

PfC issues and other Capabilities

Introduction

The document reflects a first cut analysis from the Progress Reports submitted for NCRIS capabilities 5.1, 5.2, 5.3, 5.4, 5.5, 5.8, 5.10, 5.12 and 5.13, of the demand for services and capabilities that might be supported within 5.16.

The document is a preliminary analysis intended to identify issues.

Assumptions

The analysis makes some assumptions

- At least 1Gbit networking should be provided to all significant research sites, where a site is significant if it is a site included in the facilities nominated in Capability plans or is a recognized research institute.

A policy on the nature of sites that should be connected and on what cost basis may be required

A policy regarding the equitable treatment of links which need higher bandwidths or which have unusually high costs may also be required.

- A national approach to research data retention will exist and it is likely that institutions sponsoring research activities will need to provide data retention capabilities to support that research.

Some agreed approach to sustaining research data retention is required, the comments in this document assume that data retention need not be included with every data source because the accountability for data retention does not generally lie with the data source.

In this model, while NCRIS funds facilities, it is not itself a sponsor of research.

- Any given facility and any given researcher could have access to a wide variety of compute and storage capabilities ranging from large systems, operated on a merit basis, through mid-range systems at most if not all research institutions, operated on a shared access basis, to a large number of usually smaller and dedicated departmental clusters and desktop systems.

A key issue for PfC will be the degree to which the location of such capabilities should be aligned with the sources of data, the users of data, or the owners of data.

*** Note it appears the concept of an "owner" for most research data is not well defined, certainly those professing to ownership often expect others to bear the costs of ownership.*

A second key issue will be the formation of a policy on the preferred physical and logical structure of the "national research data repository". Noting that it will exist in some form, even if that form is "completely disaggregated and un-coordinated".

***Note the reference groups initial view is that it will need to be "distributed and co-ordinated".*

This first cut analysis simply assumes that sufficient 'grid' connectivity could be achieved so that storage and compute capacity could be aggregated at institutions or facilities and appropriate pools of expertise established. This assumption will need to be tested against application needs.

Infrastructure needs

Data management, access and discovery

5.1 Evolving Biomolecular Platforms and informatics		
	Data size and sources	Moderate datasets, < 5TB, but updated daily, heavy generation on-site
	Retention/curation	By the proposed 'omic facilities
	Accessibility	Base data accessible from facilities, derived data mostly confidential
5.2a Integrated Biological Systems – Phenomics		
	Data size and sources	Small data sets, generated on-site
	Retention/curation	By the proposed facilities
	Accessibility	Through facility provided search and discovery services
5.2b Integrated Biological Systems – Collections		
	Data size and sources	Small data sets (moderate if imaging is included), generated on-site
	Retention/curation	By the Institutional 'owner' of each collection
	Accessibility	Through federated search and discovery services, operator unknown
5.3 Characterisation		
	Data size and sources	Moderate to very large, aggregate 1PB pa, generated at a few sites
	Retention/curation	By the researcher
	Accessibility	At the researcher's discretion, public access policy uncertain
5.4 Fabrication and 5.5 Biotechnology products have no stated requirements for 5.16 services		
5.8 Networked Biosecurity Framework		
	Data size and sources	Little platform specific data sets, synthesis required across a large variety of 3 rd party data sets, at up to 20 main sites, of unknown scale
	Retention/curation	Source data by 3 rd party owners, capability data by the facility
	Accessibility	Controlled access required to data delivered in standard formats
5.10 Optical and Radio Astronomy		
	Data size and sources	Large data sets, generated at observatories, 400TB pa + optical
	Retention/curation	By the observatories, and replicated internationally
	Accessibility	18 month restriction, then available through search services
5.12 Integrated Marine Observing System		
	Data size and sources	Small sets from various sensors, large sets from modelling, 120TB pa
	Retention/curation	By the institutions maintaining the sources, by researchers for models
	Accessibility	Through existing federated domain specific search services
5.13 Structure and Evolution of the Australian Continent		
	Data size and sources	Very large collated data sets per jurisdiction, large datasets from modelling, large data sets from imaging collections, 720TB pa, 20PB off-line archive
	Retention/curation	By institution for collections, by researchers otherwise
	Accessibility	Federated specific search services; currently under development

A major question concerns where data should be kept, especially given the view that the general research system lacks policies and procedures to reliably retain research data and make it accessible.

In general the table above leaves responsibility with the facility (5.1, 5.2a, 5.8, 5.10) or with the institutions related to collections (5.2b, 5.12, 5.13) and the researcher otherwise.

The 'researcher' case means that issues to do with permanent retention and publication need to be addressed in the research sponsor's context. Even though this is problematic, facility based curation for say characterisation (5.3) would create its own issues, so that it is a reasonable case to use to 'pressure' the system. If some data is lost, it is not by and large irreplaceable natural observations.

PfC will need to form a separate view on the needs and support required for disciplines outside of the NCRIS focus.

Grid-enabled technologies

Few of the capabilities reference the need for system wide authentication, authorisation and accounting support (AAA) or mention other specific compute or data grid capabilities, 5.13 being a very clear exception. Applications for grid capabilities could be expected to arise as follows.

- Plans for cohesive facilities implemented across multiple sites, as in the 'omic platforms of 5.1, the phenomics platforms of 5.2, the microscopy and image facilities of 5.3
- A less tightly integrated fusion of resources across multiple organisations in 5.8, 5.12, 5.13
- Plans for virtual presence and control in 5.3, 5.5 and potentially 5.10.

While comments within 5.8 suggest a deep integration, it may be that the provision of a shared collaboration environment in which selected experts can easily interact is the main requirement.

PfC will need to undertake a review of AAA developments, as this is an implicit requirement, and will need to develop a clearer view of options across a range of integration and collaboration issues.

High performance computing

The main investments directly in HPC arises in 5.1 and 5.10 within individual facility budgets.

Hence HPC is either not mission critical or most capabilities are assuming that it can be provided in-kind or through capability 5.16 and institutional resources. Three potential scenarios might exist.

- The first is that in the planning for one-off expensive facilities, compute power is seen to be incremental and affordable and related to researcher and institutional environments and is assumed to lie within in-kind or contextual funding.
- The second is that where the facilities operate as project oriented service providers, such as in 5.3, 5.4 and 5.5; the end user's requirements for computing capability to support theoretical or modelling analysis is not intended to be met by the planned facilities.
- The third is that an immediate focus on data and data sharing is taking priority and the planning has not taken into account the flow through implications for data processing.

An understanding as to which of these cases is present in each capability would be helpful, along with any information on the scale of the assumptions.

PfC will need to form a separate view on the need for HPC investment for peak and mid range systems and their shared and dedicated status; covering the needs of modelling, analysis and theoretical work unrepresented within the capability plans and in disciplines outside of the NCRIS focus.

High capacity networks

Within the assumption of 1Gb/s to all relevant sites, four issues emerge: coverage, dimensioning, security and the flow through effects of volume or destination pricing.

- **COVERAGE.** Network reach beyond the current AREN participants may be needed for 5.3, 5.8, 5.11 and 5.12, in many cases to government agencies.
- **DIMENSIONING.** Access to a general purpose *shared* 1 Gb/s network may be insufficient for seamless operation of the 'omics platforms proposed in 5.1 and the NANO and potentially the imaging platform of 5.3. It is also likely to be insufficient to support real-time repatriation of data or real-time interaction from the synchrotron or the astronomy observatories, especially as image capture resolutions increase.

An evaluation of appropriate network dimensioning and QoS support should be provided for these cases, and the cost/benefits of dedicated (layer 2 or 3) networks investigated.

- **SECURITY.** Integration across systems connected over the network to end points within separately managed enterprise networks may be required for 5.1, 5.3, 5.8, 5.10, 5.12 and 5.13. This requirement raises well known issues in security and the ability and willingness of participating organisations to support the arrangements needs to be investigated.

- **PRICING.** Any inequity in pricing leads to difficulties. For instance, large apparently on-net transfers have the potential to include off-net components leading to unexpected charging. Also, various universities have volume based charging regimes that inhibit collaboration. The desire to include participants not within AREN and the consequent prospect of volume charging or cost differentials is also likely to inhibit collaboration.

Expertise

The area of expertise is looming as a major issue for PfC. Contributing factors are:

- The advanced modelling, information integration and analysis tools needed to take advantage of enhanced data gathering and measurement capabilities must be continuously developed/upgraded
- The domain specific nature of these tools leads to significant software development requirements which in turn lead to expertise and staff requirements
- The need to deploy sophisticated infrastructure (such as AAA) across multiple organisations which separately neither have nor can sustain the required quantity and quality of expertise
- The apparent roll-over of Frodo and Merri projects within the scope of 5.16 in competition with harder infrastructure investments

On the first two points, expertise is needed to support the co-evolution of data generating/gathering with information services; providing for co-development of data generation and data use. This implies software development and facility investment are both required to achieve 'capabilities'.

For instance, a view of 5.13 is that it is composed of a diverse set of data generating activities, which while categorised into classes, are relatively independent of each other in terms of generating their data. The capability has about 60% of the investment into strengthening these data publishing sources and about 40% in work directed towards synthesising information products on those data streams.

In capability 5.12, which may be further developed in this area, the integrated information product already represents a key value of the activity and in fact dominates the data generation.

On data re-purposing, astronomy (5.10) would argue that enshrining data as a sharable collection at observatories is the whole point and often refers to the number of uses of the same data for different research projects. This data inter-dependency is also present in the various 'omics.

The difficulty for PfC is policy, namely if the capabilities focus investments into assets; either the in-kind contributions or some other investment process will need to make good with the people, expertise and software development needed to release the value of those investments and their associated data.

As it is 'all software after all', there is a tendency to look to PfC for a generic solution. Leaving the budget question aside, the experience to date is that domain specific tools lead the way and that generic tools (should they be possible) are much more difficult to develop. PfC is therefore likely to be unable to address software development generically or within budget.

Proposed actions

- An 'IT architecture' for the 'omics platforms be developed and the respective institutions that might need to support them be invited to comment on their ability and willingness to do so
- A study into provisioning of the extended research network system be requested from aarnet
- The APAC review be used to gain a clearer picture of overall demand for HPC
- PfC undertake a study into AAA developments and roll-out strategies
- PfC bring together policy and implementation guidelines to support data retention related to the NCRIS investments, on a case by case basis if needed
- PfC undertake a study on collaboration and grid technologies to determine how accessibility and inter-operation could be better supported across the investments (where appropriate)

5.1 Evolving Biomolecular Platforms and Informatics

The 5.1 Capability identifies Bioinformatics with a separate investment; which emphasises the predicted importance of information for this field.

The overall investment pattern is however highly complex from an ICT integration point of view.

- Each of the Proteomics, Genomics and Metabolomics Facilities is a network of nodes across 5, 7 and 5 regions respectively and involving ICT components managed by 7, 9 and 8 institutions respectively.
- Each Facility includes a separate investment in “platform-specific informatics infrastructure including personnel, hardware, software, platform-specific standards for data capture and storage, primary analysis, annotation, workflow, and user support” suggesting some intent to provide unified view of resources and systems across the sites.
- The Bioinformatics Platform is also a network of 5 regions aggregated into 3 components of which 2 are multi-institutional, with a mission to unify data access across the other three Facilities.
- The Bioinformatics Platform intends to focus on “specialized computing and data solutions, mirrors of large international databases, complex and experimental analysis software, visualisation environments, portals middleware and services of a nature or scale that the individual 'omics platforms, singly or together, could not provide”.
- In the three budget scenarios, the Bioinformatics Platform has a share of total expenditure at 13%, 10% and 0%, being a total spend of \$6.667M and \$2.667M pa in the two funded cases.

The degree to which each Facility and the Bioinformatics Platform needs to, and could, provide a seamless view of associated resources to researchers is a critical unknown.

For instance, a seamless view by research groups of all their data holdings within the participating institutions is feasible. The integration of the data with analysis systems both local and remote is a more challenging grid problem. A seamless view of resources that could generate data is unlikely.

The stated PfC dependencies: access to large scale compute and storage systems, relationships with high bandwidth providers, and development of informatics expertise; are ok. Unknowns include:

- Expectations for dedicated networking between Facilities for data replication or real-time analysis
- The short term and long term model for data retention, and the volumes to be imported/exported
- The state of connectivity to all the nominated sites

A more problematic statement is that NCRIS 5.1 can depend on NCRIS 5.16 to provide:

- “Implementation of generic middleware and systems for nationwide user authentication, authorisation and other basic Grid services”
- “Implementation of generic middleware, systems and standards for knowledge management, curation and lifecycle management and for collaboration”

A reasonable first response (for discussion) might be:

- Nationwide authentication is an institutional issue, as institutions can best manage and attest to identities of staff and students. The Facilities should however expect to provide in-house identity systems for users not at authenticating institutions in the short term
- Authorisation requires effort by resource and data owners, even if 3rd party technology is used
- PfC activities can be expected to provide expertise on AAA and other general grid technologies, however 5.1 appears to be 3 separate grids in its own right, expertise levels are a major risk factor
- While storage may be provided by PfC and lifecycle expertise may be provided, the effort of “curation” needs to be covered within 5.1 resources; as curation is a community activity

A more detailed design for a possible ICT and AA architecture for 5.1 is needed.

5.2 Integrated Biological Systems

The 5.2 Capability identifies three investments representing two different relationships to PfC.

Phenomics

The proposals for enhanced mouse and plant phenomics includes instrument based data sources (many of which are imaging devices) that therefore should have a similar relationship to PfC as discussed for Capability 5.3 (Characterisation).

Unlike the typical data use in 5.3, however, a key objective in phenotyping is to share the phenotype information as it is created for the benefit of all researchers leading to the option that the phenomics facility should maintain and curate data in perpetuity.

Data held in this way could then be made available from the phenomics facility as part of the providence trail for research performed using the mice and plant variants.

Also, unlike 5.3, the phenomics data sets are diverse in their content, having a variety similar to those in 5.1, and hence the management and curation of that data will require specific investment within the phenomics facility (as per the proposals in 5.1).

It is probable that the relationship between the Phenomics Facility and PfC will be similar to that of the facilities planned in 5.1, and the microscopy component of 5.3.

The Phenomics Facilities could be similar in ICT architecture to that needed in 5.1.

Biological Collections

The proposal for Biological Collections leading to the Atlas of Australia represents a major investment in metadata capture, data linkage and distributed information search and retrieval.

The proposal allows for the generation of separate database representations of up to 64 major collections; including metadata, imagery, geo-referencing, and potentially DNA data.

Given the development of suitable schema and the annotation of the collections, ICT technologies certainly exist that could be adapted to provide a common access portal to the data.

In terms of the relationship to PfC:

- All the cost of database entry and the servers providing web based access to that data belongs within the institution housing each collection and hence the investment profile of 5.2
- The Atlas appears to offer no particular AAA issues, however, this is based on the presumption that it is a public good resource which should be freely available
- The Atlas is however likely to support important opportunities for datamining, correlation and image processing which could use grid like capabilities to harness compute power to the data
- As the use of any fraction of the data is likely to be sporadic, and as the compute demands are likely to be determined by individual research projects, the compute requirements of research use of the Atlas should be provided by the general capabilities supported through PfC (available by merit allocation or shared access)
- However, while the collections may need to be maintained at specific sites, and while curation may need to be performed by experts at those sites, the data generated for each collection within the Atlas could be located at any site
- Hence an analysis of the scale and update rates of the likely data holdings, particularly if including images, should be made to determine if data sets should be mirrored or housed at locations of likely PfC resources able to performing datamining and image processing at appropriate speeds

5.3 Characterisation

Overall 5.3 investment represent 3 categories of need within PfC

- a high volume potentially highly interactive data source at the synchrotron
- a low volume low interactivity data source at ANSTO
- a distributed network of data sources of moderate volume and interactivity in 4 locations comprising the microscopy facilities

Imaging appears to fall into a similar category to the microscopy facilities, except that it may entail more complex access requirements and has a less well developed collaboration framework.

In many respects, the synchrotron could be viewed as a co-located version of the microscopy network, with its higher aggregate requirements arising from a larger set of image sources, and significantly faster image acquisition at each source.

Thus the question arises at each image source: should local data retention along with analysis and image processing compute power be supplied at that source.

- In the cases of relatively low speed image acquisition, the image capture process could include its transmission off-site. Reliability leads to the holding of data for some short time and, as the scale of data is modest, the cost of holding a few months of data is a minor element of overall cost.
- For very high speed image acquisition, with consequently much higher data rates, local storage combined with off-line data transfer is a far more manageable solution.
- Where image sources are co-located with user communities, local data and compute capacity has high appeal to that community.

X-Ray Techniques (Synchrotron)

- Users are not co-located with the image sources and the data is private to each researcher; hence permanent retention on site is not required for operational or purpose based reasons
- If remote presence is required, proposed data rates may be 10 MB per second data, plus 2-3 video channels and two way audio, per beam line, to arbitrary sites around Australia, which is likely to challenge the shared research network

Neutron Scattering (ANSTO)

- Users are not co-located with the instruments, so that long term data retention is not required, although the plan is to retain data for 5+ years, perhaps in perpetuity
- Data volume is 2-3 TB per year, related to 400 experiments and perhaps 300 Principle Investigators per year, and each image generation takes many hours

Microscopy

- Users and instruments are co-located at multiple sites, albeit with a wider non-site user community, leading to plans for site based compute and data systems
- Data volume is 10-50 TB per year, image generation can take hours, and the processing of each image on small compute resources can also take hours

Overall

- The design model for all the facilities should provide for local retention of data for a limited period, provision for repeated export and data deletion at the end of the nominated period
- All facilities should include a plan for local and grid based authentication/authorisation systems and show how remote resources could be accessed from the facility to assist researchers on site
- A network dimensioning analysis that supports the Synchrotron data and remote presence requirements as image resolutions scale up over the next 10 years should be provided

5.4 Fabrication

The 5.4 Capability identifies four investment proposals all of which focus on the generation of materials of one kind or another and none of which contain specific reference to services related to PfC.

The proposals relate to staffing and equipment to support four facilities:

Provisional Title	Location	Scope
ChemBioFab	Victoria	Materials – chemo – sensors; System design and integration
SoftFab:	Queensland	Bio-nano-soft matter
PicoFab	ACT	Micro/nano photonics
ElectFab	NSW	Micro/nano electronics

In each case the investment proposal is either for lab infrastructure or expertise development.

In terms of PfC issues, the facilities would appear to involve minor volumes of data generation, to do with procedures and material properties (provenance), and the computational modelling of materials appears to be treated as ‘out of scope’ as far as the facilities are concerned, perhaps to be met by resources available to individual researchers.

It would be useful if any in-kind resources intended to support computational modelling associated with the use of the facility could be nominated.

5.5 Biotechnology Products

The 5.5 Capability identifies three investment proposals none of which appear to generate any demand for services related to PfC.

The proposals relate to staffing and equipment to support

- Manufacture of recombinant proteins as potential therapeutics
- Expansion of human cells for transplant
- Biofuels

In each case the investment proposal is either for wet lab infrastructure or expertise development around wet lab infrastructure to be funded from other sources.

There appears to be little modelling or simulation proposed, only low levels of data generation and no requirement for real-time remote presence.

Again it would assist PfC planning if any in-kind resources intended to support computational modelling associated with the use of the facility could be nominated.

5.8 Networked Biosecurity Framework

The investment proposals for 5.8 include a mixture of laboratory enhancement, expertise building, data and skill registries and travel support designed to enhance teamwork within 14 different existing institutions or networks of institutions.

An overarching goal is the ability to bring all these (and other) resources to bear in the case of a biosecurity incident. Rather than comment on the specifics of the investments, the following is a summary of a recent discussion between the facilitators for 5.8 and 5.16.

With regard to the attachment 8 in the 5.8 progress report, it was noted that:

- “The highest single national priority is for a national network allowing for collaboration across jurisdictions and across sectors at a national level, and across sectors within jurisdictions; linking biosecurity nodes and centres between states; interfaced with regional delivery networks within the jurisdictions, and linked to participating universities and research institutions.”
- This network would need to include the data, analysis systems and staff of up to 20 major centres in the biosecurity area, some of which have up to 400 staff
- The data in all these sites is managed according to the requirements of each discipline and no overarching metadata schema exists
- Each discipline has its own preferred data analysis and classification schemes and these were unlikely to change in the short term
- Not all sites were aarnet3 connected and in the biosecurity area, relevant expertise often existed in organisations outside of academic communities

PfC related issues

- A minimal requirement for collaboration would be the development of a means to provide access to each data source and its associated analysis tools by any other member of the network
- A preferred level of support would be that data was supplied in agreed standard formats for both content and metadata, regardless of the site managed curation and storage solutions
- In the case of an incident, an essential capability would be the creation of a shared work space for that incident, able to hold and distribute any data from any such source, and which controlled access to nominated members of the network as well as nominated additional persons
- Given the potential to create shared work spaces, a highly desirable capability would include the routine provision of shared work spaces across the network
- The resulting collaboration environment could have an extended reach (beyond the 20 major centres) by hosting data for other components of the biosecurity network through standard web services and portal technologies

Hence a realistic overarching goal would be to support enhanced teamwork through the sharing of data and analysis between diverse and geographically distributed human experts. A goal to automate cross discipline data analysis is currently infeasible.

Given the complexity presented by the full diversity of disciplines in the biosecurity framework, a pilot activity around five sites related in the domain of plants has been suggested, for which a scoping workshop between relevant grid and domain specialists has been planned.

The incident response objective supposes a high speed network, and authentication and authorisation systems (for this purpose) reaching outside of the research community along with unknown requirements on data security and retention policy.

The claim that the development of this “network for collaboration” fell within PfC funding and was outside of 5.8 funding needs to be reviewed. A path for successful implementation is more likely to involve tendering to commercial developers.

5.10 Optical and Radio Astronomy

There are few implications for PfC from the 5.10 capability.

Its statement with respect to PfC is:

“In all the scenarios described above, there will be a requirement to access the research fibre-optic network at 1Gbps, rising to 10Gbps rates by the 2008/09. The latter is required to support Very Long Baseline Interferometry with the ATNF telescopes. International connectivity at 1Gbps is required to enable international access to data products obtained with the MIRT and Australian access to data products from Gemini and other international telescopes.

Over the coming 5 years, it is likely that the capability will have an ongoing, although modest, requirement for open access to High Performance Computing. Astronomical usage of APAC for theoretical calculations in support of observational astronomy is likely to remain at the 2-3% level. Much of the HPC, including data storage, analysis and visualisation requirements for astronomy are in the form of dedicated infrastructure, and as such are costed in the capability's infrastructure requirements above.”

On the policy side, it is worth noting that in the case of the Anglo-Australia Telescope, Gemini, MIRA, the three instruments to be enhanced by the proposed investments:

- No charge is made for time allocated through the merit allocation
- Researchers carry their own costs for travel and accommodation
- Data archive and management costs are borne by the observatory
- All data is freely available to the international scientific community via a WWW-based archive, following a proprietary period of 18 months

These data policies appear to be derived from the notion that astronomical observation is made in the public interest and are of naturally occurring publicly observable events. Hence the data is not private and the observatory can be considered to be publicly funded to keep and publish the data.

PfC related issues

Despite the accepted public good rationale, and as per Biological Collections, an investigation needs to be made of the appropriate location for the data holdings with respect to the scale of computing that could be needed to perform significant datamining or image processing on astronomical data sets.

Noting that co-location with PfC computing investments could be achieved by replication or by agreeing to a non-observatory based storage solution.

A network dimensioning analysis that supports the expected data rates as image resolutions scale up over the next 10 years should be provided.

The observatories should include a plan for local and grid based authentication/authorisation systems and show how remote resources could be accessed from the observatory to assist researchers on site.

5.12 Integrated Marine Observing System

Capability 5.12 is proposing four classes of investment, two canvassing extension to a number wide array of instrumentation within the categories of Ocean Observing (as a single node) and Coastal Ocean Observing (as a network of nodes), a third proposing access to data from the Integrated Ocean Drilling Program (IODP), and a fourth on Support Infrastructure.

The key to the relationship with PfC is nominated as the Support Infrastructure, which includes in its goals the creation of an eMarine Information Infrastructure and a backbone for remote sensing data.

Support Infrastructure

A key output of the Marine Observing system is a coherent view of the gathered data and derived information products. The effort proposed within the support infrastructure focussed on this objective is certain be required, given the diverse organisations contributing the raw and derived data products.

The analysis of the data requirements suggest an annual acquisition rate of over 120 TB per year in primary data, subject to scale up should additional variables, additional sensors or resolution enhancements be deployed.

About 20-25% (*verify this*) of this data represents raw observations that are irreplaceable, so that a risk/cost analysis needs to be performed to estimate actual data holding requirements (assessing the need for multiple copies and off-site replication on a component by component basis). For instance, the ARGO observational data is already archived internationally.

The acquisition of remote sensing data needs some analysis for network bandwidth.

The statement that no technological impediments exist is accurate, ie sufficient grid middleware and domain specialist browsing and data trawling tools already exist to support the goals.

AAA requirements are not spelled out, the proposals appear to assume public access to the data.

Ocean Observing

The Ocean Observing Node involves 6 different data gathering systems supported by a dozen organisations in various combinations, some of which are international.

The observation data rates are low (*verify this*), the large data volume is in derived products generated at sites with access to significant compute resources, hence the data gathering component presents no network issues despite the number of participating organisations.

There appears to be no need to consider an integrated 'grid' platform across the participants in order to implement the Ocean Observing Node. However, sites providing derived products might well participate in any grid infrastructure supported through PfC investments.

Coastal Ocean Observing

The Coastal Ocean Observing Nodes involve 7 data gathering systems, each supported by a large number of organisation, many of which are government agencies and some of which are international.

Again the observational data rates appear to be low volume. The means by which permanent retention will be achieved for irreplaceable data should be spelled out.

The degree to which the Coastal Ocean Observing Nodes might come to rely on compute or data storage capability funded through PfC should be checked. Many of the participants may not have access to significant in-house computing capabilities, but are also not academic institutions.

An overall ICT architecture could be developed to explain how services will be implemented, however if a loose federation of parties serving data (which is maintained in their own interest) against public queries in agreed standards is all that is required, few significant technical issues are likely to arise.

5.13 Structure and Evolution of the Australian Continent

The 5.13 Capability proposes six areas for investment: earth imaging and structure, geochemistry, simulation and modelling, virtual core library, geospatial, and access and interoperability.

While relationships with PfC are obvious in simulation and modelling, and access and interoperability, the other areas all either generate data from instruments or sensors, or accept, hold and publish data captured by third parties.

For the medium funding model, simulation and modelling is allocated 20% of the NCRIS funding for geosciences, which will be expended largely on software development of simulation codes.

About 22% of NCRIS funding will be allocated to access and interoperability which will be expended largely on middleware deployment and portal development.

The proposal contains a detailed statement of needs with respect to possible PfC activities, which includes the following:

- That compute facilities will be a combination of dedicated facilities funded within 5.13 as well as access to PfC funded shared facilities
- A strong preference for the adoption of common grid middleware perhaps through gateway servers that allows seamless access to 5.3 and 5.16 funded resources
- A requirement for distributed and centralized storage, with an identified demand for 750 TB of HSM with an additional 20 PB of off-line storage required to archive all raw geospatial data
- General satisfaction with 1Gb network links
- The need to extend network links to the Geological Surveys
- A requirement for a parameter sweep service operating across any compute capabilities provided through PfC
- That PfC should provision services that support the deployment of user-community developed software across PfC shared infrastructure

PfC related issues

Overall 5.13 appears more data and compute intensive than other capabilities, and has high growth potential for its demand in both areas. (Noting however that 5.1 may eventually be the leader in both.)

High speed network links to a range of research and non-research organisations are required, and those organisations may need to enhance their internal network and data server speeds.

The 5.13 investments appear to be developing sources of relatively independent data that do not need to integrate to provide their individual services but need to support common standards and interfaces to achieve composite services.

This suggests that 5.13 could achieve seamless access and inter-operation from a user's perspective using a hierarchy of services rather than combining the participants' IT infrastructure.

An alternative view would be that the geochemistry, virtual core library, and simulation and modelling investments all represent opportunities to integrated systems from multiple geographically distributed resources using grid technologies.

Some analysis of the cost/benefits between these views would be beneficial.

A more detailed design for a possible ICT architecture for 5.13 based around would also provide an important contrast to the design likely to arise in 5.1.