions of the Project among funding agencies and influential groups leading to a lack of buy-in

Risk Factors:
- A particular project does not have the confidence of a subsection of a community
- Lack of confidence in governance, management, or Project delivery
- Perceptions of slow engagement with areas of the sector
- Change of emphasis with regard to the policies around publicly funded research data

**Risk Mitigations:**

- Update the communications plans to ensure that the specific e-Research communities have visibility of specific projects and their outcomes before, during and after the projects are undertaken. New diagnostic strategies have been implemented and run to mitigate against failure.
- Provide a central point where progress towards the Australian Research Data Commons can be tracked by metrics such as number of collections available, and numbers of datasets accessed
- Clearly articulate the Project’s message and brand
- Engage carefully with communities to avoid perception (or reality) of not meeting its needs
- Ensure that the Project reflects the Government’s expectations through constant dialogue
- Maintain close contact with key DIISR officers to ensure they provide input to decision making, including having an observer on the Steering Committee

**Risk 2 – That the Project is not managed effectively**

**Risk Factors:**

- Lack of effective mechanisms for planning, leadership and management
- The structure of ANDS has a negative impact on coordinated delivery of required activities
- Collaboration between the Project and across locations is not effective
- EIF funding guidelines do not allow for sufficient Project staff to administer funded programs of work
- State based staff have mixed allegiances

**Risk Mitigations:**

- Put in place management and planning processes that include formal reporting and regular reviews to ensure the efficient conduct of the Project
- Regular meetings of Project staff are held to build a team approach. Communication structures are in place to facilitate working together.
- Negotiate with DIISR on appropriate administrative staffing for EIF funded programs.
- Staffing levels are monitored and adjusted as required.
- Contracts and partnerships with state based organisations that host Project staff will be put in place that ensure that staff are clear about their role.
- Ensure that ANDS-funded staff based in organisations who are ANDS sub-contractors are not placed in a position of conflict of interest.

**Risk 3: That the increased emphasis on external contracted engagements represents too big a burden on the lead agent**

**Risk Factors:**

- University processes, focussed on student and supplier engagement, are not a good fit for “funding agency” activities. ANDS’ role as a “funding agency” in many of its programs has imposed additional requirements on the lead agent causing pressure on its staff to assist ANDS.
- ANDS EOI approach generates clusters of work with tight timelines that impact on specific university functions such as the Solicitors’ Office and Finance.

Risk Mitigations:
- Seek approval for stream-lined approaches at Monash University to enable ANDS to work more effectively.
- Fund additional staff or specific work at Monash University to meet the requirements for bursts of activity.

### 7.4.2 Relationships

**Risk 4 – That the Project's external stakeholders are not effectively engaged**

**Risk Factors:**
- Stakeholders are not prepared to undertake the changes within their own organisations that are necessary for the realisation of the Australian Research Data Commons.
- Stakeholders do not see their interests in data management and those of the Project as being aligned.
- Stakeholders might feel that the wrong decisions have been made.

**Risk Mitigations:**
- Maximise the effectiveness of connections between the Project and related PFC and other initiatives, including involvement of groups outside ANDS in the ANDS Policy Forum, the ANDS Technical Forum, and the ANDS Content Forum.
- Ensure continuing wide consultation following the consultation on the Draft Final Australian Research Data Commons Project Plan.
- Membership of the Steering Committee includes key stakeholders.
- Ensure that ANDS’ engagement with stakeholders meet their research data ambitions as well as ANDS’ requirements.
- Ensure ongoing, strong engagement with the Research Sector, including current and foreshadowed NCRIS capabilities.
- All activity plans should be highly inclusive of relevant stakeholders.
- Performance measurement for the Project should include effective stakeholder engagement.
- Effective communication of benefits to stakeholders.
- Provide a clear rationale behind the decision process for project funding.

**Risk 5 – That the Project's partners do not appropriately contribute to the Project**

**Risk Factors:**
- Partner produces outcomes of low quality or does not meet the requirements of the contract.
- Partner expends funds in a way that is not consistent with the EIF guidelines.
- Lack of effective arrangements in place to ensure the contracted services are provided to an agreed service level.
- Service providers see themselves as disconnected from the Project's decision-making or strategic planning.

Risk Mitigations:
- Provide ongoing contract management to ensure the delivery of required outcomes to the contracted service levels.
- Put in place effective vendor and partner engagement approaches.
- Implement formal procurement processes to ensure that the requirements are understood and that potential suppliers meet the set criteria.

**Risk 6: That ANDS is not perceived as a long-term partner and hence our services are not taken up**

Risk Factors:
- The impending end of ANDS NCRIS and EIF funding causes a perception that ANDS-initiated services will not continue.

Risk Mitigations:
- ANDS seeks approval to expend existing funding over longer timelines (consistent with other Superscience-funded activities).
- ANDS Steering Committee seeks additional funding for 2011-12 and 2012-13.
- ANDS creates reliable sustainable services that are offered over the longer term by other long term service providers.
- Strong contribution to DIISR Roadmap process will be a mitigating factor.

**Risk 7: That there is confusion about role of ANDS versus other related service providers in e-Research sector which impedes effective service delivery**

Risk Factors:
- ANDS and other eResearch provider offerings are confused by possible users.
- Relationship between ANDS and MARCS (such as Intersect) is not clear to users.

Risk Mitigations:
- Ensure that ANDS' communications to a range of stakeholders provide greater clarity about ANDS' services.
- Ensure that ANDS' offerings are clearly targeted and that this is clearly stated.
- Seek greater clarity from other e-Research service providers about their offerings, avoiding either actual or perceived overlap with ANDS' offerings.
- Advocate for greater coordination of offerings by e-Research service providers through eResearch Infrastructure.
- Discussion with NCI, NeCTAR and RDSI taking place to ensure clarity of eResearch service offerings.
7.4.3 Impact

Risk 8 – That data providers/federators do not make their data available

Risk Factors:
- The storage needs of researchers are not met, so they will not consider sharing their data.
- Researchers do not wish to share their research data.
- Researchers do not trust the Project’s data sharing and access control mechanisms.
- Researchers are working with other collaborators who have confidentiality concerns over the data.
- Existing data federations see insufficient value in making their data available.

Risk Mitigations:
- ANDS will co-ordinate with RDSI and Institutional stores to mitigate this risk.
- Recognise researchers through peer feedback for the deposit of data into the ARDC via increased citation – would need to be recorded and measured as a performance measure by the Project.
- Effective communication of structures in place to ensure building of trust.
- Recommend that funding be linked to the provision of data via the ARDC as it becomes available.
- Provide targeted assistance to data federations to assist with integration into the ARDC.

Risk 9 – That re-users of research data do not use ANDS Services to discover, access and exploit data

Risk Factors:
- The various strategies for exposing data in the ARDC do not result in the data being easily discoverable.
- Access control mechanisms are too restrictive or complex.
- Other sources of data for re-use are more attractive or easier to use.

Risk Mitigations:
- Ensure a nuanced and multi-faceted approach to exposing the Project’s accessible data.
- Work with the Australian Access Federation to identify a simple set of standard access control policies.
- Ensure that it is easy to re-purpose ARDC accessible data.

Risk 10: That the standards and technologies that ANDS adopts are not adopted more widely

Risk Factors:
- ANDS is the only user and maintainer of actual or de facto standards, leading to inability to share maintenance and development costs.
- ANDS is the only source of development activity on particular technologies (RIF-CS, ORCA, ANDS Handle code).

Risk Mitigations:
7.4.4 Resourcing

Risk 11 – That high quality staff are hard to recruit and retain

Risk Factors:
- Limited availability of skilled staff (both within ANDS and in ANDS-funded projects) impacts ability to perform tasks funded by ANDS.
- People with second order skills end up being employed because of staff shortages.
- Limited tenure roles potentially on offer within the Project are not attractive to candidates.
- Sustained high workload leads to staff burnout within the ANDS Management Team.

Risk Mitigations:
- Commence recruitment early to mitigate delays in the commencement of activities.
- Be highly selective in recruitment and favour quality of candidates over the quantity of candidates (do not fill jobs for the sake of it).
- Encourage secondment of staff at an institutional level.
- Investigate non-traditional sources of potential staff.
- Manage staff time and monitor levels of work.
- Engage in project based activities where more leadership roles are taken by senior ANDS staff who are not management.

7.5 Summary of NeAT Projects

Paul Coddington, September 2011

This is a summary of the eResearch tools and services developed by NeAT projects, which were funded jointly by ARCS and ANDS, with co-investment from research communities. Software developed by the projects is open source and mostly available from publicly accessible software repositories such as SourceForge and Google Code.

ASeSS: ASSDA Services for e-Social Science

Aims
The ASeSS project aims to provide to the Australian Social Science Data Archive (ASSDA) community, for the first time, a simplified, unified, national access to ASSDA datasets from a variety of sources and in a wide
A variety of data formats, e.g. unit record, data cubes, electoral, historical. ASeSS also provides a standardized platform for topic and theme-specific ASSDA sub-archives to be created in, for example, historical documents, indigenous data, and qualitative data.

**Key participants**: ASSDA and ANU Supercomputer Facility.

**Outcomes**
- ADAPT web-based curation tool – in production use by ASSDA archivists.
- ASSDA data foundation services – these are the basis for the new Australian Data Archive (ADA) service which has replaced the previous ASSDA repository and portal.
- Integrated environment for ASSDA workflows – all data processing workflows have been integrated into a single environment based on Nesstar.
- ATSIDA (indigenous data), AQuA (qualitative data), AEA (electoral data), URA (unit record data) portals and a framework for more easily developing new sub-archives
- GIS services for spatial data, with scripts for data ingest from ABS and AEC.
- The deliverable to support DDI3 has had to be dropped since third-party Nesstar software will not support it before the end of the project.

**Ongoing support**: ASSDA and ANUSF.

**Benefits**
- Replacement of legacy desktop workflow tools with web-based versions will streamline and support future data collection and curation activities across all ASSDA archives.
- Improved foundational services underpinning the ADA repository and portal provide improved and faster interface, search and access to social science data sets and easier development of sub-archives for different topics.
- ASSDA tools and services now based on a modern, scalable data and services cloud hosted at ANUSF.
- Access for ASSDA researchers to modern tools for GIS and InfoViz visualization of ASSDA datasets.

**User testimonials**
“NeAT has taken us from clunky technology and manual handling of data to smooth, automated and seamless web access.” – Prof. Deborah Mitchell, Executive Director, ASSDA.

**Aus-e-Lit: Collaborative Integration and Annotation Services for Australian Literature Communities**

**Aims**
Provide eResearch services to address the needs of researchers involved in the study of Australian literature, building on the existing AustLit web portal. New services will be developed to enable data integration and search across multiple relevant databases including AustLit, Scholarly Electronic Text and Image Service (SETIS), Australian National Bibliographic Database, Australian Digital Theses, National Library of Australia’s...
PeopleAustralia and PictureAustralia; for collaborative annotation of digital resources such as documents; and for compound object authoring, editing and publishing with a Literature Object Re-use and Exchange (LORE) tool.

**Key participants:** AustLit, eResearch Lab at the University of Queensland.

**Outcomes**

- Federated search - operational on the AustLit web site, but inclusion of some databases has been delayed waiting on the National Library of Australia’s project to federate their internal archives.
- Full-text search - operational on the AustLit web site, including tools and documentation for adding documents to the AustLit full-text search.
- Annotation and compound object authoring tool - the LORE (Literature Object Reuse and Exchange) tool incorporating secure authenticated access has been released on SourceForge and the Mozilla Firefox add-ons site.
- Compound Object applications - Black Words trails in AustLit, integrated Agent network visualisation in AustLit based on LORE explore view.
- Empirical reporting services – completed and handed over to AustLit for further development and integration into the main website.
- Improved AustLit Analytics including charts, maps, timelines and empirical reporting.
- Federated Search of several repositories (AustLit, Scholarly Electronic Text and Image Service (SETIS), Australian National Bibliographic Database, Australian Digital Theses, National Library of Australia's PeopleAustralia and PictureAustralia) via AustLit portal.

**Ongoing support:** Aus-e-Lit and University of Queensland.

**Benefits**

- Users can more quickly find compound objects by title and author with new search capability
- Provision of a new full-text search service across a corpus of hundreds of works of fiction, poetry and criticism from the early 19th century to the 1930s, enabling users to quickly locate word usage and themes within hundreds of selected works.
- The Federated Search service (beta version) enables users to quickly review and organize search results from up to nine databases in addition to AustLit.
- The Annotation Service and Compound-object Authoring Service prototype will support a number of trial projects with the assistance of a small group of researchers, raising awareness of the potential of collaborative interpretation and scholarship in literary studies.

**User testimonials**

“LORE gives us a new, easy-to-use, and terribly exciting tool for grouping together and annotating resources, and then sharing the collections that we build. Our experience with Aus-e-Lit—not just its tools, but also its developers, who listened so attentively to our needs and our feedback—has been entirely positive.” - Ms Amy Cross, Dr Michelle Dicinoski and Dr Cherie Allan, Children’s Literature Digital Resource Project, QUT.
DataMINX

Aims

Provide data access and management services for the NCRIS 5.3 Characterisation community, i.e. the users and operators of high-end microscopy and imaging, neutron source, synchrotron and X-ray facilities. The essence of DataMINX is to capture scientific instrument data and the associated metadata, deliver it to appropriate repositories in standard formats, and make it easily discoverable and accessible by researchers. The services developed are expected to build on existing work in the UK on a scientific metadata model and information catalog (ICAT) and a data transfer service for the Australian Synchrotron (the VBL Gateway Service) developed by VeRSI.

NOTE: During the reporting period, the DataMINX project was restructured and was replaced with two new NeAT projects (PCA and DTS) and one ANDS project.

Key participants: Australian Synchrotron, ANSTO, AMMRF, University of Sydney, Intersect, Monash University, VeRSI, VPAC.

Outcomes

- Comprehensive ICAT system review, and review of alternative metadata repositories and portals.
- Initial pilot ICAT deployments at OPAL and University of Sydney.
- Interface of TARDIS crystallography data portal to ICAT repository.
- Initiation of pilot metadata ingest processes from data file content and local databases initiated at OPAL, USyd AMMRF node and USyd crystallography.
- Agreement for VeRSI Gateway service to be installed at OPAL, and USyd and UWA nodes of AMMRF.
- Development of prototype Data Transfer Service to extend capabilities provided by VeRSI Gateway data transfer service. Use of Open Grid Forum standards (GFD 134 and 135) and liaison with OGF representatives including standards extensions.
- Collaborative linkages established with the UK STFC (eScience, National Grid Service, ISIS neutron source and Diamond synchrotron), Open Middleware Infrastructure Institute (OMII-UK) and UKOLN (repository and preservation expertise).

Benefits

- Gained knowledge and experience with ICAT and metadata ingest ahead of proposed deployments for AMMRF and Australian Synchrotron Facility.
- Joint software and standards development with STFC, OMII, UKOLN.

DIAS-B: Data Integration and Access Services for Biodiversity

Aims

To develop an operational metadata repository for biodiversity data resources, including registration, discovery and annotation services. The use of community annotation services is aimed at improving description and quality control for biodiversity data resources. This project will provide core services
required by the Atlas of Living Australia (ALA) NCRIS capability to support management and discovery of biodiversity data resources.

**Key participants:** CSIRO, eResearch Lab at the University of Queensland, Atlas of Living Australia.

**Outcomes**

Data Integration (DI) component:
- Feasibility study and performance analysis of alternative technology options for the ALA metadata repository.
- Data harvesting tools developed for an extended range of information sources.
- ALA metadata repository and metadata and data harvesting and integration and integration with ALA web portal (note that much of this work was migrated to the ALA development team as a core ALA activity).
- Extended functionality, better performance, and greater maturity of the repository components.

Annotation and Authentication Services component:
- Danno/Dannotate annotation tools available from SourceForge. Examples and videos demonstrating use available from Dannotate web site.
- ALA portal uses Danno annotation server but based on older version of Dannote client annotation tool. Integration of current version of Dannotate into ALA portal and other ALA tools is still to be done and is expected to be completed later this year.
- Authentication tools (Emmet) available from SourceForge, however ALA has decided to adopt an alternative open source tool (CAS).

**Ongoing support:** ALA and University of Queensland eResearch Centre.

**Benefits**
- DI component underpins the biodiversity data repository used by ALA.
- Annotation service provides ALA portal with functionality for users to annotate and raise issues with information provided by ALA.
- The Danno annotation service was also used to implement the annotation client used in the LORE tool developed in the Aus-e-Lit NeAT project.

**User testimonials**

“The ALA was keen to simplify the capture of information from users while they browse the web because they can provide different types of annotations and/or add new data records. Annotations may include information for validating and improving data, coordinates, dates and species names.” - Mr Donald Hobern, Director, ALA
MACDDAP: Marine and Climate Data Discovery and Access Project

Aims
Integrate large marine and climate data sets and to deliver these data sets through a wider range of data streams, using standard metadata, data formats and access interfaces, enabling data to be more easily discoverable, searchable and accessible to researchers.

Key participants: IMOS, University of Tasmania, Tasmanian Partnership for Advanced Computing, CSIRO, Bureau of Meteorology, Arcitecta, OPeNDAP Inc.

Outcomes
- Enhancements to the GeoNetwork Metadata Entry and Search Tool (MEST), including addition of OGC Web Catalog Service (WCS) capability. New MEST is used by IMOS eMII and Australian Ocean Data Network (AODN), and incorporated into the GeoNetwork codebase.
- Enhancements to the TPAC Digital Library including an improved web portal, administrator interface and data harvester. Software also deployed at Bureau of Meteorology (BoM).
- Data aggregation harvester and aggregator completed and deployed by IMOS and AODN.
- Add support for OGC Web Map Service (WMS) to THREDDS OPeNDAP server. Included in version 4.0 of THREDDS and deployed by IMOS, BoM and TPAC Digital Library.
- Secure authenticated data access with OPeNDAP using Hyrax server. Work has been completed and improvements will be included in the next release of Hyrax later this year. It will then be deployed at the BoM to allow authenticated access to BoM data sets.
- Translation Services to convert data files to conform to standard formats and metadata specifications. Deployed at TPAC for the TPAC Digital Library.

Ongoing support: IMOS, AODN, TPAC, CSIRO, BoM. GeoNetwork, THREDDS and Hyrax are international community open source projects.

Benefits
- Enhanced functionality and improved user interfaces for portals that are currently in regular use by marine and climate related research communities, including federated search across multiple data sets, aggregation of resultant data sets, improved metadata and geo-spatial search.
- Adding support for international standards for data formats, metadata and data access.
- Improved tools for data providers to support standards for metadata, data formats and web services.

User testimonials
“MACDDAP has enabled several components of the IMOS and AODN infrastructure to rapidly progress both the robustness of the system and efficiency of the information collation, discovery and access.” - Dr Roger Proctor, IMOS eMII Director.
“I have been using the TPAC data portal to retrieve data for the last 6 years. It has been extremely helpful to my work to have the datasets I need stored in a central location and easily accessible. The new interface you’ve created is much more user friendly and retrieving data is now faster. In particular, I found it very useful to be able to search for datasets through the interactive map or using keywords.”

- Dr Helen Phillips, Institute for Marine and Antarctic Studies

“The new TPAC portal front end created by the MACDDAP project has made my task of adding new datasets or updating existing ones much easier and faster. While we have web services and data portals that are getting more sophisticated, their utility is severely limited by the general lack of metadata. The translator has greatly simplified my job of filling metadata holes in the datasets we are serving.”

- Dr Paola Petrelli, TPAC

**SISS: Spatial Information Services Stack**

**Aims**

Develop some of the component services and functional capabilities needed to realise a spatial information data commons within Australia, leveraging Open Geospatial Consortium and ISO standards, and to work with the providers and consumers of data to assist them in adopting the common spatial data access and exchange mechanisms supported by the Spatial Information Services Stack (SISS).

**Key participants:** IMOS, University of Tasmania, Tasmanian Partnership for Advanced Computing, CSIRO, Bureau of Meteorology, Arcitecta, OPeNDAP Inc.

**Outcomes**

- SISS software stack - All the components are functionally complete for the project requirements and the first release of a packaged SISS and associated documentation has been made available. Components include:
  - Web Mapping Service (Deegree and Geoserver)
  - Web Feature Service with application schema support (Degree and Geoserver)
  - Coverage services (THREDDS, Hyrax, ERRDAP)
  - Registry service (GeoNetwork) including RIF-CS harvester for ANDS
  - Information modelling tools and services (Full Moon, Hollow World)
  - Vocabulary service
  - URN resolution service
  - SISS deployments at several organisations including most of the state geological surveys, Geoscience Australia, AuScope, CSIRO, WA Geothermal Centre of Excellence, WA Centre of Excellence for 3D Mineral Mapping.
Various portal implementations are now consuming SISS service instances, including the AuScope Discovery Portal, which uses all components to source data from multiple government organisations and research groups who have deployed the stack.

Other portals utilising SISS in either production or development environments include: Virtual Rock Laboratory (UQ), Geodesy Workflow, CSIRO MDU Earth Model Portal, the Bureau of Meteorology, and CSIRO Marine and Atmospheric Research.

Various research groups have also deployed parts of SISS in support of managing and providing interoperable web access to their data, including WA Geothermal Centre of Excellence, WA Centre of Excellence for 3D Mineral Mapping, AuScope Earth Imaging (ANU Research School of Earth Sciences & Geoscience Australia), CSIRO node of AuScope National Virtual Core Library.

**Ongoing support:** All the software components of SISS are international community open source projects. Deployments of SISS are supported by the organizations hosting them.

**Benefits**

- Standard approach and software stack available for making spatial data available online.
- Many geospatial data sets from many organisations now available online from via OGC standard web interfaces, with data discoverable using metadata standards.
- AuScope Portal utilizing the SISS provides access to geoscience information holdings at multiple organizations.

**User testimonials**

“This promises Australia a wider instantaneous international reach for its geoscience information with resultant increases in mineral exploration investment expected to follow.” - Graham Butt, Geological Survey of NSW and Chair, Government Geoscience Information Committee.

“A big challenge in discovering new deposits of minerals and natural resources like base metals, hydrocarbons and rare earth minerals along with diamonds and gold is that the continent has evolved and deformed significantly since the time the deposits were formed and the present day. Software tools when connected to the SISS infrastructure can now reconstruct how the continent evolved in space and time, allowing us to develop much more accurate exploration models.”

- Dr Thomas Landgrebe, University of Sydney

**Aus-e-Stage: Collective Intelligence and Collaborative Visualisation for Creative eResearch**

**Aims**

AusStage is an online research facility for investigating live performances in Australia. The Aus-e-Stage project will develop two new visually interactive services for exploring information in the AusStage.
database. It also will create the capability to generate a new data set of immediate, on-location responses from spectators of Australian performing arts.

**Key participants:** Flinders University, AusStage partners.

**Outcomes**

- Mapping service - an interactive interface with which to search, manage and chart the geographic distribution of performance events. Also developed KML and KMZ overlays for historical maps and ABS data.
- Navigating networks service - a web interface for navigating and analysing the network of artistic collaborations embedded in the AusStage data set. Users can also download data or interest for visualization using desktop graph visualization tools.
- Research Audiences service – gathers audience input from SMS, Twitter and a web form.
- All these services are now available from the AusStage website, but further work will be done later this year to integrate them with the search interface of a new version of the AusStage portal that is currently under development.

**Ongoing support:** AusStage, Flinders University.

**Benefits**

- Performing arts researchers can better explore and visualize the data in the AusStage database.
- Theatre companies and researchers have a mechanism for accessing audience responses to performances.

**User testimonials**

“Being able to visualise and appreciate working relationships and influences between practitioners will provide greater insight into the way work is created and developed. It can provide a visual statement to support assumptions.” - Margaret Leask, Researcher/Historian, National Institute for Dramatic Arts (NIDA)

(The Mapping Events service will be used for) ... “finding answers to questions or research queries. It’s a more direct and superior way to research using the AusStage Database.” - Ross Bruzzese, Librarian, NIDA

**AusCover Workflow: Workflow Services to Enable a Large-Scale Temporal-Spatial Ecosystem Digital Information Service**

**Aims**

The AusCover component of the NCRIS Terrestrial Ecosystems Research Network (TERN) capability is focused on organising remote sensing data sources and products for terrestrial ecosystems research. AusCover will enable, for the first time in Australia, the storage of these data sets online in a form that makes them directly accessible to the user community. This project aims to provide easy-to-use workflow tools and services that enable researchers to process AusCover data sets using the ARCS grid (or cloud) computing infrastructure or other HPC resources that they have available. The intent is that the tools will hide the complexity of the underlying processing environment behind a straightforward user interface. The same
workflow tools will also assist AusCover data providers, allowing them to more easily process raw satellite data.

**Key participants:** TERN AusCover, Curtin University of Technology, CSIRO Marine and Atmospheric Research, Centre for Comparative Genomics at Murdoch University.

**Outcomes**

- Remote sensing data processing workflow system (RS-YABI) has been created, based on the YABI workflow tool (from the NeAT BioFlows project).
- Workflow service deployments – pilot system deployed at IVEC compute facility at Murdoch University, production deployment at the NCI, production deployment at IVEC will be competed later this year.
- Software modules for standard processing functions for remote sensing data sets (MODIS and AVHRR) developed for RS-YABI.
- Workflows developed to generate standard data products from MODIS and AVHRR data sets.
- Some exemplar customised application workflows developed, for identification of smoke and dust and analysis of their effects.

**Ongoing support:** TERN Auscover, CSIRO, WASTAC, hosting at NCI and IVEC.

**Benefits**

- Workflow system to enable remote sensing data providers or consumers to easily develop custom workflows.
- Automated tools for generating standard remote sensing data products for TERN AusCover and IMOS, utilizing existing compute facilities either directly or through the ARCS Grid interface.
- Researchers and state agencies can more easily analyse remote sensing data using grid and high-performance computing facilities.

**User testimonials**

“Not only does YABI provide users with simple access to large data storage capacity, it also allows users to utilise the immense processing capacity of supercomputing facilities. Processing tasks that might require months of processing time on standard desktop computers may be run in a few hours on the national super computing facilities”.

- Dr. Fears, Remote Sensing and Satellite Research Group (RSSRG) at Curtin University

“This system is improving the efficiency of research in the field immeasurably by simplifying access to data. Anyone working with remote sensing data can access and understand the information and start using it immediately.” - Dr Edward King, CSIRO Marine and Atmospheric Research.
Bioflows: Bioinformatics Workflows

Aims
The Bioflows project aims to provide a simple Web-based workflow tool that enables life sciences researchers to specify genomics and proteomics workflows that can be executed on the ARCS Compute Cloud and interface with the ARCS Data Fabric. The system will be deployable as an “appliance,” with required software, middleware and server hardware able to be installed at a site and managed remotely, if required. The appliance can interface with local high performance computing systems and/or submit compute jobs to the ARCS Compute Cloud. The appliance concept will be tested with trial deployments at the iVEC Informatics Facility at Murdoch University, the Queensland Facility for Advanced Bioinformatics and the Life Sciences Computation Centre in Victoria.

Key participants: Centre for Comparative Genomics at Murdoch University, Queensland Facility for Advanced Bioinformatics, Victorian Life Sciences Computation Initiative.

Outcomes
- Bioinformatics workflow service – provides a simple web-based interface for specifying bioinformatics workflows and running them on the grid or local high-performance computing facilities.
- A workflow server appliance that can be easily installed at a site and remotely managed or updated if required.
- Deployment of the YABI bioinformatics workflow service at the IVEC@Murdoch facility, and preliminary deployments at QFAB and LSCC. Further work is needed to provide production deployments to users at QFAB and LSCC, which is to be completed later this year.

Ongoing support: Centre for Comparative Genomics at Murdoch University, IVEC, QFAB, VLSCI.

Benefits
- Will soon have a prototype platform for supporting an easy-to-use web-based tool for specifying bioinformatics workflows and have them submit jobs using the ARCS grid interface to remote compute resources.

User testimonials
“YABI has radically transformed the way we process and analyse second generation DNA sequence data. Though a user-friendly dynamic HTML interface we can design a simple pipeline that sorts, trims and queries sequence data against databases – we put raw data in one end and obtain meaningful outputs at the end of the pipeline. Put simply, YABI has enabled us to spend more time analysing the actual data and less time ‘fighting’ it.” - Dr Michael Bunce, ARC Future Fellow, Ancient DNA Lab, Murdoch University

30 September 2011

ands.org.au
“Through this user friendly resource, we can share, analyse and store project data prior to public release. YABI is an enabling tool for scientists and students that has become essential to all of my research programs. Partnership with the CCG and access to YABI obviates the need to establish bioinformatics resources within my group.” - Dr Ala Lew-Tabor, Queensland Alliance for Agriculture & Food Innovation, University of Queensland

BioSecurity Collaboration Platform

Aims

The aim of this project is to implement a biosecurity collaboration platform (BCP) at CSIRO’s Australian Animal Health Laboratory (AAHL) to assist in the flow of complex information across the containment barrier from a variety of data sources including pathology and microscopy systems, live in-vivo animal experimental data (e.g., heart rates) and data from simulation models and historical information in both visual and written form. The collaboration platform will provide video conferencing and shared displays, as well as integrate critical technology platforms (microscopy, pathology, software applications) into a common, shared visual workspace. Following the successful implementation of the collaboration platform at the AAHL facility, additional nodes will be implemented within state government departments that collaborate with AAHL on biosecurity research and emergency response. This platform is expected to have broader applicability within the National Collaborative Research Infrastructure Strategy (NCRIS) Australian Biosecurity Information Network (ABIN).

Key participants: CSIRO, Australian Animal Health Laboratory, Biosecurity Queensland, Australian Government Department of Agriculture, Fisheries and Forestry.

Outcomes

- A high-resolution multi-site video collaboration platform providing application and data sharing, including information from instruments.
- Deployment of the BCP at the Australian Animal Health Laboratory (AAHL), on either side of the bio-containment barrier.
- Deployment at DAFF in Canberra and CSIRO Coopers Plains (QLD) for use by Chief Veterinary Officers (CVO) of Australia and Queensland.

Ongoing support: CSIRO, AAHL, DAFF.

Benefits

- Enables AAHL staff to meet, collaborate and share information more easily across the bio-containment barrier.
- Provides a pilot system for national meetings, research collaboration and biosecurity emergency response. The aim is to roll out the BCP for use by all agencies involved in the Consultative Committee on Emergency Animal Disease that provides national biosecurity emergency response.
User testimonials

“We are now able to ‘meet’ with colleagues and share information such as pathology records, high resolution maps and even view images on a microscope in real-time, easily and quickly.

When dealing with disease outbreaks, time is of the essence. A high speed, real-time interactive system will greatly improve the management of emergency animal disease outbreaks in Australia.”

- Dr Alex Hyatt, research leader, Australian Animal Health Laboratory

Data Transfer Service

Aims

Provide a service that enables researchers to easily transfer data files between a variety of end points supporting different repositories and protocols, including instruments generating data (such as NCRIS Characterisation facilities), data repositories (such as the ARCS Data Fabric or institutional repositories), data processing facilities (such as the ARCS Grid), or the user’s desktop. Transfers should be specified easily by a user through a web portal or by an automated program via an API, and the system should support immediate or scheduled transfers.

NOTE: This project was unable to progress beyond some initial work as part of DataMINX. The project was to be led by the ARCS Data Services team with the service supported by ARCS, but this became infeasible due to the ARCS transition that meant no development or deployment of new services could occur during the transition period.

Key participants: ARCS, University of Sydney, VeRSL, UK STFC.

Outcomes

- Requirements gathering from Characterisation facilities and ARCS Data Team.
- Design and implementation of a prototype Data Transfer Service backend (portal still to be developed) in collaboration with UK STFC and UK National Grid Service.
- Use of Open Grid Forum standards (GFD 134 and 135) and liaison with OGF representatives including standards extensions.

Benefits

- Joint software and standards development with STFC, OMII, UKOLN.
- Will provide a useful service to users that can be supported by ARCS, NeCTAR or RDSI.

Human Variome: Software and Data Support for the Australian Node of the Human Variome Project

Aims

The Human Variome project will create a national data repository called the Australian Human Variome Database (AHVD). The database will hold and provide access to information on genetic variations associated with human disease that have been characterised by Australian laboratories and clinics. The project will
develop services to enable submission of laboratory and clinic data to the AHVD using existing organisational workflows.

**Key participants:** Genome Disorders Research Centre, VPAC, BioGrid, Mawson Project (University of South Australia and partners), NICTA.

**Outcomes**
- Database and portal for the Australian Node of the Human Variome Project is now available.
- Data sets from three laboratories (in QLD, NSW and SA) and three different genetic diseases are being ingested into the database. Other sites are expected to be added in future.
- Software for the repository, portal and data export from data providers has been developed and is expected to be utilized by other countries participating in the international Human Variome Project.
- Ethics applications for site-specific, state-wide and national approval have been developed.

**Ongoing support:** Human Variome Project (an international project led from the Florey Institute and University of Melbourne), VPAC, data providers (genetic clinics and labs).

**Benefits**
- Provides easy access to a data set of genetic variations and associated clinical information for clinicians and researchers.
- Provides an exemplar for other countries involved in the Human Variome Project.

**User testimonials**
“A regularly updated set of disease causing breast and colon cancer mutations identified in all Australian diagnostic laboratories is a unique resource. Diagnostic laboratories believe this will encourage new opportunities for research enabling individuals in different families with identical mutations to be investigated.”
- Dr. Valentine Hyland, Supervising Scientist, Pathology Queensland.

“For bowel cancer, the curation of the international database held by InSiGHT is now located in Melbourne. In curating the database, the issue of variants of uncertain significance has become paramount due to their sheer number amongst the 5000 unique variants entered worldwide. The Australian node is one important way we, in Australia, can contribute our experience to the effort to inform the clinical significance of these variants - for Australians and the world wide community of these gene carriers.” - Finlay Macrae, Head, Colorectal Medicine and Genetics, Royal Melbourne Hospital.

**NCIRDN:** National Criminal Justice Research Data Network

**Aims**
This project aims to increase the quantity and quality of Criminal justice System research that can be undertaken using administrative data within the Criminal Justice System research community, and reduce
the barriers to accessing that data. It is also has the aim of promoting cross-jurisdictional research and collaboration.

**Key participants:** Crime Research Centre at University of Western Australia, Australian Institute of Criminology, NSW Bureau of Crime Statistics and Research, Griffith University, University of Melbourne, Australian Social Science Data Archive.

**Outcomes**

- Data Availability Register has been compiled, identifying available data sets within state government agencies. Data sharing agreements have been specified with some agencies.
- Implementation of a criminal justice data repository and portal as a sub-archive of the Australian Data Archive hosted by ASSDA.
- Curation and ingest of various agency data sets into the ADA sub-archive.
- A website to promote the project with some example data analysis tools.

**Ongoing support:** The NCJRDN consortium of state-based criminal justice research organisations, ASSDA, hosted at ANU-SF.

**Benefits**

- Identification of available data sets will assist researchers to find the data they want.
- Criminologists and sociologists have easy access to data sets from multiple jurisdictions.
- Exemplars of data analysis tools.

**User testimonials**

“It was a privilege to be in at the ‘birth’ of the NCJRDN – I expect the Network will rapidly become a key repository, if not the key repository of criminological data in Australia.”

- Dr Adam Tomison, Director and Chief Executive, Australian Institute of Criminology

**PCA: Platforms for Collaboration in the Australian Microscopy and Microanalysis Research Facility**

**Aims**

Provide services that assist researchers to use microscopy, microanalysis, electron and X-ray diffraction and spectroscopy. The project will provide a web based Technique Finder that enables users to seek out the most appropriate technique or instrument that will assist their research. The project will also develop a Data Management System (DMS) that will support data and metadata capture (with an associated metadata catalogue) from instruments and data transfer between nodes and to federated data repositories.

**Key participants:** AMMRF Sydney node, Intersect.

**Outcomes**

- Technique Finder - deployed and accessible from the Australian Microscopy and Microanalysis Research Facility (AMMRF) web site.
Data Management Service – software development completed. Support for data ingestion for microCT and atom probe instruments, additional instruments to be supported through an ANDS data capture project. The DMS has been installed at the University of Sydney node of the AMMRF.

**Ongoing support:** AMMRF, University of Sydney.

**Benefits**
- The Technique Finder allows researchers to match their requirements to the most appropriate techniques or instruments provided by the AMMRF, with information about their use.
- The DMS allows researchers to store and manage their data, with associated searchable metadata.

**User testimonials**
“The AMMRF offers a complete user experience involving stages of project registration, planning and training followed by data gathering, analysis, management and publication. The Technique Finder improves this experience by enabling researchers to identify, access and apply appropriate microscopy techniques quickly.” - Prof Simon Ringer, Executive Director of the AMMRF

“The DMS improves the manner in which we manage data. It allows us now to catalogue our images and data sets in a manner which will be searchable by others in the future and unlocks the potential for re-use of that data in years to come.”

- A/Prof Allan Jones, Australian Centre for Microscopy and Microanalysis.

**PODD: Phenomics Ontology Driven Data Management**

**Aims**
The Integrated Biological Sciences component of NCRIS contains two major Phenomics initiatives: the Australian Plant Phenomics Facility and the Australian Phenomics Network. These facilities have common requirements to gather and annotate data from both high and low throughput phenotyping devices. The PODD project will deliver a data management service that can handle multiple phenotyping platforms and data formats (text, image, video). The project will also provide the ability to manage a repository of associated metadata based on standard ontologies. A range of tools and other features will be developed to provide Web-based discovery interfaces for users, external repositories, and services and support for the automatic capture and annotation of data and metadata from instrumentation, when possible. The project also will enable the facilities to publish data or make it publicly available after a pre-determined period.

**Key participants:** Australian Plant Phenomics Facility, Australian Phenomics Network, CSIRO, Australian National University, University of Queensland, University of Adelaide, Monash University.

**Outcomes**
- PODD repository and software. While this project is focused on phenomics data, PODD can be used for any type of data if an appropriate ontology has been defined.
- Base ontology for phenomics defined, with extensions for plant and mouse phenomics data.
- Deployment for the High Resolution Plant Phenomics Centre (HRPPC) at CSIRO.
- Deployment for the Australian Phenomics Network (APN) at ANU.
- System for exporting data from The Plant Accelerator (TPA) repository into the PODD plant phenomics repository.

**Ongoing support:** HRPPC/CSIRO, APN/ANU, TPA, University of Queensland eResearch Lab, ALA.

**Benefits**

- Repository for many different types of experimental data used in phenomics research, with associated contextual metadata.
- Researchers can discover and access relevant phenomics data sets.
- System can support arbitrary ontologies so could be reused for other discipline areas.

**User testimonials**

“The NeAT funded PODD project is an important component to the Australian phenomics infrastructure, and will be an effective tool for research data management. It gives Australian researchers the ability to manage the data and contextual metadata generated from mouse models, and allows these researchers to collaborate around their data and publish it online.” - Adrienne McKenzie, Deputy Director of the Australian Phenomics Facility, Chief Operating Officer of the Australian Phenomics Network

“The NeAT grant has effectively enhanced our institution’s ability to investigate plant phenomics and find practical solutions to many of the confounding problems that will affect our generation, our children and grandchildren well into the future.”

- Dr Robert Furbank, Director of the High Resolution Plant Phenomics Facility.

**Remote CT: Remote Computed Tomography Reconstruction, Simulation and Visualisation**

**Aims**

The main focus of the Remote CT project will be to develop a three-part service for 3D reconstruction and visualisation of Computed Tomography (CT) images. The service will be deployed at the Imaging and Medical Beamline at the Australian Synchrotron and the ANU micro-CT facility. It is expected to be applicable to other tomography facilities. One component of the service will enable researchers to use remote high-performance computers for CT reconstruction, analysis and simulation that require large memory and compute power. Another component will provide 3D data visualisation of data sets and allow seamless transfer of the resulting images to a user’s desktop. This functionality will be useful for preliminary analysis of raw and reconstructed CT data, as well as for producing final publication and presentation quality images. It also will allow remote collaboration between different teams participating in CT experiments. A third part of the project will enable secure and efficient transfer of large CT data sets to and from the facilities, providing users with a transparent and user-friendly interface.
**Key participants:** CSIRO, Victorian Partnership for Advanced Computing (VPAC), the Victorian eResearch Strategic Initiative (VeRSI), the Australian Synchrotron, Australian National University.

**Outcomes**

- Remote CT software – the X-TRACT software for computed tomography reconstruction supports remote execution and parallel computation using GPUs to enable real-time reconstruction. The software can handle very large data sets, at least 4Kx4Kx4K voxels.
- A framework and software library to enable porting of Windows application software to a client-server remote execution model.
- Deployment of the Remote CT Reconstruction, Simulation and Visualization services on the MASSIVE-0 cluster at the Australian Synchrotron, a small cluster hosted at the ANU-SF for as ANU micro-CT facility, and the CSIRO TBI and GPU clusters.
- Port of most of the Windows-based X-TRACT code to Linux, including Windows and Linux job schedulers that can submit XLI jobs to a Linux cluster.
- Improvements to the Drishti 3D visualization software to support larger data sets, different input file types and improved performance.
- Two variants of Remote 3D Visualization service have been developed. Both solutions utilize either Drishti or ParaView software running on dedicated remote visualization servers under Linux, but using different remote execution platforms.
- Remote file upload and download service implemented at the AS, based on the VBL infrastructure developed by VeRSI. These services have been integrated both with Remote CT Reconstruction and Simulation service and the Remote 3D Visualization service. A remote file browsing functionality has been also developed, integrated with VBL and incorporated into the downloadable XLI/X-TRACT client software.

**Ongoing support:** Australian Synchrotron, CSIRO, ANU.

**Benefits**

- System enables users to easily access, transfer, process and visualize their computed tomography data remotely. Desktop processing and visualization is infeasible for very large data sets.
- Real-time reconstruction enables checking and modification of data collection during the experiment.

**User testimonials**

“In my research work I have collected many extensive X-ray imaging and tomography data sets using both laboratory and synchrotron sources. The availability of the software package X-TRACT has proven to be essential to my ability to pre-process, process and analyse these large data sets in a reliable and timely manner.” - Dr Andrew Stevenson, Research Team Leader, CSIRO
“Our testing of the Remote CT reconstruction facility proved the great value of using X-Tract running on a cluster. We were able to capture biomedical CT image sets, reconstruct, and then examine the images during the same session.” - Dr Chris Hall, Beamline Scientist, IMBL, Australian Synchrotron