Submission
2016 National Research Infrastructure Roadmap
Capability Issues Paper

Name/Organisation	**Australian Academy of Sciences** **National Committee for Data in Science (NCDS)**
Preferred contact	**Professor Jane Hunter** **Chair of the NCDS** **Email: j.hunter@uq.edu.au** **Phone: 07 33651092**

Recommendations for Investment in Research Data Infrastructure

The National Research Infrastructure Roadmap correctly emphasizes both the importance of data in transforming Australian research across all disciplines as well as the need for significant on-going investment in data infrastructure.

The National Committee for Data in Science strongly believes that the single major investment that the Australian Government should make to accelerate research and innovation in Australia is to invest in an ***Australian Open Research Cloud***. This would provide the Australian research community with a common, trusted cloud platform offering integrated computational, storage, data curation, analytical, visualization and modelling services. It would enable researchers to store, share and re-use data across disciplines, states and organizations.

The proposed Australian Open Research Cloud would build on existing investments in e-infrastructure but most significantly, it would leverage the knowledge and technologies generated from the multi-billion euro investment in the European Open Science Cloud[1].

The Australian Open Research Cloud would be transformative by:

- Promoting Open Data – making research data a national openly available asset that will boost Australia's competitiveness by benefitting start-ups, SMEs and data-driven research.
- Driving and accelerating innovation – the provision of cloud computing services seamlessly integrated with Big Data will unlock its potential, generate new data products and act as an incubator for new businesses and scientific activities.
- Promoting collaboration – the Australian Open Research Cloud will lead to new partnerships, better engagement and technology transfer between research, industry and the public sector.
- Providing an environment that will produce the next generation of knowledge workers, with skills in demand across a wide range of industry sectors.

The benefits for Australia's research, economy and society will be enormous. The Australian Open Research Cloud would enable Australia to become the partner of choice in international data-intensive research initiatives. The proposed world-class data infrastructure will ensure businesses, industry and the public sector reap the benefits of Big Data and will make Australia's research internationally transformative.

If such an investment is of interest to the Expert Working group, then The National Committee for Data in Science would be willing to contribute to the development of a more detailed paper proposing the establishment of an Australian Open Research Cloud.

[1] http://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

Questions

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

The key capabilities are covered.

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

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

Yes. To many researchers there is little difference between building new infrastructure here and gaining access to next generation infrastructure overseas.

From the data point of view, this involves national facilities providing access to shared international data sets as well as shared international infrastructure and services. It also involves learning from relevant international initiatives at both the generic level (e.g., European Open Science Cloud) and discipline-specific level (e.g., Astronomy, Bioinformatics, Geoinformatics) to leverage the resulting knowledge and technologies and to implement international best practice and state-of-the-art infrastructure in Australia. These are beneficial activities that should be funded.

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

Usually scale is the defining factor. For example, in astronomy a consortium of EU countries can build a much more powerful telescope than anything a single country alone (e.g. Australia) could build. To be competitive, countries need to join such consortia (or invest 10x more to develop a national facility) e.g., SKA, GBIF

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

Yes – particularly in the case of "research data", there is a need to invest in training and skills development because the infrastructure development, support and maintenance is heavily dependent on highly skilled data technologists and data scientists.

Currently NCRIS facilities do not take a leadership role in skills development. Training should be a key component of infrastructure. For example, national facilities should be encouraged to offer internship programs, or partner with universities to allow staff or PhD students to learn new skills under their guidance.

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

The skills shortage in Data Science/Data Technology is widely acknowledged in Australia and this shortage is anticipated to get much worse over the next 5-10 years.

Although providing discipline-specific scientists with data management skills is worthwhile, in many situations it is better to provide researchers with access to a pool of professionally trained information technologists, data technologists and computing professionals (e.g., software

engineers), who work alongside the discipline-specific researchers, developing data management solutions for them.

In the data capability, different types of professional training are required depending on the different roles required. Roles range from data stewards, to data technologists and software engineers to data scientists and data policy practitioners. In addition, career paths need to be defined and supported for skilled professionals working in these areas.

Broad expertise can be expanded through training programs that enhance the data management and data analysis skills of discipline-specific undergraduate and post-graduate students, early-career and mid-career researchers.

There is also a need for more cross-disciplinary training programs at the undergraduate and post-graduate level (e.g., in bio-informatics, eco-informatics, nano-informatics, geo-informatics).

The NIH in the US is currently funding a number of programs that are targeted specifically at developing the next generation of biomedical data scientists. These include:

- Development of online training resources in data science (e.g., MOOCs, Open Educational Resources);
- Institutional Training Grants - provide support for institutional programs that provide integrated training in computer science, the quantitative sciences, and biomedicine;
- Scientist and Early Career Development Awards - support mentored training of scientists who will gain the knowledge and skills to apply and develop new Big Data technologies, methods, and tools.

Although targeted at biomedical data, similar programs could be applied in other disciplines e.g., environmental data scientists, materials data scientists, social data scientists.

Another suggestion is that the ARC Industrial Transformation Training Centre program should encourage a proposal for a Training Centre in Big Data Management and Analytics.

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

Institutions will be critical in providing the training and the expertise required. Within the data infrastructure capability, one model that has been effective involves institutions or State-based eResearch facilities establishing a pool of shared expertise (comprising: data custodians, data technologists/software engineers, data scientists/informaticians, data policy practitioners) – staff who are available to migrate between projects and to work alongside scientists to help them develop their data management, analysis and visualization services. Institutions also need to define and facilitate career paths for skilled professionals working in these areas.

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

The FAIR set of guiding principles to make data Findable, Accessible, Interoperable, and Re-usable, should be applied in Australia. The value derived from the data, and hence return on investment, is maximised when the data is exposed to the maximum number of researchers. However, should businesses using data for profit, have to pay for to access research data? Selling data inhibits

innovation so should all data be freely available if its funded by the tax payer? If there is both sensitivity and value – how should this be balanced?

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

Where possible, and when a facility is still popular but can no longer be supported by Government funding, alternative options should be considered e.g., transfer of the cost of operations to industry/private partners, or migration to a more sustainable facility or perhaps a layered subscription model.

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

A key problem is that funding allocations are currently limited and timeframes are short. This makes long term infrastructure planning (and in particular, retention of good staff) difficult. The long-term stability, support and maintenance of national research infrastructure should be an important consideration in the funding model.

Other funding models include co-investment from State Governments, CSIRO, BoM, ABS, the Federal Government and Institutions. For example, the NCI facility is supported by co-investment from CSIRO, BoM and Geoscience Australia. Industry co-investment has been a challenge in the past so new models need to present convincing business cases that will motivate industry to co-invest.

One approach is to prioritize data management through SME R&D or ask industries facing skills shortages to co-sponsor training.

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

Within the Data area, if international standards are available then they should be adopted, to facilitate discovery, sharing, re-use and interoperability.

The challenge is that: many research areas do not have agreed standards; or the standards are only marginally useful for the goals the researchers are trying to achieve; or no one can agree on what the standard should be. Only when the standards in a research discipline are mature enough should they be enforced. Ideally or where possible, instruments (acquired or developed using government funding) should be generating data that can be exported or migrated to a non-proprietary open access formats.

Accreditation is relevant to 3 areas in the data capability area:

- Accreditation of data sets or trusted data repositories (what standards does it conform to, what QA/QC processes were applied to the data, data profiling, data benchmarking, etc.)
- Accreditation of data centres (e.g., WDS (Ionospheric Prediction Centre, AADC))
- Accreditation of training? – relevant but who funds it (Dept of Education?)

In each of these areas, accreditation criteria and processes need to be defined.

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

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?

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

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?

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

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?

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?

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

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?

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?

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

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?

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?

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

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?

National Security

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

Cyber-security is very different to Biosecurity, Energy security and Water security, so ideally should not be located under this capability. The challenges, technologies, infrastructure and expertise requirements are very different. Ideally *Cyber-Security* should be located within the Capability *Underpinning Research Infrastructure*, as it is closely related to: "*Trusted Communication – Access and Authentication*", *High Capacity Networks* and *High Performance Computing.*

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

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

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?

It makes sense for this capability to focus primarily on computing and IT infrastructure and comprise the following:

- High Performance Computing
- High Capacity networks
- Trusted communication (access and authentication)
- Cyber-security

Digitization – this is probably better aligned with the HASS disciplines under *Understanding Cultures and Communities.* It may also be better to describe the digitization infrastructure as "Digital Collections Development", which is necessary to enhance the accessibility and re-use of research artefacts in physical form through digitization and reduce the costs associated with managing national collections.

The following capabilities - Neutron and x-ray scattering – do not really fit in this category. Neutron and x-ray scattering are better suited to the Capability "*Advanced Physics, Chemistry, Mathematics and Materials*" – alongside ANSTO and AMMRF. The Australian Synchrotron, OPAL Reactor, Australian Centre for Neutron Scattering, should ideally also be located under the *Advanced Physics, Chemistry, Mathematics and Materials* capability , with ANSTO (see Table on page 31).

If Geospatial Systems refers to the National Positioning Infrastructure Capability, then this should be renamed to be specific. AuScope is also listed under the *Environment and Natural Resource Management* capability, as existing infrastructure in the Table on page 26.

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

The European Open Science Cloud[2]

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

The two capabilities "Underpinning Research Infrastructure" and "Data for Research and Discoverability" are closely related and should also be coordinated.

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?

See page 1 of this submission. The National Committee for Data Science recommends that the Australian Government should invest in an **Australian Open Research Cloud**.

The existing National Data Infrastructure model enabled users to discover a dataset, download it to their desktop and process it locally. The new model should allow researchers to access multiple datasets programmatically, dynamically integrate them, process the integrated datasets using services on the cloud and upload the derived data products (plus their provenance metadata) back to a repository on the cloud.

There is a need to move from data storage at the collection-level or file level to curated data that can be re-used at the record-level. Hence the focus should be less on "Discoverability" and more on services to support re-use. One suggestion is to change the title to "Data and Services for Research".

Researchers need tools that enable them to: (i) identify data sub-sets of most relevance; (ii) support programmatic retrieval of data via APIs; (iii) integrate datasets across organisations and disciplines; (iii) confidently re-use the data because sufficient provenance metadata has been provided and (iv) it is certified as high quality.

Additional services that have been identified as significant but that are currently not supported or poorly supported include:

- Metadata services that record sufficient provenance information to support re-use. To date, the majority of metadata services have focussed solely on collection-level discovery metadata (e.g., RIF-CS, Research Data Australia).
- Metadata and Persistent Unique Identifier or Digital Object Identifier (DOI) services that support the discovery and re-use of individual records or sub-sets of data

[2] http://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

collections/databases. Existing services focus on discovery and retrieval of an entire dataset or data collection – not fine-grained sub-sets.

- Standards and services for capturing the provenance of research data – including for a range of research data types (observational, monitoring, experimental, derived and simulated data).
- Digital preservation services – that automatically identifies valuable at-risk data and migrates it to new accessible formats.
- Certification services - that verify the trustworthiness and quality of data or data products provided by specific data centres or data service providers (eg WDS Certification) and facilitate the re-usability of the data.
- Linked Data National Infrastructure - services to link data across domain-specific data repositories to facilitate cross-fertilization and solve cross-disciplinary challenges.
- Semantic Interoperability, Ontological and Inferencing services to support cross-disciplinary data integration and reasoning.
- Institutional Repositories to support the long-tail of the research community.
- APIs to major data collections (such as ABS data) to enable programmatic access, retrieval and re-use.
- Data Processing Workflows and Workflow services.
- Data Publishing and Citation services.

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

- European Open Science Cloud
- ICSU CODATA – Committee on Data for Science and Technology.
- ICSU World Data System – Trusted Data Services for Global Science
- Research Data Alliance (RDA) Working Groups and Interest Groups
- GeRDI – Generic Research Data Infrastructure – Linked Research Data Infrastructure project in Germany
- International Data Citation efforts – DataCite, re3data.org, Databib, FORCE11

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?

Better overall synchronization and coordination of the Data capabilities is essential to: reduce duplication; to improve integration and streamline interoperability between facilities; and to improve outreach to research communities via a common communication and marketing strategy. Recently ANDS, NeCTAR and RDS have been discussing areas where they can partner to provide a more integrated approach to data and service infrastructure. The communication and coordination of service provision between ANDS, NeCTAR and RDS could be further improved by appointing a common governance structure or umbrella committee to oversee the implementation of jointly funded projects.

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.

One of the key challenges associated with the Research Data Capability is the tension between developing generic infrastructure that can be re-used across disciplines and demand for discipline-specific tools and services that have been precisely tailored for a single community. Although the second approach is more likely to generate useful tools for established discipline-specific fundamental research communities, it does not support the long-tail, smaller research groups.

One approach for providing research data infrastructure for a wide range of research communities is to focus investment in three categories:

1. Thematic Big data centres designed to satisfy the needs of a specific community generating data at the TB or PB scale (eg Genomics, Marine, Climate)
2. Virtual Laboratories – that don't necessarily produce Big data but need domain-specific data repositories and provision of domain-specific services for QA/QC, curation, analysis etc.
3. Institutional repositories – common interoperable repository software that will support mid-tail and long-tail researchers in institutions and that can be shared across institutions. This requires both local investment but also access to a shared, common approach and coherence of skills across institutions. Jisc in the UK are currently running a Shared Repository Pilot Project[3] across a number of universities, which is relevant and should be monitored.

Significant investment has been made in categories 1 and 2 above. This needs to continue, expand into new disciplines and also focus on developing links across data centres and VLs to support cross-disciplinary research. To date, limited investment has been made in Category 3 so ideally this should be the focus of future investment.

[3] https://www.jisc.ac.uk/rd/projects/research-data-shared-service