

Research Data Services

Room 506, Axon Building,
The University of Queensland, QLD, 4072



Wednesday, September 28, 2016

Dr Alan Finkel AO,
Australia's Chief Scientist

Dear Alan,

We have great pleasure in attaching the Research Data Services submission to the National Research Infrastructure Capability Issues Paper.

We look forward to consulting with you and the panel over the coming months in relation to the comments and suggestions contained in the submission.

Regards,

A handwritten signature in black ink, appearing to read 'Brian Anker', written in a cursive style.

Mr Brian Anker
Chair, RDS Board

Submission

2016 National Research Infrastructure Roadmap

Capability Issues Paper

Name	Mr Brian Anker
Title/role	Chair, RDS Project Board
Organisation	NCRIS Research Data Services (RDS)

1. General Commentary

The Research Data Services Advisory Board, the University of Queensland (as lead agent) and the RDS Project Office are all committed to build on the on the key achievements of RDS and its predecessor RDSI and work with fellow NCRIS capabilities and in particular ANDS and Nectar in ensuring data is accessible, shared collaboratively and of maximum value for both the public and private sectors. This alignment of activity should go beyond just the current joint activities and shared projects and embrace the creation of a jointly-developed investment opportunity for the Australian Government, research community, individual institutions and agencies, and state governments. This initiative would allow for a smooth transition from the current three project approach to a jointly-developed investment opportunity providing confidence in continuity of service where needed, managed migrations to new services, and service retirement where appropriate.

The purpose of an investment in *Data for Research and Discoverability* is to place Australia as a world leading locus for data intensive research. The initial NCRIS RDSI Project delivered a national infrastructure where there was none, able to hold and make available, at scale, significant collections of research data. The subsequent RDS project has worked closely with research communities and partner capabilities including Nectar and ANDS to build flexible and powerful data sharing and management platforms upon these resources, enabling researchers to find, access, use and reuse data collections that were previously either unavailable to them or difficult to access.

RDSI and RDS have been background entities operating through a federated structure of nodes. This has led to considerable cohesion of purpose among previously disconnected eResearch groups, but with the downside that connectivity to the research communities served is indirect and through intermediary service providers.

The experiences of the Super Science Initiative and subsequent NCRIS strategic funding have provided significant and lasting benefit to research communities and have matured to the stage where the existing eResearch infrastructure environment, with data storage coupled organically rather than by optimal design with compute and curation functions, faces challenges in dealing with a rapidly changing e-infrastructure ecosystem.

To provide the optimal outcomes of *connecting the data environment*, there are four key transformations that can and should be achieved:

1. Data generated across the whole of the NCRIS investments 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 should be accelerated - 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. Facilitation of collaboration for borderless research nationally and internationally – research communities should work in an agile, discoverable, and 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 confident in a robust underlying environment of reliable data, tools, and methods.

These transformations require partnerships between the NCRIS data generating and discovery capabilities, research communities, NCRIS domain capabilities, underlying capabilities, research institutions and industry.

2. Specific Commentary

Question 1: Are there other capability areas that should be considered?

Nil Response.

Question 2: Are these governance characteristics appropriate and are there other factors that should be considered for optimal governance for national research infrastructure.

There is an opportunity to create an inclusive governance mechanism that encourages both strategic and tactical engagement across the various levels of endeavor thus allowing for strategy development to be interactive with program development and project delivery to ensure real time communication, feedback and reporting. This nimble and informed governance structure would build on previous experiences and aim to minimise transaction costs while creating a responsive and demand-driven service delivery culture.

Of critical importance is that a governance structure should foster an organisational culture that clearly demonstrates and embraces the key principles of having:

- A true culture of collaboration, responsiveness, and engagement across the sector,
- An ability to proactively encourage and seek out co-investment both in terms of dollars and in-kind, especially from Universities,
- A desire to participate both with the public and private sector organisations,
- A commitment to flexibility of its service mix to cater for a dynamic and changing technological environment,
- A legal framework that ensures a low transaction cost for all players,

- An ethos of partnering with institutions and government agencies to provide input on strategic data issues while not being dictatorial or disenfranchising in its approach, and
- A willingness to embrace and work with new entrants, stakeholders and providers (both nationally and internationally) to create value for its community.

Other factors for governance that should be taken into account include:

- governance for scarce resources versus governance for whole of system benefit will have different characteristics and membership. For example governance of scarce resources might require appropriate allocation schemes for time on an instrument or access to nationally funded resources; whereas governance for whole of system benefit, including addressing national research priorities, increasing the value and access to data, accelerating research and creating a borderless collaborative research environment will require representative advisory groups that build strong collaborative ties across organisational boundaries and disciplines.
- governance should generally consider both effective collaboration, and whole of system coherence (including forecasting and planning), requiring that the NCRIS investment as a whole is optimal and can attract co-investment by institutions where that co-investment is aligned and/or complementary.
- all governance entities across the system should have a clear requirement to both develop and foster a culture of collaboration as well as ensuring that the necessary interoperability and international connections exist to support borderless research.

It is also important to highlight the point that national investment needs to work hand-in-hand with local investment to match supply to demand as well as to leverage local commitment to national solutions, so any governance arrangement must aim for cohesion between these very different entities.

At a project level, RDS, Nectar and ANDS have been working closely together to harmonise current activities. Looking to the future there are clear opportunities for these projects to continue this close collaboration and furthermore consider developing a single investment opportunity. In this regard, please see further commentary on this construct in the “Other comments” section at the close of this document.

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

There is a direct correlation between international partners and the quality of research. It is thus vital that Australian researchers are easily able to collaborate with international partners and share research data. In support of this Australia should be an active participant in international collaboration support and data initiatives. In addition to alignment in the way in which researchers interact with infrastructure (ensuring a common/compatible user experience), it is valuable for infrastructure providers in Australia to coordinate with international approaches around development, support and provision of services. This ensures that Australia uses its money wisely on adapting/complementing, with developing *de novo* as a last resort.

Question 4: What are the conditions or scenarios where access to international facilities should be prioritised over developing national facilities?

In data and collaboration support, national and international capability and resource provisioning must be considered together where it makes sense – Australia must contribute to and engage with

international initiatives, such as the relevant international Science Platforms, international Virtual Laboratories (science gateways) community, and the Research Data Alliance.

A core principle should be that the default approach is that a collaborative solution is adopted (either international, national, inter-institutional, or inter-project) and that taking an alternative approach should require some justification.

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

Yes. In order for Australia to take full advantage of its research data and collaboration infrastructure and to demonstrate international leadership in key areas, research workforce skills should be considered research infrastructure, adding significant value to infrastructure assets.

There is a wider need for data management, analytics and software skills in the research system as a whole, in particular in developing research data skills in students, early career researchers, and research leaders. The approach we advocate is one of support and guidance of others in developing those skills local to the research itself. This could involve providing access to cohorts of skilled professionals in these areas to lift capability within the research system as a whole. For example, researchers in the genetics area will very often need to work with bioinformaticians. One possible form of training would be to help such researchers interact more effectively with bioinformatics experts so as to more effectively use the time of this scarce resource.

This creation of inter-disciplinary teams where participants leverage each other's skills will magnify the benefit, for example with bioinformaticians learning as much from geneticists as geneticists will gain from access to bioinformaticians.

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

National research infrastructure in eResearch must include a professional workforce skilled in providing services and partnering with researchers, as well as assisting through provision of a broad skills program in the following areas:

- Data stewards - skilled in data management, curation, collection development, and the application of data policy;
- Software engineers and Data and Collaboration technologists - skilled in the assembly and development of infrastructure, platforms, and tools to facilitate researcher data collaboration;
- 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.

The cohorts would collectively have underpinning skills 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.

Australia currently has unique advantages in this area, but it is not yet systematic. A strong partnership between NCRIS, the research institutions and the research communities needs to be provided systematically, enabling the use of eResearch tools like shared data collections accessible through Virtual Laboratories in undergraduate teaching programs. There is also a range of more domain specific services that could be provided and/or coordinated centrally, such as bioinformatics, marine, terrestrial systems etc.

Question 7: What responsibility should research institutions have in supporting the development of infrastructure ready researchers and technical specialists?

Research institutions have specific responsibilities for their research publications, data, and other research outputs, and thus it is the case that they need to provide the policy procedure, technical infrastructure, and professional services to meet their obligations, and achieve their research ambitions. However doing so individually loses coherence, and national advantage.

It is thus important that research institutions are active partners in a national approach to infrastructure and are supported by professional services that enable them to make the investment decisions that suit their priorities, in the context of a nationally coherent approach.

To achieve this, infrastructure funded and developed under the NCRIS program must be built in a manner which is responsive to the needs of both researchers and their institutions. This includes having sufficient interoperability and international connections (where relevant) so that the preferred and more attractive pathway for institutions is to invest in a manner which is interoperable with the national infrastructure or to co-invest in the national infrastructure. An essential element of this is that national infrastructure must be seen as being something that can be relied upon with long term funding certainty. The recent approach of year on year funding of NCRIS has demonstrated that without this certainty there is no incentive for co-investment and a disincentive to engage as the durability/sustainability of service is uncertain.

Question 8: What principles should be applied for access to national research infrastructure, and are there situations when these should not apply?

Day to day access to NCRIS capability resources should be informed by a variety of approaches to scarce resource management, each appropriate for that particular resource type. However access to the outcomes of the use of infrastructure should be informed by a principle that the outcomes of use of publicly funded infrastructure should be publicly available. In particular there should be a requirement that data generated by any NCRIS facility should be generally available in a high quality form, informed by national and international policy.

Question 9: What should the criteria and funding arrangements for defunding or decommissioning look like?

Nil Response.

Question 10: What financing models should the Government consider to support investment in national research infrastructure?

Nil Response.

Question 11: When should capabilities be expected to address standard and accreditation requirements?

While the issues of monitoring, compliance and accreditation processes could be complex, capabilities should strive to meet the accepted standards for their discipline as soon as they become available.

The data generated by NCRIS facilities should be of high quality and support the ability to reproduce research and enable translation to commercial, public policy and further research. Australia needs to fund the creation and sustainable operation of infrastructures that facilitate capabilities to respond to this requirement, with appropriate accreditation of components where this is relevant and supports quality, reproducibility and translation.

Question 12: Are there international or global models that represent best practice for national research infrastructure that could be considered?

Australia is considered internationally to have best practice in research data infrastructure – its approach to partnership with institutions on data management, its ability to support the full lifecycle of data from capture to management, to collaboration, to analysis and publication. While Australia can and should do better, its strengths should be maintained, and its eResearch infrastructure leadership should engage with overseas equivalents to ensure two-way transfer of good practice.

Question 13: In considering whole of life investment including decommissioning or defunding for national research infrastructure are there examples domestic or international that should be examined?

Nil Response

Question 14: Are there alternative financing options, including international models that the Government could consider to support investment in national research infrastructure?

As the research data generated by NCRIS facilities becomes more and more valuable, there will be a temptation to build subscription models for access to this data. There will also be a temptation to implement user pays models for the consumption of storage and cloud. Neither have been successful and both inhibit innovation. Thus, models of financing inputs should be explored, but in keeping with many reports, the data and tools produced by research should be as freely available as possible to support the widest level of translation.

Health and Medical Sciences

Question 15: Are the identified emerging directions and research infrastructure capabilities for Health and Medical Sciences right? Are there any missing or additional needed?

Nil Response.

Question 16: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Nil Response.

Question 17: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Health and Medical Sciences capability area?

Nil Response.

Environment and Natural Resource Management

Question 18: Are the identified emerging directions and research infrastructure capabilities for Environment and Natural Resource Management right? Are there any missing or additional needed?

Nil Response.

Question 19: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Nil Response.

Question 20: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Environment and Natural Resource Management capability area?

Nil Response.

Advanced Physics, Chemistry, Mathematics and Materials

Question 21: Are the identified emerging directions and research infrastructure capabilities for Advanced Physics, Chemistry, Mathematics and Materials right? Are there any missing or additional needed?

Nil Response.

Question 22: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Nil Response.

Question 23: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Advanced Physics, Chemistry, Mathematics and Materials capability area?

Nil Response.

Understanding Cultures and Communities

Question 24: Are the identified emerging directions and research infrastructure capabilities for Understanding Cultures and Communities right? Are there any missing or additional needed?

Nil Response.

Question 25: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Nil Response.

Question 26: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Understanding Cultures and Communities capability area?

Nil Response.

National Security

Question 27: Are the identified emerging directions and research infrastructure capabilities for National Security right? Are there any missing or additional needed?

Nil Response.

Question 28: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Nil Response.

Question 29: Is there anything else that needs to be included or considered in the 2016 Roadmap for the National Security capability area?

Nil Response.

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?

There are a number of important components of a rich research environment discussed in this section that comprise infrastructure that underpins broad areas of research or in some cases all research, including Access and authentication, High capacity networks, and digitisation. These touch not only every NCRIS capability but every activity engaged in by any researcher anywhere, particularly in the cases of networks and identity.

There are also significant, if somewhat more complex, areas where development of our national Access and Authentication system can be addressed. We have the opportunity to provide a leading global position on developing robust and reliable methods of providing persistent unique identifiers to individuals or projects, an area of nascent development internationally. We are also well positioned to work in partnership with the Australian Access Federation to develop a national access and authorisation framework (described as Access in this paper) capable of dealing with a variety of service types, a capability which currently exists in a simple form. The areas of identity and authentication are highly topical, difficult, and provide immense boosts to collaborative activities but it would be premature to describe Australia's current ecosystem of identity providers and consumers as a coherent whole.

The other items are valuable areas in which strategic funding will provide significant national benefit, but would more usefully be placed either in another section (Geospatial systems, and Neutron and x-ray scattering) or combined into a more broadly defined area. In particular High Performance Computing could be considered being really one end of the Data for Research and Discoverability spectrum and an important data partner. High Performance Computing needs to be closely aligned with Data for Research and Discoverability.

Question 31: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

It is essential that HPC, along with the investment in Data for Research and Discoverability are readily able to enable the research community to collaborate with their partners of choice in a coherent and seamless manner. As well as requiring a tight coupling of HPC and Data for Research

and Discoverability, this will require access methods via the AAF to be fully integrated into the broader international frameworks such as eduGAIN - something which is currently in pilot by the AAF and requiring further work. The European Science Cloud and the NSF Advanced Computing Initiative are current examples where there is likely to be value to be gained by engaging. The key element here however is that yet to be defined or currently unknown international initiatives will likely emerge over the course of the next 10 years. Sufficient agility needs to be built into the guidance system for NCRIS to enable capabilities to engage as required as opportunities that are in the interest of Australia emerge.

Question 32: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Underpinning Research Infrastructure capability area?

It is important that data generated in these capabilities generate nationally significant research data assets that support a broad class of nationally significant challenges.

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?

Research software infrastructure, including but not limited to sophisticated research methods and tools that support the integrity of data, knowledge exchange/dissemination/translation is another priority area for international collaboration, enabling researchers to utilise leading edge data and tools with data infrastructure that improves access to methods and data. This will enable Australian researchers to gain a significant competitive advantage. Initiatives that Australia should engage with include USA's NSF-funded Science Gateway Community Institute and a range of discipline specific science gateways, and Horizon 2020 initiatives such as Elixir.

In addition to the ongoing investment in the new combined capability, a significant capital investment should be made in a/set of integrated collaboration environment(s) available to all Australian researchers. This would include making available significant local storage for Australian reference data sets, co-located compute and discipline-appropriate collaboration and tool environments such as the Virtual Labs and Science Clouds and, where appropriate, local mirrors of overseas reference data sets. This should be complemented by experts able to assist research teams to make (the best) use of such an environment, noting a diversity of eResearch skills across the research base. Such a facility would have a transformative effect on Australian research by providing a coordinated and strategic allocation of appropriate resources tailored to large and small scale activities, and build on our decade of collaborative infrastructure investment in technology and data.

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

The approach to the transformation of data practice outlined in this submission is happening at the same time as a large transformation of the same form is happening in Europe – notably through the European Open Science Cloud, but also in supporting open collaborative laboratories through the Science Gateways program, and building data interoperability solutions through the mechanisms of the Research Data Alliance. RDS advocates the development of collaborations with these projects, in addition to a watching brief to ensure that Australia interfaces with relevant programs globally as they develop. Solutions from commercial cloud vendors also offer the opportunity for Australia to adopt a multi-cloud strategy resulting in more flexible supporting infrastructure for research particularly into the future.

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?

It is very important that an investment in data is holistic in nature so that data generation, data (re-) use and data publication are considered together. Secondly the investment should be done in a context where other investments in data generation must be coordinated so that there is an overarching transformation in the Australian data environment that dramatically transforms Australia's ability to partner in data intensive research addressing nationally significant challenges. It would also optimise reliability and interoperability, achieve reproducibility of research results and encourage engagement with industry.

The placement of high performance computing in the underlying capabilities section and separate from data for research and discoverability is one manner in which the eResearch Infrastructure system could be viewed, however high performance computing is an important and integral part of the data system as both a data generator and source of data products. It is essential that any roadmap recognises that high performance compute is an integral part of the data system.

The underlying premise of a Data for Research and Discoverability investment is that Australian research data has value, and we are aiming to multiply that value through alignment of existing services, development of new services, and broad accessibility of research data outputs to both the domestic and international research communities. Recent history has shown Australia is not immune to attempts at unauthorised access and one function of the Australian eResearch capability should be to develop, in collaboration with relevant Centres and groups, an overall security strategy that reflects the importance of data, tools, and methods to Australian researchers and the sophisticated nature of external threats. This strategy should be developed with input from external expert agencies that have experience in considering these issues and have developed robust responses, such as the Australian Signals Directorate.

Other comments

If you believe that there are issues not addressed in this Issues Paper or the associated questions, please provide your comments under this heading noting the overall 20-page limit of submissions.

As mentioned above, at a project level, RDS, Nectar and ANDS have been working closely together to harmonise current activities. At this stage of the road-mapping exercise it is premature to commit to a specific governance model for any new data-related initiatives currently under consideration especially noting that the projects (RDS, Nectar and ANDS) have not had any organised conversation around governance models, preferring to focus on the what and why rather than how questions, to this point.

However, given the scale and impact of a possible single investment, and the excellent progress made in discussing approaches, it would be valuable to commence building consensus with stakeholders on future governance and delivery. This work would, of course, be informed by experiences learnt from the current structures within RDS, ANDS and Nectar and also importantly provide an opportunity for a timely and smooth transition to a new construct while recognising existing legal arrangements are in place between the Department and Lead Agents and Lead Agents and delivery agents, nodes, and other parties.

It is possible to envisage a new structure for investment created as a partnership or joint venture (incorporated or otherwise) between existing lead agents with the flexibility to encourage other participation as appropriate. This vehicle could emerge as the three existing projects continue to deliver upon current agreements but with a joint “board” overseeing and guiding collaboration and positioning of the new service. Experience would suggest that private sector participation on any board would be important however just as important is representation of the public research sector both in terms of institutional understanding and embedding tangible collaboration. A single institutional Lead Agent approach, although having some benefits, may not be seen by the sector as truly independent and may indeed encourage a level of unnecessary redundancy in the system. These and other matters would need be considered further in progressing the design of any new construct between RDS, Nectar and ANDS.

A possible way forward therefore would be for RDS, ANDS and Nectar to commit to bring forward for Departmental consideration an agreed position on governance, structure, scope, role and participation in a new combined project. This approach would be timely (with action commencing in the near future) but also recognise that legal frameworks are already in play across the three existing projects with activities currently being actioned.

RDS is committed to working collaboratively with ANDS and Nectar and other stakeholders with a view to providing a viable project governance framework for Departmental consideration and input at the next Draft Roadmap stage, noting that any structure must be guided by the principle of being a new structure serving the broad research communities nationally.