Submission
2016 National Research Infrastructure Roadmap Capability Issues Paper

Name	**Glenn Moloney**
Title/role	**Director**
Organisation	**NeCTAR**

Introductory Comments:

NeCTAR would like to commend the expert working group and expert teams on the breadth of considerations within the *National Research Infrastructure Roadmap Capability Issues Paper*. Reflecting on the issues paper, Nectar, ANDS and RDS have identified that the aims of a joint investment in *Data for Research and Discoverability* should be **improvement of the value and use of data** through a range of targeted activities. This will be achieved by facilitating partnerships that enable the development of sustainable, reliable, and interoperable national infrastructures.

In particular, NeCTAR encourages:

- Development of a nationally interoperating **Active Data** infrastructure which accelerates the generation of new knowledge from research data by Australian research and industry;
    - Supporting enhanced **knowledge sharing** and collaboration across research disciplines and industry;
- **A responsive investment in a responsive infrastructure** which can support new directions, facilitating the innovation that is important to every sector of the economy; and
- A focus on **infrastructure partnerships** with NCRIS research domain and mission focussed investments leading to improved sustainability, co-investment and strengthened partnerships in knowledge translation.

NeCTAR's current investments have been focused on addressing the needs of national research communities, and the nature of relationships between data infrastructure and these communities, research institutions and other NCRIS investments must be a key consideration in delivery models.

In this response, we have addressed the key questions related to the *Data for Research and Discoverability* and the *Underpinning Capabilities* focus areas in detail. We have also provided some responses to several of the general questions at the end of this document.

Data for Research and Discoverability

Question 33 Are the identified emerging directions and research infrastructure capabilities for Data for Research and Discoverability right? Are there any missing or additional needed?

The purpose of a joint investment in *Data for Research and Discoverability* is to place Australia as a world leading locus for data intensive research. NeCTAR, ANDS and RDS have been meeting to consider options for how increased and more closely aligned collaboration across the three projects could develop to support a NCRIS *Data for Research and Discoverability* investment.

There are four key transformations that we agree can and should be achieved by a joint investment:

1. Data generated by the whole of the NCRIS investment should provide **a world leading data advantage** to enable research and industry to work on challenges in new ways because piecemeal, discipline specific, and tightly controlled data loses opportunity;

2. **Innovation is simpler** - researchers should be able to create the data tools and services that they need, from reliable, configurable foundations, without needing to build from the ground up. In providing this, the innovation burden is greatly reduced, and a platform for collaboration is provided that enables partnerships across research and industry;

3. **Collaboration for borderless research** nationally and internationally – research communities should have an environment to work in a data rich environment with all of the underlying services to enable partnership; and

4. **Enhanced translation of research** through reliable and available outputs of research – including data, methods and models - enabling translation across industry, policy, and problem areas. Institutions and research communities can build research strategy and partnerships without the reputational risk of unreliable data.

In particular, the NeCTAR program has identified from engaging with research communities:

- **The need for infrastructure that is responsive** to the evolving needs of research communities, national research priorities, and the ability to respond to emerging international and industry opportunities.
- **The need for research priority led infrastructure** in planning and development, enabling national access to tools supporting advanced methods, software platforms and virtual laboratories, cloud-based storage and computation.

Data for Research is Active Data.

NeCTAR encourages consideration of a focus on an approach which supports more active interrogation of data by Australian research and industry.

A new generation of **Active Data Infrastructure** in Australia will accelerate the generation of new knowledge by Australian researchers and industry by supporting:

- Shared access to pre-deployed advanced research informatics capabilities; and

- Access to integrated computing capacity to support new and novel analytic capability emerging from research and industry.

Such an approach will significantly increase the value to research and industry of data generated by NCRIS infrastructures and other national research activities. Active Data infrastructure can be underpinned by infrastructure investments in coupled compute and storage, including cloud compute and storage capability as well as co-deployed HPC and storage facilities.

Virtual Laboratories: A number of the NeCTAR Virtual Laboratories have proven the value of access to pre-deployed, sector prioritised analytics and modelling platforms and the competitive edge this provides to Australian research. The experience of the NeCTAR Virtual Laboratories has highlighted:

- The proven value of providing access to pre-provisioned advanced research informatics capability;
- The potential barriers to applying advanced analytics to data held in traditional, passive data repository infrastructures which do not support the ability to apply computing over the data - at scale; and
- The need to partner with research communities (and NCRIS research-domain investments) to prioritise and support access to the specialised informatics capabilities of each research domain.

Australian Science Clouds are an *Active Data* investment: This approach to a national data infrastructure, which recognises the value of coupling research-domain focussed analytical and modelling tools with an underlying cloud compute and storage infrastructure, is recognised in the NeCTAR Agility Fund proposal to establish an *Australian BioScience Cloud*, an *Australian Marine Sciences Cloud* and an *Australian Ecosystems Science Cloud*. These also demonstrate the potential value arising from a deeper partnership between an NCRIS data investment and the research-domain and mission focussed NCRIS investments.

Active Data - Internationally and in Industry: Internationally, it is being increasingly recognised that research data infrastructure architectures should embrace the need to support increasingly sophisticated analytical and informatics capability to accelerate the generation of knowledge. The EU-led ELIXIR initiative is re-imagining a future Europe-wide interoperable bioinformatics infrastructure which provides access to deep analytical capability co-located with distributed bioinformatics data repositories. ELIXIR is exploring a model based on a network of OpenStack-based cloud sites - with opportunities for future interoperability with the NeCTAR Research Cloud.

Similarly, industry adoption of ***Big Data*** practices leverage large scale cloud computing and storage platforms **and** scalable computing platforms to apply powerful data analytics capability to complex data holdings. It was the emergence of Hadoop like scalable computing platforms and the use of cloud-style resources which accelerated the ***Big Data*** revolution in commerce and industry.

A Strategic Infrastructure to accelerate an Australian research informatics revolution: Research methods are becoming increasingly digital, as methods are being encoded in software to cope with the increasing scale and complexity of research data. Many research communities in Australia and internationally are exploring and adopting next-generation analytics and informatics capabilities, including large-scale machine learning technologies, automated image classification and advanced

statistical processing. These are arising in research domains from the Humanities to the Big Sciences. A national Data infrastructure which better supported analytics and informatics would accelerate this transformation in digital research methods in Australia. Improved access will both accelerate new research and broaden the exploitation of those assets (data and expert methods). A range of tools is needed, including those that enable the analysis of data that is crucial to research; ensure the integrity of data; and support knowledge exchange/dissemination/translation.

Were such an Active Data infrastructure approach to be adopted in Australia, NeCTAR would recommend this be underpinned by a strategic investment in building key **skills partnerships across research, eresearch, computer science and industry** to ensure Australian research and industry can exploit the opportunities arising from this digital informatics revolution.

Meaningful access to Data: An **Active Data Infrastructure** will facilitate meaningful access to data to enable the next generation of collaborations, of much larger groupings of researchers and other stakeholders nationally and internationally. Infrastructure is needed that enables the rapid and responsive delivery of storage, computation, and software platforms at national scale. Infrastructure allowing researchers to innovate rapidly and construct specific architectures (virtual laboratories, science clouds) which address specific challenges or interoperate with international initiatives as needed.

The Nectar Virtual Laboratories and Research Cloud programs have delivered national impact to a number of research communities, as detailed in the NeCTAR Impact report (https://nectar.org.au/about/reports/). *NeCTAR Impact* details the values of this investment in providing an online infrastructure that supports researchers to collaborate and share ideas and research outcomes; connect with colleagues in Australia and around the world; and ultimately contribute to our collective knowledge, in order to make a significant impact on our society.

From experience with this program Nectar has identified the opportunity and need to aggregate infrastructure planning and operational expertise to further support research communities, collaborations, and national priority areas. The opportunity and demand is for community led research data, research tools and expertise development. The immediate challenge is for the national e-infrastructure as a whole to respond. In engaging with a number of collaborating national platforms (BPA, TERN, IMOS) , and supported by NCRIS Agility Funding in 2016-2017, NeCTAR is partnering to explore national science/research cloud development. This “Cloud” concept - encompassing multiple clouds including commercial cloud, international science clouds, Australian research cloud, storage and computation, and cloud enabled research software and data - is the friction reducing platform that enables rapid access and interoperation at national scale, across borders, and across infrastructure investments.

Knowledge Sharing platforms to support research-industry collaboration: A substantial opportunity exists to deliver the benefits of NeCTAR services to support improved collaboration and knowledge sharing between Australian research and industry. The NeCTAR Virtual Laboratory program has delivered widely-accessible, research domain-focussed, knowledge sharing platforms. NeCTAR supports the Virtual Laboratories to enhance research-industry collaboration and industry access to streamlined

research tools, data and knowledge. The NeCTAR Research Cloud is a widely-accessible, interoperable national platform for hosting research tools and data online. The NeCTAR cloud has delivered a platform which significantly reduces barriers to rapidly sharing research data, tools and knowledge across institutional and national boundaries. For example, the Cancer Therapeutics Cooperative Research Centre (CTx) is Australian Government supported organisation focused on the discovery and development of novel cancer drugs for children and adults. During a period of finalising CTx's and Australia's largest pre-clinical deal to date, the selection of the right partner and right solution without business interruption was mandatory. The choice of NeCTAR as this partner highlights the quality and value of Australian research technology services and the additional benefit from investing in the widening the services going forward.

Nonetheless, barriers remain to seamless research-industry collaboration on the Research Cloud and in the Virtual Laboratories. The current identity and authentication services provided through the Australian Access Federation still provide barriers for seamless industry access and the Super Science provisions have previously created uncertainty within research communities on the terms for industry access to Research Cloud resources. For example, this has led the Virtual GeoSciences Laboratory partners to deploy a separate instance of the Virtual Laboratory on commercial cloud resources for use by industry partners. Clarity on access policies and streamlining of access systems will be important to realising the potential benefits to enhanced research-industry collaboration.

Question 34: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

In order for Australia to take full advantage of its research data and collaboration infrastructure, there is a need to adopt common solutions/platforms that leverage innovation across borders. These increase the agility of Australian researchers to connect, leverage and contribute to international outcomes. Collaboration with international research infrastructure is vital in two areas:

- International Virtual Laboratory-like investments, including the NSF-funded Science Gateways program and the EU Virtual Research Environments program; and
- The blossoming of significant international investments in *Science Clouds* over the past several years.

International networks of Virtual Laboratories, Virtual Research Environments and Science gateways: The success of Australia's NeCTAR-funded Virtual Laboratories reflects the international significance of similar initiatives, with USA's NSF-funded science gateways and Europe's Horizon 2020 funded Virtual Research Environments achieving significant outcomes. Collaboration with organisations such as the Science Gateways Community Initiative and International Coalition for Science Gateways will leverage off international momentum to build internationally interoperable ecosystems of these digital platforms.

International network of Science Clouds: NeCTAR's Australian Science Clouds will elevate international recognition and enhance linkages with international science platforms, including Europe's ELIXIR program and the UK's CLIMB microbial bioinformatics cloud, with the European Open Science Cloud, and with other international research clouds involved in the OpenStack Science Clouds working group.

Due to the Nectar Research Cloud, Australia is recognised as an international pioneer in deploying cloud computing platforms for research. In addition, Australian research communities having advanced access to cloud infrastructure, have established an international reputation in Virtual Laboratories and cloud-enabled research software platforms. This unique leadership position provides the opportunity to elevate Australian Science Clouds internationally.

Question 35: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Data for Research and Discoverability capability area?

In any design process there is a tension between providing a solution which is able to respond to all requirements, available resources, and competing drivers (such as national vs local benefit). It is highly desirable therefore to have a flexible and responsive architecture which is able to modify and cater for a variety of outcomes, particularly in the areas of compute and storage which are undergoing significant ongoing development. Creating an investment framework which reflects changes in demand will provide both solutions and strategic direction which meet specific and measurable requirements.

Research Cloud: As an immediate need, the original NeCTAR Research Cloud infrastructure, established under the Super Science EIF funding, will exceed the 3 year replacement cycle in the near future, with much of the equipment now older than 5 years. A renewal of capacity will be required to support the priority needs of the instrument and data-intensive NCRIS capabilities and existing research users of the Research Cloud aligned to national research priorities. Based on NeCTAR experience, direct sector co-investment in the combined capital and operating budgets can be expected. Both NeCTAR and RDS propose a more streamlined, interoperable future investment in cloud computing and storage infrastructure.

A Multi-Cloud Strategy - Leveraging Sector-based and Commercial Resource Providers: As the NeCTAR Research Cloud infrastructure was established under the Super Science investment, it was not possible to leverage commercial cloud resource providers under the NeCTAR program. At that time there were also no large public cloud providers operating on-shore facilities in Australia. Following establishment of significant on-shore commercial cloud providers, a potential refresh of the research cloud infrastructure to support a *Data for Research and Discoverability* investment should consider the opportunity presented by commercial cloud providers.

NeCTAR is currently developing a *Multi-Cloud Strategy* to inform a possible investment in a national research cloud platform.

We note that careful consideration needs to be given to the balance of investment in access to sector-based resource providers and commercial cloud providers. Nectar would argue that there would be substantial risk at this time in planning for Australian research communities, universities and NCRIS capabilities to rely entirely on commercial cloud providers for hosting their research data.

These risks are also recognised across industry, with the emergence of *Hybrid Cloud* approaches identified as the fastest growing segment in cloud. *Hybrid Cloud* (or *Multi-Cloud)* utilises private, or community cloud infrastructure as well as public cloud providers. Up to 50% of Fortune 500 companies now have access to their own internal OpenStack-based cloud infrastructure, as well as access to public

cloud providers. Major global companies are exploiting OpenStack cloud infrastructure, including Walmart, Paypal, eBay, Sony, and Time-Warner Cable, especially where they have complex data analytics requirements.

NeCTAR would advocate for a program of investment that supports access to sector-based providers as well as partnerships with commercial providers. NeCTAR has established links with major cloud providers, including Amazon Web Services, Microsoft Azure and Rackspace.

Virtual Laboratories: NeCTAR also regards the establishment of the current cohort of Virtual Laboratories under the Super Science scheme as an investment in the development of national infrastructure. These have proven to be of recognised high value, sustainable and to have successfully elicited ongoing sectoral investment. There is very broad support for the Virtual Laboratory program among the research-domain focussed NCRIS investments and high levels of ongoing support from research institutions and research agencies.

Noting that much has been learnt from the original investment in what was then regarded as a novel infrastructure investment, NeCTAR would recommend a program of further national investment in the development of Virtual Laboratory-like infrastructure. Such an investment would further consolidate research community and sector investment in national platforms for research analytics. These should be aligned with investment in data infrastructure and other NCRIS research infrastructure investments. The Virtual Laboratory model has a proven ability to attract significant co-investment from the sector.

An Australian Data capability, as identified by the issues paper, must serve both "Research and Discovery". These infrastructure concerns may compete in some cases and an appropriate balance of competing concerns must be ensured. Specifically, the data infrastructure must support both:

- rapid innovation, development and immediate research need; and
- broader access to sophisticated digital research methods and exploitation of data assets.

A joint or coordinated investment in *Data for Research and Discoverability* should ensure an appropriate balance. We believe the "Ideal research data system" illustrated in the issues paper may represent a mature research community environment which supports broader, trusted access and reuse. However, it does not represent rapid innovation needs, direct storage/compute/cloud infrastructure support for specific or large-scale research community needs, R&D of research infrastructure (development projects, establishment work, pre-validation), research community led software architecture development, nor necessarily the ability to interoperate with emerging international infrastructures.

Underpinning Research Infrastructure

> **Question 30: Are the identified emerging directions and research infrastructure capabilities for Underpinning Research Infrastructure right? Are there any missing or additional needed?**

These capabilities are critical to Australian research, with clear benefit arising from and wide support for a national peak investment in High Performance Computing. Research infrastructure relying heavily on networks and identity. Provision of access to peak compute jointly with cloud compute adds considerable value to the research community.

The NeCTAR Virtual Laboratory and Research Cloud programs have developed strong mutually supportive arrangements with a number of the identified underpinning capabilities. To maximise the value of the investments in these areas it will be important that future arrangements are supportive of such alignment and considerate of the mutual dependences created.

High Performance Computing is a critical component of a national research and innovation system. NeCTAR agrees that a significant national investment is required in well-managed national Peak Computing facilities along with Tier 2 HPC facilities. NeCTAR also supports the proposition for consolidation of governance of the national peak computing facilities.

We note the close relationship between the NeCTAR programs and the national Peak HPC facilities. NCI and Pawsey are operators of two nodes of the NeCTAR Research Cloud; providing access to computational and data resources which complement the investments in the HPC computing facilities.

NeCTAR shares with the national HPC investments a keen understanding of the importance of support for research computing and digital methods closely coupled with data infrastructure. This is reflected in NCI's role as a key enabler and champion of a number of the NeCTAR Virtual Laboratories.

While we understand the need to consider the particular needs for targeted investment in Peak HPC facilities, we recognise the importance for strategic alignment and coordinated delivery of HPC and cloud capabilities. We advocate that the national research capability will be well served by a well-aligned and deeply connected *Data for Research and Discoverability* investment and the national *High Performance Computing* investment. Many national research communities and national data holdings are well supported by the peak HPC investments. The ability to easily share access to data, analytics and modelling capabilities across the two investments would substantially raise the value of both investments. Internationally, we are seeing increasing interest in the optimal exploitation of peak HPC facilities and cloud computing infrastructures.

NeCTAR also identifies sector interest in exploring opportunities for exploiting distributed cloud resources to support aspects of Tier 2 HPC requirements. The University of Melbourne and Massive have recently deployed HPC capabilities which leverage or integrate with NeCTAR cloud platforms.

Australian Access Federation (AAF): The national authentication and identity systems offer by the **AAF** have been crucial to the uptake and accessibility of the Research Cloud and the Virtual Laboratories. NeCTAR and the AAF have jointly funded successful initiatives to extend the capabilities of the AAF to support emerging needs from the Virtual Laboratory research community.

In particular, NeCTAR infrastructure has a high level of reliance on AAF capabilities and supports that the AAF should be resourced to:

- Further extend capabilities to meet challenging needs for nationally federated and delegated authorisation capabilities;
- Interoperability with Industry standard authentication protocols, including OpenID Connect; and
- Support internationally federated research identity exchange capabilities such as EduGain.

Digitisation: It is agreed that digitisation would benefit from national coordination and funding, and this should be coupled to the proposed *Data for Research and Discoverability* capability.

Geospatial System needs would benefit from further clarification of focus areas, but clearly should also be coupled to a national *Data for Research and Discoverability* capability.

General Considerations

Question 3: Should national research infrastructure investment assist with access to international facilities?

It is important that national research infrastructure investments in a data capability facilitate international collaboration, to elevate recognition on the world stage and enhance linkages with international science platforms. As the international research community moves towards achievement of a global infrastructure for data, Australian infrastructure needs to both meet the needs of local users, and act as a reference point for international partnerships.

The ELIXIR program provides one example of the value generated by international linkages. ELIXIR, the European infrastructure for biological information, is building a portal to bioinformatics resources world-wide for users in academia and industry. Australia's EMBL-ABR is currently collaborating with ELIXIR and others in the collection and dissemination of Australian bioinformatics tools as well as broader tools of relevance to the Australian life science community. National research infrastructure investment in international initiatives such as this will ensure that Australian researchers can contribute to the design, development and innovation that occurs when developing digital research platforms, in addition to collaborating on usage to facilitate increased international linkage of data.

Question 5: Should research workforce skills be considered a research infrastructure issue?

Research workforce skills have been repeatedly highlighted in sector review, and are a research infrastructure issue. To enable Australia to continue to leader internationally in research data and collaboration infrastructure, this is a key area where focussed attention and stable investment is essential.

Question 6: How can national research infrastructure assist in training and skills development?

National research infrastructure in eResearch can assist through provision of programs supporting development of underpinning capability in the following areas:

- Software engineers and Data and Collaboration technologists - skilled in the assembly and development of infrastructure, platforms, and tools to facilitate researcher data collaboration; and
- Data scientists/informaticians - skilled in working alongside researchers in extracting meaning from data; and
- Data policy practitioners - people skilled in the development and application of data policy in research institutions, research facilities, and the research system.

- Data stewards - skilled in data management, curation, collection development, and the application of data policy.

The cohorts would collectively develop essential expertise in:

- Data infrastructure, services and management;
- Software infrastructure development;
- National/central operations, coordination & support in cloud fabric management & services (supporting cost effective national distributed infrastructure partnerships);
- Storage hardware and software configuration management and deployment; and
- Development & operation of combinations of these infrastructures and delivery at scale.

There is also a range of more domain specific services that could be provided and/or coordinated centrally, such as bioinformatics.

Similarly, NeCTAR's Australian Science Clouds will elevate international recognition and enhance linkages with international science platforms, including Europe's ELIXIR program and the UK's CLIMB microbial bioinformatics cloud, with the European Open Science Cloud, and with other international research clouds involved in the OpenStack Science Clouds working group.