

National Collaborative Research Infrastructure Strategy Capability 5.16 (Platforms for Collaboration)

Summary of Workshop 1, October 23, 2006

Paul Fritze (incl. notes by Stephen Young and Peter Nicholson)

Overview – Mike Sargent & Rhys Francis

The NCRIS goal is to provide world-class facilities to researchers. This is beyond a simple competitive process – it involves seeding cultural change mechanisms, collaboration, strategic approaches and a use of principles, rather than filling in forms.

NCRIS principles seek maximum national benefit, world-class research, non-exclusive access, making isolated data accessible and considering whole of life costs. It's not about cost shifting to NCRIS.

“Data is the infrastructure” on which research is built - it's about networks, data, collaborative workspaces, and sometimes intensive computation. We need to tackle generic solutions and cultural issues.

The nine NCRIS capability areas identified for investment have addressed the long haul view, equitable access regimes, good governance, an investment plan – a ‘living strategy’ under continuous review rather than being driven by funding rounds.

Most research communities are starting to engage, but some have not understood the system. NCRIS must avoid allowing an inner clique to develop.

Most initiatives are ‘networks of capabilities’ greater than any single entity. They involve the challenges of distributed generation, ownership, storage, accessibility control, audit trails, ability to annotate, etc.

NCRIS Platforms for Collaboration is a separate capability focused on a collaborative framework servicing the other capabilities. It's about data management, analysis, grid enabling technologies, technical and HPC expertise and high capacity networks.

Current PfC work in progress¹ includes Authentication and Authorisation workshops, APAC review, exploration of the concept of an eResearch Toolkit, Data Management Survey.

Background to problem

A tidal wave of data and global sharing is changing research process and culture. The scale and complexity of this is ‘novel’ and we cannot rely on old models². The amount spent on data is growing as a percentage of research.

Categories of data are key, e.g. reference datasets, institutional curated datasets, private datasets, federations of data and metadata. The CAUDIT survey indicates there is not a lot of centralised data. Other references include PMSEIC data community³, NH&MRC code for research⁴.

The question is what to do, how to agree and who should pay - what should NCRIS do?

PfC Data Management survey (Paul Fritze)

The picture emerging from surveyed initiatives⁵ portrays a common data lifecycle model in which data is transitioned between states: source, treated, analysed, preserved, published and end use.

¹ NCRIS PfC <http://pfc.org.au>

² Microsoft 2020 Science <http://research.microsoft.com/towards2020science/>

³ AAS submission to PMSEIC <http://www.science.org.au/reports/data-for-science-11august06.pdf>

⁴ Australian code for the responsible conduct of research
<http://www.nhmrc.gov.au/funding/policy/code.htm>

⁵ Analysis of PfC DM survey <http://pfc.org.au:80/cgi-bin/twiki/view/Main/DataWorkshop1>

Initiatives were clustered under different cultural perspectives: the research institution, academic scientific/medical community, academic Social Science/cultural community, Government organisation, professional support service and 'commercial' research organization⁶. The different perspectives were characterised by policy statements and standards applying to different phases of the DM cycle. These can be seen to operate at multiple levels, e.g. group, institutional, national, or international. Research culture is an additional less formal level of influence, e.g. sometimes described as 'ad hoc' management practices.

Individual engagement in data management strategies will reflect the perceptions of each researcher of their current institutional/research environment, as shaped (only in part) by service providers and policy makers. Processes for monitoring research community perspectives will strengthen the alignment of policy and support with researcher needs.

Short presentations

SIRCA - Michael Briers

SIRCA deals in financial markets and news data sourced from a range of suppliers. Content data is owned by providers with SIRCA being the custodian. SIRCA also generates and owns metadata about this content. End users are researchers.

SIRCA has made a sustained investment in data management capability through successive grants but is moving to wean itself off the 'boom and bust' cycle of public funding and is adopting user pays principles. Sustainability includes maintaining and finding work for an internal IT department.

Current moves are to grid computing, improved networks, more real time processing, development of Web service tools, visualisation and annotation.

SIRCA maintains close contact with end users, providing 24/7 support and ontological views to facilitate end use. An expansion to international involvement is focusing effort on ISO standards, quality control processes and resilient systems that are required for broader exposure.

MMIM Project - Michael Georgeff

The need here is to combine patient data held in multiple institutions and agencies, e.g. hospitals, PBS, Birth & Deaths. Data is heterogenous, non-compliant with no common standards or terminologies. Significant privacy and ownership constraints cause problems for sharing. The requirement to work with de-identified data complicates record linking. It is hard to get researchers to agree.

MMIM is a federated system providing discovery metadata and query services that provide links to Local Resource Repositories in institutions. Data quality and preservation is managed at institutional level. There is a requirement for rich discovery/query services, e.g. data mining, analytic tools and Web portal applications for users once data is found. Metadata exists at both analytical (LRR) and discovery levels.

Attempting to force standards on institutions would not work, so the focus has been on simple linking. The successful approach has been to let standards emerge and incrementally implement them. MMIM started with a pilot in 4-5 hospitals in 2005; it's now 10-15 and moving to the Australia-wide Cancer Grid.

HE Physics – Glen Moloney

In the High Energy Physics world you have join with the international community although competition still exists between institutions. Australia is part of two large High Energy Physics collaborations: Belle and ATLAS.

Belle is an old style experiment involving 300 collaborators, 50 institutions in 50 countries, with data and computing located in the Japanese facility. APAC NSF is providing an SRB infrastructure giving Australian researchers access to the data and ability to contribute analysed products back. It's trying to bring a more federated approach to a running experiment.

ATLAS is a next generation experiment that will start a massive data deluge in 2007. Australia is a node in a global grid involving 192 institutions and 200k CPUs. Australian partners are bound by the requirements

⁶ SIRCA is actually a non-commercial funded agency that is working towards a more self-funding model.

of the international collaboration, i.e. you don't have a lot of say. A lot of support has come from AARNET, APAC, VPAC, etc.

Working groups

Biomedical group

Biological data are semantically very rich. Ontologies are different in each part of biology, don't overlap and are expanding. Different people have different definitions of 'data', e.g. as publications, diagnoses, archived datasets for further analysis, metadata, internally generated or harvested from other providers, subject matter, unit data, aggregated data and reports.

All forms of data have related issues of IP, copyright, privacy, access control, discovery, authentication, curation, quality control and in particular, preservation, e.g. how replaceable is data, why should it be kept? The real issue, however, may be one level down – how do you actually do things in hospitals, with publications, instrumentation.

ARC funding policy could have say in research quality, publishing, citing, contribution, finding.

- ***NCRIS could provide some kind of definition of data in “matrix form”.***

Publications data

To NH&MRC, data is about research papers, getting results out and effective end use, e.g. Cochran collaboration as critical assessment of repositories of papers.

Funders have a role in policy for open access to journal articles (NH&MRC are about to put out a statement to “encourage” open access).

- ***NCRIS could lead establishment of ownership principles for publicly funded data.***

Health data

Health related data has issues of privacy, linkages, unique identifiers, consent and ethics. Data is hard to understand and terminology is a problem, e.g. what is meant by 'non-smoker?' Identity management is important – it's about individuals and data. Different levels of access are required e.g. for research, clinics or teaching. Data is highly heterogenous, widely distributed and “has to stay in hospitals”. This all suggests a federated system where data linkage and access are the core problems.

- ***It's up to researchers to define, identity, authorise and manage access, but NCRIS could provide a framework and technology to help structure and support discovery and re-use. A service that links via metadata across datasets – beyond what a single entity can do.***

There is a problem of mutual recognition of ethics committees across institutions and States (NH&MRC are considering some kind of “harmonisation” across institutions)

The health system has “enormously bureaucratic distributed management”. A key issue is involvement of hospitals – you just can't get the people together in the health system the way you can in the universities. To engage Health agencies, you need to allow diversity and gradually get them to come together. A purely centralised model won't work. Researcher buy-in is a problem - an issue of confidence in data custodians

- ***NCRIS could provide framework for negotiation between entities or change mechanism that would allow for hierarchy of authority, e.g. show hospitals what authorisation might look like.***

Bottom line end use requirement is to find data, know who to talk to about it and how to understand its meaning – must user friendly to researchers.

- ***NCRIS could provide a National level ‘trusted house’ service to manage linkage, authorisation and metadata management of health dataset in institutions and agencies, such as hospitals. (e.g. providing intelligent gateways between existing State data linkage projects).***

Small scale Biology research data

Biological research involves data from instruments rather than patients. Researchers own the data but do little in the way of data management policy. Data management is not covered by research funding.

Institutions could support individual researchers/small datasets, e.g. Arrow Dart, Melbourne/Monash, Versi – need to investigate these small research approaches and get buy-in by universities.

ABS is developing infrastructure, software tools and services and looking for collaborations (NDN⁷) - one of a number of approaches that need to be mapped.

- ***Need for data management ‘toolkit’ that can be adopted by institutions.***
- ***NCRIS could provide frameworks and technology to facilitate sharing of instrument and biology research data across research collaborations: Middleware and toolsets to facilitate research collaborations seamlessly and securely.***

Bioinformatics instrument research data

Distributed large datasets; volume doubling every 10-15 months, AAA important. There is a need for storage, mirrors, DM people, HP networks and some HP computing

- ***NCRIS could provide a national approach to bioinformatics for data streaming.***

Possible duplication with 5.1, 2, 5?

Quantitative/scientific group

Requirements for TeraByte architectures are an issue with larger datasets demanding better tools & bandwidth, access control, sharing solutions, AAA, metadata framework/policy, attention to longevity. Data classifications include raw, quality, time ordered, credibility, uniqueness (replacability), cost of assembling. ‘Discoverability’ is essential to enable collaborative research.

There are no common standards across science datasets – a need for technical standards, ontologies and principles – particularly in the move to supercomputing. Discipline communities are driven by international standards, but the interface between communities needs coordination.

- ***NCRIS could play role gathering standards for Science data sets.***

Dataset ownership is defined by the collaboration itself through policies and access. For Government agencies, focus is on its client base, e.g. industry, with unclear responsibility for providing access to research community. Need to understand the organisation’s charter. e.g. to put data in public domain, cost recovery models. State-based agencies can be closed networks making external dataset access difficult.

There is a need for change in institutional and researcher culture from competition to working in a more collaborative space. IP issues are becoming difficult. Need buy in by senior staff. PhD training is in everyone’s interests.

Broad-band access is an issue for agencies (SIRCA, BOM), e.g. no access to AARNET. Different charging models occur across research sectors.

- ***There is a need to revisit AARNET’s charging model to see how it can carry industry R&D.***

There is a lack of specialised skills required for scientific data management. Service level may be different for researchers and ordinary users.

Bottom line is a need to identify where best practice is happening, i.e. practice that is consistent with international standards and it ‘works’. It will be driven by the research community, but related to technical service provision, e.g. grid community – particle physics model.

- ***NCRIS could identify large science ‘reference datasets – best practice, international standards.***

Social Science Humanities data

SS/Humanities data involves lots of collections and individuals with no clear decision process - similar to small scale Science research. Data is about events and people, analysing and commenting by evolving group of researchers. It is complex and involves a lot of non-electronic data [for NCRIS a digitisation proposal is ‘out of scope’].

⁷ National Data Network <http://www.nationaldatanetwork.org>

Problems reported include access, preservation, cultural change, access to funding, perceptions of benefit by researchers. Ownership is a major problem. Creators are the data managers who decide what's important to keep, but have no expertise to do management in a federated environment. While university researchers own data, in Government organizations the Commonwealth does. Data ownership may change over the life cycle, e.g. the institution takes over ownership when researchers move on.

Requirements include cohesion across datasets, data findable by other interests, perpetually accessible, AAA, 'incentives', identity management. Generic preservation services exist but we probably don't know what they are. Priorities are preservation, support services, cultural change and toolkits.

- *NCRIS could provide mapping of datasets in the community and mechanisms for identity management.*
- *NCRIS could provide a stewardship framework. Some type of federated access and standards for collection – a 'preservation infrastructure'. E.g. a national registry for catalogues, a national body for expertise in data management.*

Require support services of some sort, for academic users and collection managers. e.g. 'toolkit' and standards. Expertise in social Science and data management is required, probably a distributed model.

Concluding thoughts - Rhys

Areas of interest for NCRIS:

There is a need to provide a 'home' for small datasets whose sum is worth more than parts - a federated preservation environment that may involve collection agencies. The desktop level is more an institutional responsibility, e.g. for backup infrastructure, but service providers may play role getting data into more accessible form, e.g. Water Resource Network⁸. NCRIS may facilitate extension of existing storage service providers to others.

We need a way to identify (large) primary datasets to be sustained, but there is no agreed process for deciding National value. Such judgments are outside NCRIS, but some sort of framework and process of prioritising is required.

Grid approaches (automatic, real time access, fine grained data) require National AAA infrastructure that is probably five years off. This is beyond institutions and relevant to NCRIS. Providing a framework for public data access is more straight forward.

Under consideration:

- *Health data linkage service.*
- *National approach to Bioinformatics data management and services.*
- *AAA and related trust federations currently being developed.*
- *Grid e-Research toolkits currently being explored.*
- *National models for annotation, preservation, services and expertise.*
- *Models of data identifiers, objects in repositories, e.g. in design of toolkit. NCRIS is only interested in the fabric – not how to do in each discipline.*

Preservation infrastructure services are vital and complex, organisations have a role to fund, with NCRIS as backup for others.

Next workshop:

To be smaller and more focused, looking particularly at what we don't know and services above the institution.

⁸ Water Resource Network <http://water.nml.uib.no/about/>