

E-research National Perspectives

Adrian Burton, Australian Partnership for Sustainable Repositories

Abstract

"eResearch", where the worlds of academia, computational power and high-speed telecommunications intersect. This is the world of networked information, networked researchers, and networked computing environments, where single researchers and single institutions can longer compete alone. What then are the national and international perspectives for policy and infrastructure to enable and support eResearch? What are the trends and working models elsewhere? What is happening from a national perspective in eResearch?

These are the questions to be addressed by Adrian Burton, the Leader of the Australian Partnership for Sustainable Repositories (APSR). Adrian will look at national and international government policies, initiatives, trends and issues, and discuss them in the context of the current and future work of APSR and the higher education sector more broadly. He will also outline some of the work of the APSR eResearch Technical Consultancy service established during 2005.

This is one of three papers on eResearch which are proposed as a group for a single session of EduCause. The three are (i) Adrian Burton and David Berriman: National perspectives on eResearch – policy, infrastructure, trends, and demonstrators, (ii) Margaret Henty and Danny Kingsley: Readiness and responsibility for managing research data: institutional perspectives and (iii) Anna Shadbolt: Tracing a sustainable path for data intensive research communities: An institutional case study from the University of Melbourne.

What is E-research?

E-research is an loosely-defined concept that adds to the commonly understood word "research" all the connotations (both positive and negative) of the prefix "e-". E-research describes a traditional quest for knowledge and understanding, yet this "research" is somehow transformed by the application of contemporary information and communications technology (ICT). In one sense it is always an "aspirational" term, a vision of a continually receding horizon.

The allure of e-research is not only the possibility of doing research faster, more easily or even more efficiently, but actually new types of research in new fields with new methodologies. Bio-informatics might be an example here, and it is interesting to note

that there is not yet a place for this research field in the Australian Research Classification Codes.

E-research is more than a researcher at a computer or even a researcher at a supercomputer. At the very heart of the eResearch vision is *joining up* the researchers at computers and supercomputers to create collaborative experiments and investigations on an international scale never before possible.

E-research also describes the paradigm change created by the advent of instruments that create vast amounts of data about our environment. At one end of this spectrum are a relatively small number of peak national and international scientific instruments such as particle colliders, high definition satellites, and telescopes that create super-human volumes of data. At the other end of this spectrum is a huge number of small sensors and personal devices such as digital sound and video recorders used by researchers. To this we add the sea of data being collected outside the research sector such as census data or traffic cameras used by researchers. Consequently data, the raw material of research, is being produced at unheard of rates. By some projections, in the year 2010 we will be producing in a single year more data than the total ever produced since the beginning of history.¹

Fortunately, to process this data, ICT is also delivering ever more sophisticated software applications, more powerful processor computational power, greater data storage capacities, and networks with greater capacity. These engineering tools enable the researchers to manipulate this tidal wave of data into more orderly networked data sets and online collections of the inputs and outputs of research.

However, there is another dimension to e-research which requires this raw data to be identified, described, structured and given attributes. This is the “value-add” to raw data that transforms it into information. And if this is done according to standards, it paves the way to grids of information, inter-disciplinary cross-pollination, visualization, analysis, data fusion, data mining and SDA (search discovery and access). These are the current goals of e-research.

A national and institutional infrastructure to help achieve these e-research goals must therefore provide or enable:

1. expertise in ICT
2. digital data creation
3. online collections of research inputs and outputs
4. applications and tools
5. data storage
6. high speed networks
7. high performance computing
8. grid computing
9. enrichment of data
10. enrichment of the data (information) environment

¹ From Data to Wisdom (reference below) P.23

These categories have been used to plan, fund, and support e-research infrastructure in various countries around the world. Such infrastructure is known by various names, such as cyber-infrastructure, e-infrastructure, or information infrastructure. This paper proposes a broadbrush sketch of national e-research infrastructure in these broad categories with particular reference to some landmark from the USA, the UK and Australia:

- [The National Science Foundation \(NSF\) Strategic Plan](#)²
- [NSF's Cyberinfrastructure Vision For 21st Century Discovery](#)³
- [Long -Lived Digital Data Collections: Enabling Research and Education in the 21st Century](#) (National Science Board)⁴
- ["Our Cultural Commonwealth"](#) The report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences.⁵
- ["Developing the UK's e-infrastructure for Science and Innovation"](#)⁶
- ["An E-research Strategic Framework"](#)- the Interim Report of the E-Research Coordination Committee.⁷
- ["From Data to Wisdom"](#) Prime Minister's Science Engineering and Innovation Council (PMSEIC) Data for Science Working Group Report.⁸
- The National Collaborative Research Infrastructure Strategy (NCRIS) ["Strategic Road Map"](#)⁹

Some elements of e-research infrastructure are well-established, namely high speed networks, data storage, and high performance computing. Priority for, investment in, and coordination of these three elements (grid) should (and almost certainly will) continue in Australia. This paper does not focus on these well established elements, but rather on some of the more uncharted waters outlined in these reports.

Expertise in ICT

All these major reports stress the need to develop a skilled workforce. Two types are generally identified:

1. ICT and research data management skills for researchers in all disciplines
2. encouraging the tertiary sector to produce engineering and IT graduates with specialist skills in data management, information science, and e-research technologies.

² <http://nsf.gov/pubs/2006/nsf0648/nsf0648.jsp>

³ <http://www.nsf.gov/od/oci/CI-v40.pdf>

⁴ <http://www.nsf.gov/pubs/2005/nsb0540/start.jsp>

⁵ <http://www.acls.org/cyberinfrastructure/OurCulturalCommonwealth.pdf>

⁶ <http://www.nesc.ac.uk/documents/OSI/>

⁷ http://www.dest.gov.au/sectors/research_sector/policies_issues_reviews/key_issues/e_research_consult/interim_report.htm

⁸ http://www.dest.gov.au/sectors/science_innovation/publications_resources/profiles/Presentation_Data_for_Science.htm

⁹ <http://www.ncris.dest.gov.au/NR/rdonlyres/91C5DFB3-10E5-4A09-A861-6973B2912417/9519/NCRISStrategicRoadmap.pdf>

For example the eResearch Coordinating Committee in Australia has called for “an investment in human capital”. The PMSEIC Data for Science Working Group report also made core recommendations about skills for data management:

That data management expertise becomes a core skill for researchers, including graduate and postgraduate science students across all disciplines, and that they receive data management training as part of their education

The NSF’s report “Long Lived Data Collections” underlines the need for more specialist graduates:

The Foundation, working with collection managers and the community at large, should act to develop and mature the career path for data scientists and to ensure that the research enterprise includes a sufficient number of high-quality data scientists.¹⁰

The NSF is following up these planning priorities with funding priorities. For example the US\$50m Cyber Enabled Discovery and Innovation program has as one of its priorities “educating researchers and students in computational discovery”.

Implementing this in the Australian context would seem to require some coordination. The very nature of NCRIS has resulted until now in a greater focus on infrastructure than human expertise. Expertise development within the research community might be coordinated by the NCRIS PfC, which included “technical expertise” as a priority in the NCRIS roadmap.

Digital Data Creation

The underlying infrastructure for digital data creation in the scientific domains is scientific instruments or sensor networks that create research data. The NSF’s strategic plan includes the intention to identify and support “the next generation of major equipment and facilities to enable transformational research”¹¹. Australia’s NCRIS roadmap focuses on identifying and funding medium to large scale instruments and facilities for selected research fields.

The UK’s e-infrastructure report underlines the importance of good quality data. The report indicates that the process of creating data is key to getting good quality data with appropriate metadata. It stresses therefore the importance of automated metadata creation and software to enhance the quality of the data creation process.

The UK’s report includes digitization on its digital data creation agenda. It calls for “a strategic approach to digitization and repurposing of data as a means of enabling access and new forms of analysis”. The “Cultural Commonwealth” report of the American Council of Learned Societies describes the importance of supporting and coordinating the

¹⁰ Op cit p. 48.

¹¹ op cit. p.9

digitization of legacy materials particularly in the humanities disciplines.¹² Australia's e-research agendas do not include digitization.

NCRIS investments in instruments and research facilities include a responsibility to create the highest quality data possible, directly from the instruments.

Online Collections of Research Inputs and Outputs

Online research data and collections are the basic building blocks of a new research information infrastructure for all disciplines. For example the report of the American Council of Learned Societies concludes that "extensive and reusable digital collections are at the core of the humanities and social science cyberinfrastructure"¹³.

Digital data collections have also been identified by the NSF as fundamental to a new scientific research paradigm:

It is exceedingly rare that fundamentally new approaches to research and education arise. Information technology has ushered in such a fundamental change. Digital data collections are at the heart of this change. They enable analysis at unprecedented levels of accuracy and sophistication and provide novel insights through innovative information integration. Through their very size and complexity, such digital collections provide new phenomena for study. At the same time, such collections are a powerful force for inclusion, removing barriers to participation at all ages and levels of education.¹⁴

The Systemic Information Infrastructure investments by DEST in Australia have seeded some discipline specific repository networks in medicine and marine science, as well as some foundation work in institutional repositories, which will be followed up this year by the Australian Scheme for Higher Education Repositories (ASHER) scheme. It also funded the development of the APAC Data program for large scientific data sets.

There is still more work to be done in supporting national reference collections and important community collections. The only program currently funded to address some of these issues in Australia is the NCRIS PfC.

Enriching data and the data environment

Enriching the data requires the application of structure, description, and attributes to the "raw" data and, if these are done according to commonly agreed standards, the whole data environment can be enriched by any number of common services such as SDA services (search discovery access), analysis, visualization, fusion, data submission and presentation services etc.

¹² Op cit p.5

¹³ Op cit p.38

¹⁴ Op cit p.9

As an example, the Report of the Working Group on Search and Navigation for the E-Infrastructure Strategy underlines the importance of metadata as one of its “central issues” for SDA:

Metadata is structured information that describes the key characteristics of information objects. Resource discovery is critically dependent upon metadata, which itself is essential for effective machine-to-machine interaction. In order for search and navigation to move forward it is necessary to encourage information providers to adhere to metadata standards, for new, richer metadata standards to evolve, and crucially for all information providers to expose their metadata effectively so that information can be indexed and searched.¹⁵

An important role for national eResearch infrastructure is the identification, adaption, and promulgation of these metadata standards for describing collection and data-sets as well as standards for exposing this descriptive metadata.

Then community, national, and international SDA services can be built as part of an enriched data environment.

The above is an example of enriching data and the data environment for SDA, but a similar process applies for many other types of value-added functions, visualization, data fusion, data mining. The first step is describing, structuring and giving attributes to research data, then provide common or aggregation services to create a much more sophisticated information environment for research.

Digital Sustainability

The UK’s e-infrastructure strategy for Research allocated a whole working group report to preservation and curation.¹⁶ This e-infrastructure report includes digital preservation as one of the central issues for e-research infrastructure. These reports identify digital sustainability as one of the major risk factors for the longevity of e-research infrastructure. The JISC funding plans to 2009 include the Digital Curation Centre.

The NSF have also issued a report on the curation of scientific data¹⁷, and have followed this up with a funding call for projects that bring together diverse skills in the fields of science, information science, library and archiving.

There have been some programs that support digital sustainability in the Australia through the SII, but there seems to be no plans to continue this beyond 2008.

¹⁵ P. 5.

¹⁶ <http://www.nesc.ac.uk/documents/OSI/preservation.pdf>

¹⁷ “To Stand the Test of Time: Long-term Stewardship of Data Sets in Science and Engineering”
<http://www.arl.org/bm~doc/digdatarpt.pdf>

Conclusion

There are opportunities to continue to develop infrastructure for e-research in Australia. These include consolidating existing strengths (HPC, data, networks) and extending our national capability in new areas of e-research infrastructure.