

NCRIS Capability 5.16

Progress Report

Platforms for Collaboration

Facilitator: Rhys Francis

Date: 14 November 2006

Contents

1	Introduction	3
1.1	Context	4
1.2	Reference group and web site	5
1.3	Consultations	5
1.4	Service providers	5
2	National e-Research Infrastructure	6
2.1	Overview	6
2.2	NCRIS Investments	7
3	Progress with consultations	8
3.1	High Performance Computing	8
3.2	Advanced Networking	8
3.3	Expertise	9
3.4	Authentication and authorisation services	10
3.5	Data management services	11
4	Outline of Investment Structure	14
4.1	Overall Mission	14
4.2	Context	14
4.3	Potential Activities	15
4.4	Budget Issues	15
5	Project Plan	16
6	Expenditure to date	18
	Appendix I Reference Group	19
I.1	Membership	19
I.2	Terms of reference	19
	Appendix II Investment Principles	20
II.1	General Comments	20
II.2	Specifics	20
	Appendix III Research Information Infrastructure projects	21
	Appendix IV AARNet	43
	Appendix V APAC, the National Facility, the National Grid	48
	Appendix VI State based APAC Partners	52

1 Introduction

Modern research is increasingly dependent on technological platforms that enhance the research community's ability to generate, collect, share, analyse, store and retrieve information.

Some research can only be progressed because of the capabilities provided by these ICT platforms.

While termed "platforms for collaboration" in NCRIS, they are also described using the term "cyber-infrastructure"; and their application within research activities, and the consequential change in the way research may be carried out, is often referred to as "e-Research".

Notably, both e-Research and cyber-infrastructure are areas of rapid development, both in technology and social terms; a situation likely to continue for the foreseeable future. As a result we can expect an increasing pace of change and an ongoing flow of new opportunities to enhance the quantity, quality and productivity of research efforts; noting always that some research is otherwise impossible and that the improvement in infrastructure and the ability to ask more demanding questions go hand in hand.

Within this context, five key areas were identified within the NCRIS roadmap¹:

"Platforms for collaboration include the following sets of inter-related components:

- *Data storage management, access, discovery and curation* to improve interaction and collaboration;
- *Grid enabled technologies and infrastructure* to enable seamless access to the facilities and services required in various research fields;
- *Support skills* to assist researchers in developing and using this infrastructure effectively;
- *High performance computing* to allow analysis, modelling and simulation; and
- *High quality network access through high capacity bandwidth* to permit interaction with diverse data and computing resources."

The consultations for 5.16 and the other NCRIS capabilities confirm these as critical capabilities.

While undertaking the consultations, several statements of longer term intent (or goals) for an Australian national e-Research infrastructure have been suggested, statements such as:

- All networked research resources should be accessible through single sign on
- A common electronic access and authorisation framework should apply to all research data
- A common high speed network should connect all researchers and all research resources
- "Peta" scale (and beyond) capabilities should be available and made accessible

With visions of such scale, a number of issues present themselves within the context of planning within the NCRIS Platforms for Collaboration activity. Most notably:

- Institutions within the system will spend on the same activities, and spend far more than NCRIS, so that a clear understanding of the NCRIS role within e-Research infrastructure is important
- Underpinning components, such as researcher authentication and e-Research toolkits may also need to be considered for investment to ensure the infrastructure is easier to adopt and apply
- The inter-relatedness of components means that priority cannot be simply addressed by selecting some needs over others, a funding distribution will need to relate to some weight given to each need, and the corresponding readiness of solutions, technologies and associated products
- The state of readiness is also different across communities suggesting that support for adoption as well as delivery may be needed, and that the distribution of funding might change over time
- The funds available to NCRIS are expected to be insufficient to meet the overall need, so that extensibility of the investment structure is important to allow additional funds to be applied

¹ National Collaborative Research Infrastructure Strategy, Strategic Roadmap, DEST, February 2006.

1.1 Context

The context for the investment in Platforms for Collaboration is global in scale, and a variety of initiatives around the world provide guidance.

Broadly speaking this ground work has been recently and expertly covered by the e-Research Coordinating Committee (eRCC), some details of which are available through its discussion paper², which says, inter-alia:

“Successful research is increasingly team-based. It is also increasingly necessary for research to be carried out across disciplines and across geographic boundaries, as researchers attempt to address more complex issues where boundaries are less relevant.

Developments in information and communications technologies (ICT) are enabling large amounts of data to be manipulated and transferred very quickly across long distances on advanced networks. Developments in ICT are also changing research methodologies and enabling formerly inaccessible problems to be addressed. And in both ways, ICT developments are enhancing our ability to approach complex problems.

A number of countries are taking significant steps to enable researchers to utilise advanced technologies. Many researchers in Australia, whether in universities, research organisations or industry, have access to advanced computing and network capabilities linking places in Australia and overseas. However, as ICT develops very quickly, Australian researchers will require specific support to exploit the full potential to share data and information and work more collaboratively. “

and

“Status of Australia's e-Research Infrastructure

A focus on excellence is critical to ensuring internationally competitive outcomes from research and the high returns that can arise. The nature of excellent research is increasingly characterised by collaboration on a national, and more often, a global scale.

To conduct top quality research, Australia needs to maintain a world competitive electronic research infrastructure (e-Research Infrastructure) and have researchers who can use that research infrastructure in an increasingly sophisticated way.

...

Australia's strategic investment in e-Research infrastructure is consistent with similar initiatives in leading research communities, such as cyber-infrastructure in the US, e-Science in the UK, European e-infrastructure and GRID Canada.“

The eRCC committee recommended a range of activities should be supported, only some of which might meet NCRIS requirements as ‘infrastructure’ and which were intended in any event to be in addition to infrastructure investments. Therefore while 5.16 will ensure its investments align with the directions and intent set out by the eRCC, the activities the eRCC proposed will need separate funding.

As well as the views developed and represented by the eRCC, the development of the 5.16 investment plan intends to align with international developments and be explicitly informed by:

- the needs of NCRIS investments which are summarised in a companion report from the facilitator³
- the discussions and recommendations of the PMSEIC working group on Data for Science
- consultations in data management, authentication and authorisation, and e-Research toolkits
- reviews and inputs from the existing networking, high performance computing and grid service providers

² An e-Research Strategic Framework, A Discussion Paper, DEST June 2005

³ Investment Plans and Platforms for Collaboration, DEST, October 2006 (see pfc.org.au)

1.2 Reference group and web site

A reference group has been established and has met three times, and is expected to hold a further three full day meetings over the remaining course of the planning activity. The terms of reference and membership is given in Appendix I. A number of email lists and a web site have been established to support the development of the investment plan (see www.pfc.org.au).

1.3 Consultations

The investment plan is being informed by a number of specific consultations and several activities which were initiated separately to the 5.16 consultations, as detailed below.

The report of the e-Research Coordinating Committee	Provides an assessment of the expertise and assistance research groups need to effectively adopt the developing e-Research infrastructure
PMSEIC Data for Science working group	Provides an agreed strategic view of relevant issues and the associated systemic requirements
NCRIS investment plans	Provides requirements mostly related to data management and grid based collaboration systems
DEST review of APAC	Provides an assessment of the need for supercomputing, grid deployment and data transport services, and the role APAC might play in delivery of those

Documents related to these are available separately, although the PMSEIC Data for Science working group report will not be available until after PMSEIC meets in December.

Discussions related to the NCRIS investment process have led to a set of principles for 5.16 investments which are provided in Appendix II.

Early work also identified a range of additional issues which would need to be explored. These relate to the support needed within a national framework for research collaboration that permits:

- Institutions to easily recognise each other's research staff
- Facilities and data to be easily accessed, while supporting appropriate security, privacy and confidentiality requirements
- Researchers to easily locate and access resources relevant to their research goals

Broad based consultations are therefore in progress as follows (see pfc.org.au for more details):

- AAA – to identify a trust federation that could be permanently established and operated to support identification for research as well as other higher education institutional purposes
- Data management – to identify the categories of data and data management missions, so that systemic requirements can be identified and a role for NCRIS investment agreed
- e-Research toolkits – to identify the state of play in the tools and technologies that researchers might deploy to take advantage of e-Research infrastructure (from the desk top)

1.4 Service providers

In addition to formal interactions with AARNet and APAC, additional investigations are planned or already in progress with other participants around their ability and interest in supporting relevant services and service delivery requirements, including:

- National institutions: the Bureau of Meteorology, CSIRO, ANU, Geosciences Australia and the National Library of Australia
- Regional institutions: the state based APAC partners, research intensive institutions, and e-Research activities; to ascertain the extent of existing services and development plans
- Other service providers: such as Aus-cert and a variety of "collections" agencies

2 National e-Research Infrastructure

2.1 Overview

At present, e-Research infrastructure service delivery is dominated at the national level by AARNet (expending circa \$40M pa) and APAC (expending circa \$20M pa).

However, this picture misses the interests of other key participants. For instance:

- The Bureau of Meteorology (BoM) and CSIRO jointly invest more than \$10M pa in HPC and data services within the High Performance Computing and Communication Centre, independently of APAC, and both invest substantially in related infrastructure outside the HPCCC
- Many of the largest data holders in the country, Geoscience Australia, BoM, many state and federal research agencies (such as in the primary industries, marine and health areas), and the humanities overall, are not perceived to be connected to a common national e-Research infrastructure
- Substantial and ever increasing data management capabilities exist around the country, including at least five petabyte scale near-line storage systems, and nearly all future research data will be born digital, and yet a co-ordinated approach to national data management has not arisen
- Several regional research networks now exist, creating a federated network model
- DEST initiated Research Information Infrastructure projects representing an annual spend of \$17M (average over last three years) are only weakly connected with the service providers and there is insufficient support for transitioning developments into products and services
- A shibboleth trust federation has been created but with uncertain future while significant nation wide infrastructures, such as the APAC National Grid and CSIRO's enterprise management services, remain separate to this trust federation
- Many state governments intend to invest in e-Research activities and yet the co-ordinated development and delivery of services against the needs of research groups lacks a home
- The APAC state partners spend an additional \$20M pa on activities historically identified as outside APAC's core interest, activities which then remain uncoordinated at the national level
- NCRIS capabilities themselves will spend an estimated \$10M pa on e-Research infrastructure within the various capabilities, with no available means of co-ordination
- LIEF grants are in excess of \$30M pa with some component in e-Research infrastructure, and Universities themselves invest in eResearch infrastructure, all of which is carried forward independently and without strong coordination

These points appear to show a system developing in piecemeal fashion, with weak co-ordination and few anchor points for research communities; there is certainly no 'one-stop shop' for developing and supplying e-Research solutions to meet the needs of a research community or a research facility.

To provide a view across the national space, some capability and summary statements are appended:

- Appendix III provides a summary of the current Research Information Infrastructure projects
- Appendix IV contains a summary capability statement for AARNet
- Appendix V contains capability statements from the state based APAC partners
- Appendix VI contains a summary of APAC, the National Facility and the National Grid

Overall, awareness of the need for co-ordination and the importance of standards is growing. The policy barriers to collaboration and the impact of uncoordinated investment are less well understood.

Finally, the aggregate e-Research infrastructure expenditure in just the entities named above exceeds \$100M pa, and of course all research institutions manage major (and in aggregate much larger) ICT budgets, the components of which need to combine with and become part of the overall e-Research infrastructure.

Thus, as the vision for e-Research infrastructure hinges on inter-operability, universal access, coverage, and strategic coherence; investment priorities for Platforms for Collaboration will need to support improved leadership and improved co-operation as well as providing key shared services.

2.2 NCRIS Investments

A detailed report is available⁴ covering the implications for platforms for collaboration that arise from the proposed NCRIS investments. Highlights from the report are as follows.

e-Research infrastructure investment

About \$10M pa of NCRIS funds in the investment plans may be being directed towards specific elements of e-Research infrastructure. The activities within that \$10M pa are mostly focussed on data access and platform integration across multiple sites, with little in modelling, HPC or networking.

Also, aggregating with 5.16 yields a funding rate about half that of the SII funding rate over the last three years. The result may be a reduced rate of general e-Research infrastructure development and significant difficulty in providing additional services

Data Management

Overall significant attention has been paid to the capture, curation, and access, for “community” data.

In general the investments leave responsibility for data with the facility (5.1, 5.2-phenomics, 5.10) or with the institutions related to collections (5.2-atlas, 5.12, 5.13) and the researcher otherwise (so that data is then managed according to the work practice of the researcher’s environment).

Grid

The requirements for different grid-like capabilities can be distinguished as follows:

Grid service	5.1	5.2	5.3	5.4	5.5	5.8	5.10	5.12	5.13
Integrated multi-site platform	X	X	X						
Distributed shared view of data	X	X	X			X	X	X	X
Remote access to compute power	X	X	X	X	X	X	X		X
Real-time multi-site data analysis							X		
Virtual presence and control			X				X		
Access Grid like collaboration				X	X	X		X	X

Expertise

A significant part of the this expenditure will relate to e-Research expertise, which will be in short supply. Mechanisms to grow and share key human resources need to be provided.

High Performance Computing

The overall needs of modelling and theoretical research, and the increased demand that may arise from research activity associated with the NCRIS facilities is likely to easily exceed supply.

The benefits of speed vs capacity and the best means of user engagement need to be investigated.

Networking

A desire has emerged for subscription based research traffic from all NCRIS funded facilities to all Australian researchers; from shared facilities through to every researcher’s desktop.

As investments are made, the network backbone is likely to become less homogeneous (some links with higher speeds), so that a revised pricing framework may need to be developed and agreed.

⁴ Investment Plans and Platforms for Collaboration, DEST, October 2006 (see pfc.org.au)

3 Progress with consultations

Consultations around e-Research Toolkits and the broader concept of the grid have yet to take place.

3.1 High Performance Computing

Any given researcher and research community will have access to a variety of compute and storage capabilities, including large systems operated on a merit basis; mid-range systems at institutions operated on a shared access basis; and smaller and often dedicated departmental clusters and desktop systems. Consequently, the HPC needs of a research community are difficult to estimate, and the degree to which central or co-ordinated provision is relevant, is even more difficult to estimate.

However, the comments on HPC made during the APAC review provides several observations, perhaps most importantly related to the scale of investment.

- Additional expenditure is required if the peak facility is to be retained at the historical level
- A more frequent purchase rate is needed to improve the return in Tflops delivered against dollars, under the assumption that a significant overlap in the operational periods of systems is manageable within the machine room infrastructure
- Given the already competitive nature of access, more resource overall should be provided to allow for the broader clientele envisaged under NCRIS

The different requirement for rate of computing (peak Tflops/s or capability) versus the amount of research computing that can be supported (Tflop-years or capacity) is left unclear.

While the annual average for both, and the corresponding rates per dollar spent, are all significantly improved if the frequency of purchases can be increased, which of the two measures one is aiming at dominates the strategy for provision.

Overall, some of the implied need associated with the NCRIS investments may be able to be met by significant capacity rather than requiring peak capability, however this is untested.

3.2 Advanced Networking

The NCRIS investments suggest a generalisation of the existing research network that needs to be better understood, both in terms of detailed requirements and in provisioning options.

We can note overall that NCRIS investments:

- include many examples of grid like systems providing shared access to infrastructure operating across multiple sites with a national distribution
- includes in 5.3 and 5.10, examples of very high real time data generation rates that may require special analysis and treatment
- represent a substantial move to develop federated views of widely dispersed data (in all but 5.4 and 5.5), which may lead to service level requirements designed to support rapid search and access

The manner in which the research network can be generalised to encompass a range of government agencies and state based research institutions needs to be determined as does the manner in which costs can be met by volume and destination independent subscription charges.

Also, the actual user experience in access to the research network, depends ultimately on institutional infrastructure and institutional policies, that may also need to be considered. Apart from 5.3 and 5.10, each platform proposed under the NCRIS investments might reasonably operate over a quality of network connectivity that could be met with 1Gbps tails. How that reaches sites on campus and research desktops in general needs to be further discussed. Also, to support multiple investments with such connectivity, some of the research intensive institutions will need significantly higher aggregate connectivity which may also create network issues within those institutions.

3.3 Expertise

One of the three commonly identified issues in moving towards a more coherent development of e-Research infrastructure concerns the development and provision of appropriate expertise.

An analysis of investments in e-Research infrastructure across the NCRIS capabilities, identifies a level of spending that will ultimately translate in to a demand for expertise. The investment plans themselves show that the various communities are at different stages of development towards an e-Research perspective; which means they will necessarily have access to very different levels of such expertise. This will be true more broadly.

Also, e-Research involves the use of multiple and entirely unrelated specialisations, such as curation of data, advanced networking, or parallel software for supercomputing. Added to this, grid capabilities and middleware are a rapidly evolving set of specialisations in their own right (such as searching, authentication and authorisation). Research groups cannot possibly cover this space.

Some important factors related to e-Research expertise are:

- Expertise management is enhanced by building groups of specialists rather than relying on unrelated individuals
- Different expertise and different levels of expertise are required during different stages of a communities migration toward e-Research
- Eventually some expertise needs to be embedded in communities (eg. data curation) and some needs to be embedded in service providers (eg. network management)
- Along the way, flexible collaborative teams are needed so that the infrastructure can evolve as the requirements are better understood

Also, as communities become more e-Research oriented, they tend to co-evolve services for data generating and gathering, with services for information analysis and re-use. This happens because each community needs to develop a consensus on the standards required for inter-operation.

Standards development is always a long iterative processes, which means that researchers will undertake bespoke software development in order to continue their research within the evolving context. Further difficulties then arise as reliance is placed on such software leading to a need for improved software engineering and particularly software productisation expertise.

Many of these issues were identified by the e-Research Co-ordinating Committee and the basic perspective developed by that committee remains valid and is reflected here.

A solution to expertise is beyond the budget of Platforms for Collaboration and a broadly based consultancy activity is most likely out of scope from an infrastructure investment point of view.

However the problems arising from the fragmentation of e-Research infrastructure through divergent directions in investments is exacerbated by this missing expertise and in particular the leadership that could be expected from high levels of expertise.

Consequently, some actions need to be taken and the investment plan for 5.16 cannot be silent on expertise and leadership.

- Some form of recognised peak body or association needs to be formed, the exact role needs to be resolved, but some initial suggestions are made in a later section.
- Investments in 5.16 need to be structured to enhance and provide access to pools of expertise where they can be found or where other parties will co-invest

It should also be noted, as a rough estimate, perhaps 50 new experts in various areas of e-Research infrastructure may need to be appointed into NCRIS platforms over the next five years, in addition to expertise developed within service providers.

3.4 Authentication and authorisation services

Another of the three commonly identified issues in moving towards a more coherent development of e-Research infrastructure concerns authentication and authorisation. A uniform approach to authentication is critical if ad hoc collaborative groups are to be easily supported.

- Any consideration of national e-Research infrastructure leads immediately to the need for a shared method of identifying researchers across institutions that eliminates creating new identities, new logins and new passwords every time a person is added to a project or a collaborative activity
- A second need then appears around systems that allow owners of resources to say who can do what; and to define such authorisations for access in terms of roles and categories of people
- The third component can then be identified that relates to the provision of access control systems that enforce those decisions reliably and with certainty within applications
- However, before any of this can be deployed systemically, a range of framework and policy decisions need to be made, for which community agreement is needed, so that trust is established and the system that provides authorisation is easy to use

The consultation process within Platforms for Collaboration brought together representatives from the Open Access to Knowledge Law Project, the Middleware Action Plan and Strategy project, the E-Security Framework for Research project, the Meta Access Management System project, the APAC National Grid, AARNet and CSIRO, to review what was known and what plan might be developed.

Agreement has been reached on a path forward noting that the technology and product support for each of the needs identified above is in different stages of development and yet all need to be in place for acceptance of the system to be wide spread. Overall:

- A trust federation will be established to provide the foundation on which identity and authorisation servers and transactions can rely. This trust federation will at least cover the entire Higher Education and Research sector, and may federate with other national trust systems where possible, such as may be implemented for various government agencies
- The trust federation will support PKI and shibboleth, and may support other standards as required
- To support the formation and ongoing operation of the trust federation, discussions will be convened within CAUDIT to consider the approach to compliance that should be taken, in terms of policy, framework, escalation, advice, remediation, and training
- Effort within the projects listed above will be directed towards developing a range of next steps, including: a Shibboleth federation policy; a suitable minimum set of attributes and mechanisms for extensions; bridging between the APAC National Grid and the trust federation; the development of some simple use cases; and the working up of the required directories and policy enforcement components to support those use cases

How the trust federation will be structured legally and organisationally needs to be determined.

Such a trust federation is a key system service essential for collaboration, and as it will simplify and support operational aspects of the various platforms and all of the access regimes envisaged by the NCIRS capabilities, it appears a natural fit for investment by 5.16.

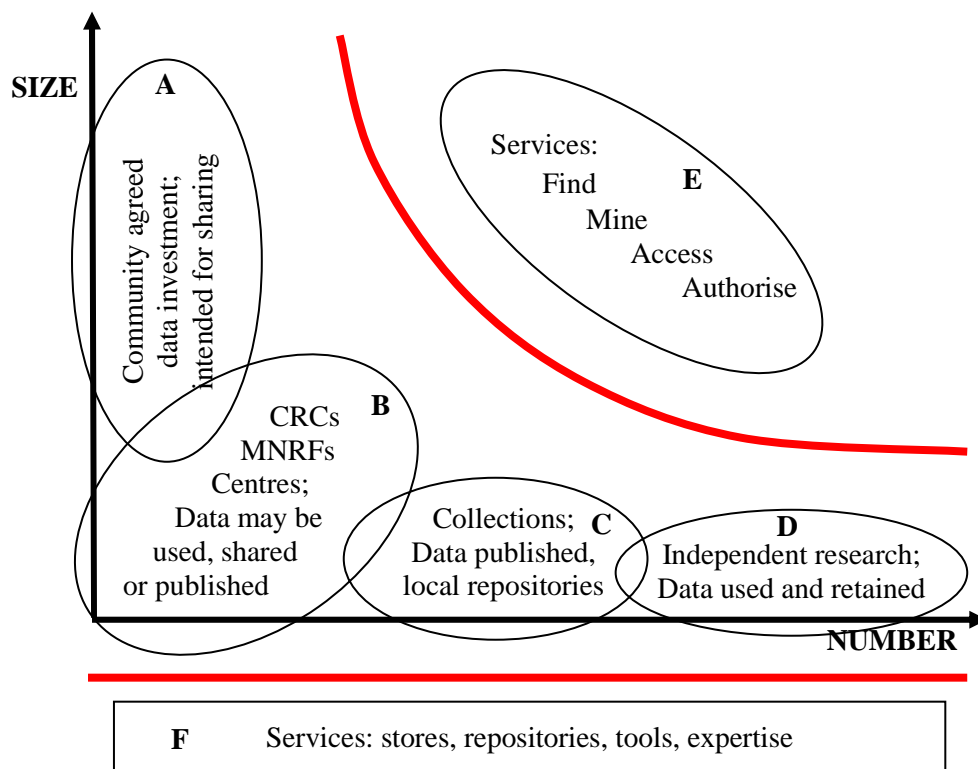
However the trust federation will also be able to support services related to a range of non research activities (such as student and staff resource access) and may also federate to non research organisations to assist resource sharing, suggesting either some form of subscription funding may be appropriate or that funded value adding services should recoup some of the base level operational cost.

The participants in the AAA consultation intend to provide a detailed plan with governance, budgets, funding model and service roll-out within the 5.16 Investment Plan.

3.5 Data management services

The last of the three commonly identified issues in moving towards a more coherent development of e-Research infrastructure concerns data management.

Consultations and surveys undertaken by Platforms for Collaboration suggest that data volumes are growing rapidly and that a significant fraction of research data is held in isolated forms and not easily accessible, and that a number of different missions exist around data, depicted below.



Within these cases, two kinds of data can be distinguished: private and public data. Private data is usually held by a researcher through self generation, or by acquisition from surveys, an instrument or device. Public data is data which has been placed in a public repository for general public access. Often, researchers also make some of their private data available on a controlled access basis, e.g. by copying to collaborators, but do not normally provide public access to all the data they hold.

In some disciplines where data is derived from significant public or shared investment, a common data holding protocol applies in which each researcher is granted exclusive or controlled access to the data collected on their behalf for a limited time. After that time the ownership of the data reverts to the institution in which the researcher is based, or is made public through a repository.

With reference to the diagram, for the different missions, different properties and responsibilities are present, and the role of NCRIS needs to be determined.

- A. Several communities have established practices that capture and share data of general value. These tend to form large datasets that are generated or gathered as the result of significant investments; such as in astronomy, high energy physics, earth observations and bioinformatics. The size of these data sets tend to be very large, typically in terms of tens to hundreds of terabytes and sometimes petabytes, with high growth rates.
- B. At a slightly smaller scale and more numerous, various organisations e.g. CRCs and MNRFs, have the gathering or generation, curation and publication of data as part of their mission; or use data for internal purposes, the results of which may be published in the form of information services; or they provide research support services that generate data. The data often is privately held, and even when it is published, it is in the form of processed data, the raw data may not be made

available to the public. The size of these data sets tends to be in the many gigabytes to terabyte range.

- C. Continuing to decline in scale and increase in number, many research organisations, departments, teams, and even individuals have established data collections the contents of which are intended for publication and access. These publicly available data sets (donated by researchers) are normally kept in institutional or personal archives or web sites. The size of these data sets tends to be in the multiple gigabyte range.
- D. At the end point of this curve, nearly all researchers generate data or store data on the desktop, much of which is only ever intended for individual use, and much of which is inaccessible to other researchers and only visible through the publication of derivative research results. Most of this raw data, would have to be uploaded to a web site or institutional repository to make it accessible. The size of such data sets can be quite variable depending on the devices or computational models that generate the output.
- E. A separate class of mission relates to organisations that seek to provide access to research data, either within disciplines or more broadly, and which may or may not hold the primary data themselves. These organisations provide catalogues, abstracts or thumbnails, and can search and possibly mine data over which they hold indexes. These sources can only provide access to public data or private data researchers have decided to make publicly accessible. The aggregate size of data accessible in this way would be in the range of terabytes to petabytes.
- F. A final class of mission relates to service providers seeking to support the retention, curation, access and analysis of data sets. Generic service providers necessarily operate only on publicly accessible data, although some specialised services could operate on controlled access data. Their added value lies in the co-location of data sets and the consequent ease of datamining and linking that becomes possible. Many institutions might operate a similar service on their own behalf for their retained private data. The size of repositories tends to be in the range of hundreds of terabytes to many petabytes, and exabytes would be conceivable in some settings.

This discussion highlights that data management services will be provided by a variety of sources and the development of investment decisions for data management by NCRIS in general and 5.16 in particular needs to be based on principles that fit within a broader framework.

Fortunately such frameworks are under development, and 5.16 planning can make the following assumptions in line with current thinking by the AVCC, NH&MRC and ARC.

Institutions should:

- Develop and implement a policy on data ownership
- Provide guidelines to researchers on ownership, what to keep and researcher responsibilities
- Maintain durable records on what research data has been held and ensure that research data is under the control of the institution where the work was performed
- Provide secure systems for holding data and for granting access to that data

Researchers should:

- Determine what data to keep, considering research community practice and any project or legal requirements
- Ensure research data is retained (for at least 5 years from publication of results) using institutionally provided mechanisms
- Ensure at the end of employment (for whatever reason) data retention passes to the institution
- Maintain confidentiality where it exists

Also, data actually moves through a life cycle, and gains value from the process, so that the use of the term data here is broadly based to include digital objects.

In consideration of the above, the following general principles will be applied in determining the services and support 5.16 might provide.

- Investment by 5.16 should be directed towards data which is of value and preferably lasting value and available for public access, either immediately or ultimately
- In general, investments (by other sources) intended to support data acquisition should be expected to also fund or arrange retention and access services for that data for the lifetime of the investment
- Retention of privately held data or data to which access is highly limited is the province of the institution(s) sponsoring the research

Taking these principle into account we arrive at some possible investment goals.

- 5.16 could support public data retention when investment periods expire, where that data is important or irreplaceable; across all categories in the diagram above. However the permanent preservation of data from category A is a difficult problem given the potential scale, but is also not expected to arise within the current investment horizon for 5.16
- 5.16 could provide the means by which collections of public data (generated in categories C and D) could be preserved and made accessible, on the basis that such a service adds considerable value to the data and will not otherwise be easily or quickly provided. This would relate to data specifically nominated for public access and deposited by researchers into public repositories
- Such a “collections service” could be extended to also include public data deposited into repositories from activities in category B
- Of the two sets of services depicted in the diagram, those related to re-use (E) and those related to retention (F), support on the re-use side (E) is clearly systemic, beyond the bounds of individual institutions, complementary and value adding to individual data holdings
- Therefore 5.16 could focus on data search, data mining and access control services in E and perhaps national data location and movement services in F

There is another issue which may need to be considered in developing the investment plan, that being the benefits of co-location of data and compute resources. At present network bandwidth is perceived to be sufficiently high to transport data to compute resources. However, the ever increasing amount of data may make it more efficient to install processing power close to the sources or repositories of data. This suggests a future intention to build towards a number of e-Research support facilities that provide for the co-location of compute and data storage resources, is a desirable goal.

Underlying this analysis is a governance question, namely which body will have the authority to decide on the longevity of data sets, particularly if they are considered to be “nationally significant” and worth long term archival. Clearly any funder can do so for the duration of their funding, the issue arises when investments expire and some transfer of responsibility is needed. In addition, sponsors of research will need to be clear which data generated from their sponsorship is private, limited access, immediately public or ultimately public.

These issues, along with the implications of the stages in the data life cycle as well as requirements for supporting infrastructure such as universal identifiers will be discussed in further workshops.

4 Outline of Investment Structure

The Australian national e-Research infrastructure can be expected to take the form of a combination of national and regional activities due to the decision making processes of governments and institutions.

A foundation principle for the investment plan is to strengthen and build on co-operative arrangements so that an increasingly coherent level of support can be provided to researchers, and their collaborations and communities.

The goals of the investment plan for 5.16 are to provide e-Research infrastructure services which are:

- of broad value across all research communities
- of value to those communities whose research needs led to the NCRIS capabilities
- of value to the services delivered by the platforms created or augmented by NCRIS investment

Hence 5.16 investments will focus towards shared services that no single community could afford or justify in its own right, or where piecemeal development would hinder rather than aid collaboration.

The intention is that such services should be delivered through agencies where the service is the primary mission, so that over time, other capabilities and research investments can source (potentially dedicated) services from this infrastructure and provide appropriate funding to it for that purpose.

At a functional level, this demands a highly co-operative approach across suppliers of infrastructure and services so that the user experience is more cohesive in terms of data management and access, the sharing of compute capabilities, the controlling and setting of authorisations, and an ease of remote interaction through readily available “always-on” collaborative environments.

4.1 Overall Mission

The term eRI, for e-Research Infrastructure, has been adopted herein as a place holder for this overall goal and the entity which might occupy that position. That entity could be a development of the existing APAC partnership or it might be a new entity, the governance arrangements are deliberately left as a subsequent discussion at this time.

eRI is proposed as the peak body tasked with developing and sustaining the forums in which co-operation can be achieved and which over time provides policy and standardisation frameworks that deliver a nationally coherent e-Research infrastructure.

Its mission will be to identify, develop and deliver nationwide and world class services and expertise, that support effective e-Research within and across all research disciplines.

This will include services and expertise related to:

- data capture, management and retention
- data publication, discovery and re-use
- data analysis
- computational modelling
- collaboration systems
- grid inter-connectivity
- networking

4.2 Context

The development of e-Research and related infrastructure requires a mixture of:

- Policy advice and implementation
- Consulting to bring together requirements and solutions that meet user needs within sustainable architectures and deployable services

- Provisioning of services and directing service development through service providers

The stakeholders relevant to eRI include researchers and communities of research, research facilities, research support services, institutions and collaborations of institutions, and their funders.

The manner in which these entities will be included within the governance of eRI and/or the service providers funded by or related to eRI will be considered at a later stage in the planning.

4.3 Potential Activities

The following activities represent a core set, further work may identify additional services or responsibilities within proposed service areas.

e-Research Infrastructure (eRI)

- Providing a framework for investment and allowing for further activities and funding
- Promoting e-Research and the deployment and use of e-Research Infrastructure
- Mapping user/application requirements, priorities and expectations, to assist services meet needs and transform the landscape
- Providing policy advice regarding e-Research deployment and adoption

e-Research Consulting (eRC)

- Developing and deploying staff within eRI and from other relevant entities (service providers and user communities) in a co-ordinated and expert e-Research consulting service
- Providing a team solution to researcher, research community and research facility support

The e-Research Trust Federation (eRTF)

- Providing the central support (including auditing) for PKI and shibboleth infrastructures

The Australian National Data (Collections) Service (ANDS)

- Providing retention services for important collections
- Hosting a range of access, location, and datamining services

The Australian National Grid (ANG)

- Supporting access and interoperation for compute and data services
- Hosting instrument/facility/data integration to the grid
- Providing advanced collaboration and visualisation environments

The Australian National (Computing) Facility (ANF)

- Providing a peak HPC/data capability service, accessed by shares and merit allocation

The Australian Earth Systems Science Facility (AESSF)

- Providing share based access to a peak capability service for Earth System Science research

The Australian Research and Education Network (AREN)

- Providing a broader, more capable, subscription based research and education network

4.4 Budget Issues

The total current activity in these areas (including activities funded through SII relevant to AAA) represents about \$35M pa from DEST and a further \$65M pa from participants. In addition, the first two items have a new breadth and are therefore not currently supported within that funding.

The NCRIS provision of \$15M pa, if the sole funding line, requires participants to take a greater share of the burden, or implies a reduction in activity. The intention in further planning is to find ways to support all of the above activities if possible.

5 Project Plan

Month	Activity	Details	Changes	Responsible
May, June	Start-up	Agree reference group, role and scope, contract, engage DEST/facilitators		Rhys, Ian
July	R/G Meeting #1	Establish role and scope, identify actions, review 1 st deliverable		Rhys, DEST
23 Jul	1st deliverable	PfC implications from Progress Reports		Rhys
August	5.1 IT Architecture Activities R/G Meeting #2 Networking BoM/ACCESS	Engage help for review of IT requirements in 'omics Scoping work around proposed workshops Implications of initial NCRIS feedback, scope tasks Establish activity to provide network input to the plan Establish activity to include BoM/ACCESS input to the plan	Delayed	Rhys, Ian, Paul Davis Reference Group Reference Group Rhys, Ian Rhys, Ian
September	APAC review Networking 5.1 IT Architecture 5.8 Collab scoping AAA workshop #1 Collab workshop#1 Archer scoping	Engage reviews and results Progress information from aarnet Finalise outline of 'omic IT requirements, trial with IT Directors Assist develop scoping for bio-security collaboration environment Bring parties together to gain views on 1, 2, 5 and 10 year visions Bring parties together to gain views on 1, 2, 5 and 10 year visions Outline information management needs across capabilities	Delayed Replaced	Rhys Rhys, Peter, aarnet Paul Davis Markus Buchhorn Daniel, Nick, et al Daniel, et al Ah Chung, et al
October	R/G Meeting #3	Review Progress Report and associated inputs (as above)		Reference Group
6 Oct	2nd deliverable Data workshop #1	Provide written comment on the first 9 Investment Plans Bring parties together to gain views on 1, 2, 5 and 10 year visions		Rhys Linda, et al

NCRIS		Platforms for Collaboration	Progress Report	
19-20 Oct	Network Review #1	Bring parties together to gain views on networking requirements and options	Delayed	Peter, et al
30 Oct	3rd deliverable	Attend NCRIS Committee meeting as required		Rhys
		Progress Report including overview of existing infrastructure	Delayed	Rhys
November	AAA workshop #2	Develop content for path forward	New	Daniel, Nick, et al
	Collab workshop #2	Develop content for path forward	Replaced	Daniel, et al
	Data workshop #2	Develop options for investment related to data management	New	Linda, et al
	Archer workshop #2	Consolidate information management options across capabilities	New	Ah Chung, et al
	NCRIS Investments	Consolidate requirements from NCRIS investments		Rhys
	Service providers	Discuss options for service supply, aarnet, apac, others		Rhys
December	R/G Meeting #4	Review outputs from work to date, determine any additional activities		Reference Group
	eRI Members	Scoping governance arrangements with eRI members	New	Rhys
	e-Research Toolkit #1	Consider what tools and technologies could be promoted to assist e-Research uptake	New	Rhys, Daniel, Paul et al
January				
February	AAA workshop #3	Finalise proposal for trust federation		Daniel, Nick, et al
	Data workshop #3	Finalise proposal for any data management investment		Linda, et al
	e-Research Toolkit #2	Finalise any initial tools and technologies to be supported	New	Rhys, Daniel, Paul et al
	R/G Meeting #5	Review draft investment plan		Reference Group
23 Feb	4th deliverable	Draft Investment Plan		Rhys
March	R/G Meeting #6	Provide DEST with input on Investment Plan		Reference Group
		Make presentation to NCRIS Committee		Rhys
30 Mar	5th deliverable	Revised Investment Plan and Final Report		Rhys

6 Expenditure to date

A summary of expenditure to date against the original budget headings along with an estimate to the end of the consultation process is provided below.

9-Nov-06

NCRIS - Platforms for Collaboration

LIFE TO DATE FINANCIAL SUMMARY

01 May 06 - 31 October 2006

	Budget	Total to date	Estimate
Income			
DEST funding	250,000.00	100,000.00	250,000.00
Total Income	250,000.00	100,000.00	250,000.00
Operating Expenditure			
<u>Facilitator Expenses</u>			
Remuneration/on-costs	\$157,531	\$50,340	\$157,531
Administration support	\$6,922	\$2,141	\$6,922
Travel	\$23,000	\$8,543	\$22,543
Incidentals	\$2,000	\$245	\$1,145
<u>Reference Group meetings</u>			
Travel	\$24,000	\$3,724	\$11,245
Accommodation	\$9,000		
Incidentals	\$3,947	\$458	\$1,458
<u>Activities & Consultations</u>			
AAA Workshops	\$7,200	\$1,488	\$4,488
Data Workshops	\$3,600	\$3,447	\$6,447
Collaboration Workshop	\$7,200	\$0	\$4,000
Network Review	\$3,600		\$0
Archer Workshop	\$0	\$831	\$831
IT Architecture (5.1-5.13)	\$0	\$0	\$0
<u>Miscellaneous</u>			
Binding/Printing	\$2,000	\$0	\$2,000
Total Expenditure	\$250,000	\$71,217	\$218,610