Status Report on the National eResearch Capability

The Australian Department of Education is developing a report on the status of national eResearch infrastructure in Australia, funded under the National Collaborative Research Infrastructure Strategy (NCRIS) program and related programs, as of year's end 2014. The main elements of this eResearch Infrastructure have developed after investments made between 2006 and 2014 in nine Projects and Services as follows:

1. ASSDA Services for e Social Science (ASeSS) - This project was established in 2008 to deliver a national provision for the collection and preservation of computer readable data relating to social, political and economic affairs and to make these data available for further analysis.

2. **Australian Access Federation (AAF)** - This project provides a framework and support infrastructure to allow trusted electronic communications and collaboration within and between universities and research institutions in Australia and overseas.

3. **Australian National Data Service (ANDS)** - This service seeks to transform Australia's research data environment through better managing, connecting, enabling discovery and supporting the multiple use of data.

4. Australian Research Collaboration Service - also known as ARCS this service sought to provide the research community with world-class tools and services to advance research capabilities. It was wound up in 2010 (?) and its role was subsumed into the NeCTAR and RDSI component capabilities.

5. **National Computation Infrastructure and Climate High Performance Computing (HPC) Project** - This project has been building an internationally significant high performance computer facility which will prioritise Australian research in climate change, earth systems science, and national water management.

6. **National eResearch Collaboration, Tools and Resources (NeCTAR)** - This project is creating national infrastructure to provide Australian researchers with access to a full suite of digitally enabled data and analytic and modelling resources, relevant to their research, at their desktop.

7. **The National Research Network (NRN) Project**, - This Project comprises eight Component Projects to build network infrastructure to extend and enhance the Australian Research and Education Network (AREN).

8. **Pawsey Centre** - this Project has been building high performance computer facilities which will prioritise research in geosciences and radio astronomy, and process data including that produced by the Australian Square Kilometre Array Pathfinder radio telescope.

9. **Research Data Storage Infrastructure (RDSI)** - This Project is building a national network of data storage that will improve the availability, management and sharing of data, to support data intensive research.

Of the above, the seven marked in bold are a particular focus for this Status Report.

The Status Report will take a strategic view of investments to ascertain how relevant and critical they are in supporting Australian researchers, the relationship between these investments and other research infrastructure and research investments, and how the capabilities and expectations of users have evolved and been addressed over time. The findings of the Status Report will inform future consideration of research infrastructure investment and delivery, and in the short-term will be used as input to the 2014 efficiency review of NCRIS projects and facilities.

Considering these investments and the services and projects that they have developed and/or are delivering, you are asked to consider and respond on the following matters and issues. It is understood that you will be more familiar with some projects/services than others. It will be helpful, where you are considering a response with reference to a particular service or project, to identify it as relevant in your response. You may wish also to comment on interdependencies and synergies among the service and project components.

SUBMISSION FROM THE NATIONAL COMMITTEE FOR DATA IN SCIENCE (NCDIS)

APPROPRIATENESS

1. Was the investment and support for the project(s) and/or service(s) from the Australian Government a response to market failure? More specifically
    a. If the project(s) and/or service(s) had not been supported by Australian Government investment would it have been developed in and by the relevant community(ies)?
    b. If not why not?
    c. If so with what possible differences?

A small proportion of the projects and services generated may have been developed using ARC LIEF funding, NHMRC funding or institutional funding but, to a much lesser extent than NCRIS has enabled, and with more limited availability. The wealthier research-intensive universities (e.g., Go8) are more likely to have had the funds to build/acquire their own e-Research infrastructure. The majority of the NCRIS-funded eResearch infrastructure is unlikely to have been developed or made broadly available, for a variety of reasons including:

1) A lack of funding /resources. Generally ARC and NHMRC funds can't be used to fund software development and researchers are reluctant to allocate any funding to support data curation/publishing;
2) Lack of motivation by communities to support/prioritize some services (e.g., data/metadata capture and sharing, online analytical/modelling services) or to collaborate on the development of shared infrastructure;
3) Conflicting agendas between research institutions (universities, CSIRO) , state agencies, state governments, national agencies, discipline-specific centres/networks;
4) Limited awareness of the benefits or necessity of core infrastructure (e.g., single sign-on, metadata standards, data identification and citation services) that are needed across the board.
5) Many of the services developed (e.g. ANDS and NeCTAR) were not available via the commercial sector – and if developed via commercial entities, they would have been much more expensive.
6) Currently data intensive science is prohibitively expensive on commercial clouds due to online costs of data access.

Moreover, development would have been very patchy – it would not have been at the scale possible via NCRIS. Further, it is unlikely that the government-funded research agencies would have participated to the extent that they did (e.g. the contribution of over 8PB of data to the NCI RDSI data store from BoM, GA and CSIRO). Universities, state governments and research agencies would not have contributed the extent of co-investment that they did.

2. In your view, did the investment change or transform the conduct of research?
    a. Specifically, to what extent was greater research collaboration enabled/fostered?
    b. Were other kinds of collaboration enabled?
    c. Where you believe there has been a transformative impact on research, can that be readily demonstrated, and how?

Yes – the NCRIS investment in eResearch has transformed the way that research is conducted in Australia, but not in all disciplines and to a varying extent across different disciplines.

The disciplines in which there has been a transformative impact include: astronomy, climate science, earth system science, oceanography, ecological science, environmental sciences, bio-informatics/genomics, geo-sciences (geochemistry and geochronology).

This is because the investment to date in national eResearch facilities/services has tended to favour those communities that are large, well-organized, have strategic leadership and a common goal/problem, generate "big data" automatically via shared instrumentation (satellites, synchrotron, sensor networks), have a history of using computational modelling as a research tool, and adopt the same community-wide services/workflows for processing data.

In fields such as earth sciences, climate sciences and geophysics (earth observation data), the NCRIS-funded eResearch infrastructure has also led to greater collaboration between the research and government agencies. Collaboratively-developed national data collections are now accessible in machine readable formats (e.g., the Australian Geoscience Data Cube, ALA) and can be processed and analysed at scales and resolutions never before considered possible. NeCTAR Virtual laboratories have facilitated access to both open source software services (analytical and modelling services) and licensed software services that have lowered both the skills and cost barriers and have led to considerable uptake and expertise building.

For the earth, climate, environmental, space sciences, the NCRIS-funded services have enabled a more rapid transition to data intensive, quantitative science. Additional CPU's and storage have enabled multiple scenarios to be modelled at higher spatio-temporal resolution, relevant uncertainties to be quantified and above all, allowed provenance capture and reproducibility and transparency of science.

One key beneficiary has been the Australian Government itself in that the data that has been unlocked via e-Research infrastructure is now the basis for evidence-based policy advice on issues of national priority (e.g., Water Resource Management, Reef Plan, AURIN).

Disciplines in which further effort is required to encourage the adoption of shared e-Research infrastructure and an open data culture - to transform research conduct include: humanities and social sciences; health/medical sciences; materials sciences. Issues that have contributed to a delay in the deployment/adoption in these disciplines include:

- Lower levels of IT/eResearch knowledge, skills or confidence (humanities/social sciences);
- Extreme diversity of research problems, methods and data types within these disciplines (particularly in the humanities, materials science) – the wide range of sub-disciplines and smaller communities - act as barriers to standardization of data/provenance capture, e-Research methods or sharing/re-use of research outputs;

- History of commercialization of databases in some areas (e.g., materials databases)
- Confidentiality of data (medical/health sciences)

3. In your view do the results to date satisfy the following kinds of expectation about this investment?
    a. Yours.
    b. Your Stakeholders: ie if the main expectations were by a stakeholder group on whose behalf you implemented a service and/or program how do you understand these expectations to have been met (or otherwise).
    c. The funding parties' expectations as you understand them.

Generally, the results to date have satisfied (or even exceeded) expectations in terms of the delivery of services.

The areas in which results have fallen below expectations (of all stakeholders) (primarily because the problem is more complex than originally anticipated) include:

- the numbers of real users of many services;
- ease of accessing and seamlessly integrating different eResearch facilities – from an end-user's perspective and from the perspective of the long-tail research community;
- guarantee of future availability, sustainability (ongoing maintenance and support) of the service;
- robustness/performance of the service;
- ability of services to capture provenance at the detail required to enable re-use, repetition, verification and validation;
- many researchers/individual scientists are still unwilling to share data and metadata because of
    - competitive nature of research and the advantages of exclusive access to data
    - effort required to perform data curation
- problems with AAF authentication that have inhibited wide-spread utilisation (especially by non-university users in the Virtual Laboratories).

CONTRIBUTION TO TECHNOLOGY PLATFORM

4. Can the project(s) and/or service(s) be seen to have contributed to a platform of ICT support for the research sector
    a. as a whole?
    b. in parts or selected disciplines?

In some disciplines, the projects/services have contributed to a comprehensive ICT platform. For example, at NCI the combination of: petascale computing; fast access to spinning disk provided through RDSI; data discovery through ANDS; and support of virtual laboratories from NeCTAR (Climate and Weather, Virtual Geophysics Laboratory) have established the foundations for an integrated platform to perform data intensive science at scales never before possible in Australia, if not globally.

However, the investment to date in national eResearch facilities/services has tended to favour certain disciplines (astronomical and space sciences, geosciences/geophysics, climate sciences, biological sciences (genomics), protein crystallography, ecological sciences, marine sciences). It has mostly supported those communities who are large, well-organized, have a common goal/problem,

generate “big data” automatically via shared instrumentation (satellites, telescopes, synchrotron, sensor networks), use computational modelling as a research tool, and employ similar community-wide services/workflows for processing data.

Research sectors in which there is significant potential for expanding the uptake and development of eResearch infrastructure include: humanities and social sciences; economics; engineering; built environment and design; materials science. These comprise communities with multiple, smaller, sub-disciplinary groups and sub-disciplinary problems and research methodologies. Generic approaches to ICT services are less likely to be successful and more tailored approaches are required. Some of the NeCTAR eResearch tools provide good, successful examples of such focussed communities, tools/methods and research data formats (e.g., OzTrack, FAIMS, Aust-ESE).

5. Was this contribution one which intersected with, or complemented ICT support funded by state-based programs and/or your and other research institutions?

The majority of projects/services have involved a collaboration between national and state e-Research programs as well as multiple research institutions. For example, the NCI is now a partnership between GA, CSIRO, NCI and ANU.

One of the major benefits of the NCRIS-funded eResearch program has been its ability to motivate previously disparate organizations and communities to form collaborative partnerships and to co-invest in staff/facilities.

However, a major challenge has been satisfying the different/conflicting funding requirements of partners e.g., State Governments (who want industry/commercial outcomes), research institutions (who want Science/Nature papers and further grants) and the Australian Government who wants efficient, cost-effective, widely adopted sustainable research infrastructure.

6. In future, will you be planning or seeking to leverage the national eResearch investments to date by co-investing your own organisation’s planned investment in infrastructure and/or services?

Without doubt, many organizations/institutions already funded by NCRIS to develop/host eResearch facilities, will use this past investment in them as a basis for applying for future investment. If there is a further round of NCRIS-type funding, the agency responsible for administering the funding and selecting host organizations will need to assess the performance of the existing host organizations *particularly from end users’ perspectives*, before making decisions about future funding. If there is no further NCRIS funding, it is likely that some institutional *caretaker* funding may be allocated for basic support and maintenance of critical services but expansion of existing facilities will have to depend on future ARC LIEF or NHMRC grants.

GOVERNANCE

7. Do you consider that the governance arrangements for implementation of the investment(s) with which you are familiar were appropriate? If not, how could they have been improved?

There should be more women and mid-career researchers on the governance committees/boards of the eResearch Projects/Facilities.

Members of the boards/chairpersons of these projects are often on multiple committees/boards – of institutional, national, state and discipline-specific e-Research organizations and facilities.

Governance arrangements should seek to minimize potential for conflicts of interest by reducing cross-representation and increasing the diversity and equity when selecting members of boards and committees.

TAKE-UP

8. How well has the infrastructure enabled by this investment (or these investments) been taken up?
    a. Could this be improved?
    b. What barriers have acted to inhibit uptake?

In some disciplines, it has been very widely adopted. In others, significant further effort/investment is required to increase uptake by:

- Engaging with the non-adopters to identify why the infrastructure is not meeting their requirements;
- Making the available services easier to discover, access and use, by non-IT experts.
- Improving the feedback, communication and trust between end user communities and infrastructure service developers/providers e.g., employing "community liaison officers" who mediate between the end users and the service providers and who provide user community feedback to the service providers.
- Employing user interface designers skilled at designing intuitive front-end of systems.
- Streamlining the interfaces between different services (computational, storage, data curation).

Another barrier to uptake is skills shortage, in particular a shortage of data scientists and e-Research software engineers with multi-disciplinary skills (e.g., biology, software, statistics, data curation/databases, information systems). It would be good if there was government support for undergraduate/post-graduate training programs to vastly improve the skills base available. Courses in data science are now becoming available in US (Master of Information and Data Science, Berkeley[1]) but are virtually non-existent in Australia.

LESSONS LEARNED

9. From your experience and understanding to date, what lessons have been learned about the optimum delivery of this kind of infrastructure?
    a. Do you think any of the issues in delivering infrastructure are systemic in nature? Please describe.
    b. If these issues are not readily described as systemic, how are they better described?

Many of these facilities have been developed independently in silos and there has been little investment or effort in coordinating or integrating these facilities or ensuring communication between facilities to reduce duplication and maximize interoperability e.g.,

- ANDS Metadata stores and RDSI storage;
- AARnet CloudStor and NeCTAR Research Cloud Nodes.

[1] http://datascience.berkeley.edu/

There needs to be better overall synchronization and coordination of the eResearch facilities to: reduce duplication; to improve integration and interoperability between facilities; to reduce apparent competitiveness between facilities; to improve outreach to research communities via a common marketing/information strategy across all facilities - so it is clear who is responsible for which particular services and who researchers should contact with regard to specific services.

One suggestion is to re-establish an entity similar to the Australian eResearch Infrastructure Council which was disbanded two years ago. AeRIC was active in seeking to ensure integration across all of the individual programs and to minimize duplication. AeRIC also helped raise awareness of the activities between the individual programs/services.

In the past, DIISR also facilitated meetings and cross-collaboration between the various capabilities (e.g., AuScope Grid, EMII, TERN) – ensuring harmonisation and sharing of developments. DIISR's past involvement in the eResearch Australasia Conference also facilitated outreach/awareness-raising and coordination of the various NCRIS/SuperScience activities.

COMPARABLE DEVELOPMENTS/COMPARISONS

10. Considering the approach taken by Australia in undertaking these investments how would you rate that approach by comparison with
    a. International developments in comparable contexts?
    b. Infrastructure development in other "industries" or sectors?

In many aspects, Australian infrastructure development is more advanced and better funded than many other countries.

However, one of the biggest problems in the eResearch domain is the lack of sustained funding to support data, models and services that have been developed using short-term (e.g., 3-5 year) project-based funding. New business models need to be identified to enable long term support and sustainability of such data and services.

One approach that has been employed internationally with some success and that has been partially implemented within Australia, involves establishing a number of National Data & Computational Centres aligned with National Institutions that are funded long-term. The role of such centres is to provide accredited data repositories and data/computational services associated with a particular high level national research priority:

- Bureau of Met – National Climate and Water/Environmental Data Centre
- Australian Antarctic Data Centre – Antarctic-related data sets
- Bureau of Statistics – Census data plus social science datasets
- National Library – humanities-related data centre that hosts/serves literary data and services, oral histories, creative works, heritage/archeological data and historical data

Where a relevant organization does not exist, it may be necessary to establish one e.g., National Language Archive/Centre. Such national data centres would need to be closely aligned with high performance computational centres but would provide a cost-effective way to host and maintain shared eResearch infrastructure in the long term – particularly for smaller sub-disciplinary communities, regional universities and long-tail researchers.

Australia's eResearch infrastructure could also be better integrated with international networks/international data centres to help solve global challenges. To date, the majority of the funding has been invested in establishing the national facility. Providing funding to engage with similar international efforts and establish links from national to international facilities would be beneficial both in terms of leveraging infrastructure (data and services) and exchanging knowledge.

INFLUENCES AND CONSTRAINTS

11. In considering your responses to the above, would you make any comment about the extent to which the development of the infrastructure as enabled by the relevant investment(s) was affected by
    a. Budgetary constraints, both associated directly with the investment and also in the case of associated issues which were not within the anticipated scope or area of effect of the Australian Government investment(s), eg inadequate complementary investment, or other resource challenges.
    b. Structural issues, including the organisation of research in the sector or in institutions
    c. Technology developments or deficits.

One constraint was the speed at which significant amounts of funding had to be expended. Given that each of the programs had to go through planning, proposal development, submission, evaluation and approval phases, the timelines to develop and deliver services were very tight. This situation was exacerbated by the time it then took to resolve contractual arrangements. In some projects, the time between notification of the funding and actual start date for implementation/employment was over 12 months. Such delays make it very difficult to attract and retain the best staff, particularly when projects are only funded for 12-18 months (e.g., NeCTAR).

Another constraint was associated with State-government co-investment. In some State-based programs and State-based agencies, the State-government funding/co-investment had to be expended specifically on hardware or on sub-projects associated with State economic benefits or industrial/commercial outcomes. Such constraints do not always align with the target performance indicators associated with co-invested NCRIS or University funding. Moreover the State funding constraints and State agency agendas may lead to decisions being made for political or commercial reasons rather than based on criteria such as technical efficiency, technical performance, researcher track records or the track records/expertise of potential service developers.