Intersect Australia Ltd
PO Box H58, Australia Square
Sydney NSW 2000

T +61 2 8079 2500
F +61 2 9262 3040
www.intersect.org.au

ABN 67 131 752 657

National Collaborative Research Data Infrastructure and tools

NSW considerations on state/national data storage infrastructure.

30 October 2009

Version 2: "an initial model"

markus.buchhorn@intersect.org.au

INTERSECT

Table of Contents

This document is very much a draft outline, based on feedback to an initial set of questions, and will need to be further evolved and informed. It is also set within the context of a national planning framework that is itself evolving. This version of the paper provides an initial summary of views and seeks further feedback from the NSW research community.

1. Executive Summary

The need to effectively manage research data, during its collection, analysis, publication and eventual re-use is becoming increasingly evident to the sector. The federal and state governments are offering funding to assist organisations, institutions and other agencies to tackle the issues. In NSW many groups have agreed that a strategic and collaborative approach will provide the greatest benefit to its participants. The federal funding through the SuperScience initiative of around $145m provides a catalyst to develop a plan for the next 5+ years and into the future. This paper summarises the issues, requirements and opportunities, and proposes a three-pronged approach going forward to reach a well-considered, appropriately-governed, and user-informed framework running on a trustworthy and robust infrastructure. It will be carefully connected into any federal and state planning processes, especially the EIF/SuperScience process currently emerging.

2. Introduction

Research data is becoming an ever more valuable asset for researchers in Australia, and the volume of data being collected is increasingly dramatically every year. The way it is captured, managed, stored, accessed, presented and preserved is becoming an issue for researchers, their institutions and agencies, and governments at all levels.

There are many ongoing organisational, regional, and state-wide aspects that will benefit from shared knowledge and a strategic approach to the management of research data. When dealing with the longer term there is a need to consider the ongoing responsibilities and costs which implies the support of enduring organisations.

This document sets out a range of issues, requirements and opportunities related to the management of primary research data across NSW research institutions and agencies. It has been informed by the research community and by the management of the organisations in which they work; the universities, research centres, government agencies and other bodies.

The major trigger for this document is the significant recent federal funding for research data infrastructure and services, and ensuring it is effectively allocated, within the context of a much longer-term strategy. This paper provides advice to the federal initiatives on behalf of the NSW research community, as well as articulating a process that NSW research organisations will immediately undertake to develop a state plan for the management and appropriate sharing and re-use of research data. This process will be closely connected with federal processes as appropriate, and managed under a formal project-management and governance framework.

It is important for the acceptance of any large-scale infrastructure investment strategy in this space to gain the confidence of the research community. It must lead to services that are seen as trustworthy, reliable, robust and long-term stable. These depend on a solid foundation, of people with

skills and well-designed and -operated infrastructure, which themselves depend on a proper governance, planning and budgeting framework.

3. Background/Context

The federal government has invested significantly in infrastructure and services to support research data management and accessibility. The major channel recently has been the NCRIS Platforms for Collaboration capability, which manages the ARCS and ANDS programs, which between them have been funded with around $190m. This provides the single largest investment in research data infrastructure in the world.

The design process outlined in this paper should build upon the work carried out by ANDS for the Australian Research Data Commons (ARDC – *ref. TADC paper*). It should also be informed by, and provide information to, the ARCS Technical Working Group on research data services and collaboration tools.

Major funding opportunities for infrastructure are listed below; all have particular purposes and constraints that need to be balanced within the overall plan:

- NCRIS, both within individual capabilities that have prioritised certain discipline areas, as well as PfC which manages ARCS and ANDS on behalf of the entire research sector. NCRIS can fund capital as well as operations, and will finish during 2011/2012.
- EIF, through the 2009 federal budget SuperScience initiative, which has provided funding to ARCS and ANDS, but will only fund infrastructure and its development but not its operation. The EIF funding from 2009 will cease in 2011 for ANDS and 2013 for ARCS.
- The federal funding programs under the ARC and NH&MRC provide funding mostly for infrastructure and development but rarely for any level of operations.
- Institutions and research agencies have recurrent budgets which could be applied to capital and operational costs if they see it as a strategic priority, but of course takes away from other purposes.
- In NSW, the State Government provides the State Leveraging Fund, which is targeted at attracting other funding to support NSW research initiatives.
- Commercial organisations may see benefit in providing in-kind support for the storage and management of research data.

While the various sources have particular constraints and priorities on what their funds can be used for, there are also wider contexts and imperatives which will inform infrastructure planning. These include:

- The Australian Code for the Responsible Conduct of Research (ACRCR) sets out broad requirements for the longer-term access to research data.

- The ARC and NH&MRC are starting to adopt the emerging approaches from the US and UK that research data collected through public funding must be made (appropriately) publically accessible.
- Some international collaboration projects with an Australian presence are supported by agencies in the US, the UK and elsewhere, with data-sharing guidelines or policies.
- Some projects are constrained by ethics and privacy frameworks, which can be at odds with the rules of the funding sources.
- Some journals require "open-access" to the research data that is evidence for a publication.
- Some datasets are acquired from primary data providers under commercial relationships and may have additional access or curation requirements.

4. Research Data Storage/Management Purposes

4.1. Purpose principles

When considering the development of a significant research data infrastructure, providing a rich set of services, it becomes easy to attempt 'everything'. Even with significant funding available, it will not fund 'everything', nor will it fund it in perpetuity. This pragmatic consideration alone means that plans must be carefully scoped and tied to particular purposes.

There have been major efforts, within Australia, the UK and the US, to map the space of research data management needs and services, and then identify priority areas that can be funded and lead to operational services. ANDS and ARCS have guiding business plans, through both NCRIS and EIF funding, which identify major areas of development.

However these plans do not define boundaries around areas such as policy frameworks, content "ownership" and research priorities. This immediately leads to questions such as whose research data could/should be supported, at what stages in their lifecycle, under what guarantees, and even which organisations should be supported and how. Not all "research data" comes from "research agencies", e.g. government, industry and private datasets that are used to inform and support research. If there are limitations to be imposed these must be defined early, to set out the rules for participants. Some organisations may sit on the boundary of the research space, e.g. those that have collecting imperatives or are targets of donations of historical materials, such as cultural organisations, are often not fully funded to actively acquire, or provide ongoing storage and curation, of data assets but are expected to support research initiatives. This could lead to unintended cost-shifting.

It is widely agreed that all research-supportive data should be supported at some level within a state and national data management framework. Data itself will evolve through a lifecycle of processes, and there may be different stages at which data is stored on the proposed infrastructure, for many different reasons. Stages in data processes include

  - Raw data collection (and storage), preceding any calibration or initial analysis process

- collaborative access to raw or calibrated data during a project,
- wider sharing and re-use beyond a project, becoming 'reference' data
- archival for policy and publication purposes,

with ongoing management of the data and its metadata throughout all of these stages by appropriate authorities. These management processes become highly discipline-specific at some level, and move into a separate governance domain.

All of these processes also need to be linked, and ideally integrated, with the publication process. Publications are data in their own right, and there are increasing imperatives for the evidence underpinning a publication to be made accessible.

4.2. Classes of storage services

The research community has a diverse range of needs for the storage and management of research data. It is useful to broadly categorise the services under particular purposes, or classes.

Storage classes make different kinds of promises to users. For example: "Scratch" or "Drop-box" storage can imply short-term bulk storage for active projects, possibly high-performance for reading/writing but with limited reliability guarantees. "Collaboration" storage can imply medium-term storage for active projects, and likely to have significant access control requirements, but certainly higher levels of trustworthiness. "Archival" storage implies long-term accessibility, hence firm operational support, and may have various levels of performance of access e.g. it may be stored offline. "Mirror" storage implies copies of authoritative information kept elsewhere, typically brought closer for performance issues or institutional requirements.

Any planning process for infrastructure must catalogue the storage classes, their functionality, performance and promises, and needs to be informed by the needs of the research community, and existing practices.

5. Generic and Specialised services

There is a continuum of services that can support data-related activities, ranging from simple concepts such as 'store' and 'access' through to more complex issues such as 'manage', 'curate' and 'discover', up to discipline-specific services related to e.g. 'merging', 'analysis' and business-specific services such as 'preservation' for the ACRCR requirements and 'reporting' (with 'access') for ERA/HERDC requirements. Some of these may be totally common across all disciplines and all organisations, with minor tweaks for context, through to highly-context-specific such as an astronomical cone-search service.

It must be noted that 'generic' does not imply a sense of simplicity or low quality, but the concept of an underpinning infrastructure layer that is common across diverse communities, e.g. across disciplines or across institutions.

Funding sources (e.g. ARC/NHMRC grants, NCRIS) for 'services' may target certain discipline opportunities, but are unlikely to provide ongoing operational funding. Some groups have suggested that institutions may set research priorities and underwrite operational services for research activities

at some level (maybe at the 'generic' level), or may choose to focus on where it adds strategic value ('specialised' level) in the expectation that generic services are funded from broader frameworks.

Other groups have suggested that there is no strong boundary between them, or that services may too easily evolve from specialised to generic, or that the mesh of dependencies between them make the distinction meaningless. However, there appears to merit in the classification when discussing operational funding. Generic services may be described as those that have a well understood cost model for recurrent funding, whereas specialised services have greater elements of uncertainty. Given that discipline needs are set within institutional priorities, it suggests that generic ongoing services should be supported through institutional operational funding, with specific targeted funding for specialist services that are consistent with the University's strategic objectives.

It is clear though that data must be visible to both classes of services, allowing for discoverability and re-use in other contexts.

6. Use cases

To inform the design of the infrastructure, and the classification and prioritisation of services requires gathering the needs of the research community, in universities, government agencies, and other organisations. The deeper that gathering process is carried out, the better-informed the planning will be, and at the same time the better the community engagement will be.

There are various ways such use-cases may be collected. There is broad support for a multi-stage process, with an initial (online) survey to indicate an outline of needs, followed up by more detailed engagement through pro-forma questions, and for larger-scale initiatives a much closer analysis can be carried out. Organisations such as Intersect are well placed, and trusted, to provide support for this process.

Design of the survey and pro-forma need to be done carefully, to capture not just basic needs (such as data volumes, file formats and particular processing tools) but also information related to rights and other terms of use, as well as discipline-related specifics. It should capture a sense of relative priorities, to inform the deployment process. The process should take several perspectives, including that of the researcher in relation to their workflows, of the data in relation to its curation and lifecycle, and of the institutions and other organisations that will be expected to provide ongoing support and have administrative process built on research outputs.

7. Infrastructure and Services Design

There are widely-understood benefits in a coordinated strategy for storage infrastructure, which means there ultimately needs to be a technical framework developed that takes into account every level of service provision from the desktop through the institutional, state-wide and discipline services, and how they connect into national frameworks. It needs to address the higher-level 'services' view and map that onto a physical infrastructure. It also needs to consider the need for support services, which enhance the capability of the research community to make use of the infrastructure.

For the research community, the two crucial elements are that firstly, the services are robust, trustworthy and effective, and secondly that they provide user-friendly and useful interfaces to their applications. The back-end technology is important, but fundamentally most researchers will not particularly care where or how the data is stored as long as the interfaces provide the required services. Regardless of the implementation it should be transparent to users, no matter what model is employed – the back-end must be informed but transparent to a researchers who says 'store this'.

The process to develop the design should be run as a formal project commensurate with the scale of the proposed investments, with a steering group, project manager, and project plan. Intersect could provide the coordination of this project. As a process it can run in parallel to the planning of the governance and management framework. However, in the short-term it needs to account for the evolving context in the ARCS/EIF process.

This design process needs to start quickly and be able to evolve. The consensus is to start with a simple set of services, and to add functionality over time as maturity and trust are established and the operational model is tested. To evolve the framework there must be a process to publish any proposed extensions or changes as early as possible to allow local groups to plan around future developments.

Some infrastructure already exists at institutions and other organisations, and is funded, or is about to be funded. That investment needs to be recognised, not just for financial accounting of 'leverage' but also the institutional strategic priorities implicit in such investment. As part of the technical design, there will be a need to link with institutional infrastructure; some level of abstraction and well-defined interfaces will simplify the communication, and enhance the engagement between the state, national and local infrastructure.

8. Governance and Management

8.1. Oversight and coordination, at the state level

A strategically planned infrastructure with significant federal, state and institutional investment needs a solid, transparent, appropriate and well-understood model for oversight and coordination, with a high level of trust/assurance.

At a minimum, there must be a way for participants to participate actively in the governance of the infrastructure. There needs to be a formal core structure for governance, with advisory groups on policy and technological issues. Some committee-based structure for governance, with representation from participants, is required, however it should avoid creating yet another organisational entity if possible, rather it should build off existing organisations. Such a governance structure must also take into account that there are different classes of stakeholders, such as those related to infrastructure/service providers, or related to particular content providers.

Like the technical infrastructure design, it may be beneficial to start quickly with an interim model, and allow it to evolve, again taking into account the evolving national context. But it must be planned for the longer term, beyond any initial investment for capital, and interim models can become entrenched and hard to change.

One central group, such as Intersect, should be tasked to provide coordination and operational management, under appropriate and representational oversight, and clear effort made to balance needs of all scales across the sector.

8.2. Relationship between national and regional frameworks

There are 'infrastructure' and 'service' providers at multiple levels with Australia, from the institutional to the national. All of them have their own governance and management mechanisms, such as the joint ventures underpinning ARCS and ANDS, and incorporated entities such as Intersect. Any state-wide governance structure has to interface with those frameworks, and has to rely on those other ventures interfacing effectively amongst themselves

To ensure that the relationship between any NSW program and federal programs is optimal and effective, a group such as Intersect is crucial, to provide a single point for the state where the various programs can be combined and aligned.

8.3. Managing Risks

The governance and management framework must deal with risk; an appropriate register of risks at all levels must be established, maintained and monitored, regardless of how the services are ultimately delivered.

9. Planning

Having arrived at an overall architecture, there need to be plans for at least three distinct aspects: the overarching/longer-term strategy, the deployment phase(s), and the ongoing operations and refresh. These will necessarily overlap. They will also need to evolve over time, as maturity and expectations grow.

9.1. Strategic Plan

The Strategic Plan must include the business model and indicative budget, participation expectations, and a broad outline of the services framework over time. It needs to be reviewed annually, and set out at least a five-year timeframe commensurate with the initial NCRIS/EIF funding. It must also identify the broad principles for ongoing support and evolution beyond that timeframe.

To keep the infrastructure relevant means ongoing review of the services framework, of the market offerings, and refresh of the underlying technological platforms. There must be an asset replacement schedule in the budget, as well as maintenance planning. These will ensure that there are sufficient operational funds, underpinning the operational plan, to support the infrastructure and to provide for timely upgrades.

The business model must identify sustainability mechanisms, including subscription models and fee-for-service models. Participation expectations should entail a longer-term commitment, for up to the timeframe of the strategic plan initially (i.e. 5 years).

9.2. Deployment and Commissioning plan

The demand for research data infrastructure is already very high, and the expectations have been raised by programs such as EIF, the ARC2009 NSW grant for HPC and storage, and the ARC2010 proposal for large-scale state-wide storage. There is wide agreement that researchers need operations to start as soon as possible, implying a deployment plan that rolls out basic functionality very quickly. That plan should outline a professional project management approach which includes detailed time lines, milestones, deliverables, regular reports, and audits. This process is likely to be strongly influenced by the processes for the ARCS/EIF-infrastructure plans.

The deployment plan must take into account existing institutional infrastructure, and allow for the deployment of perhaps centrally-managed or -coordinated equipment at institutions.

While the use-case gathering outlined above will provide more detail, some commonly identified priorities include:

- Storage for large data sets,
- More general data storage mechanisms to protect/rescue data from researchers' desktops and make it more easily available as appropriate.
- Archival storage for data that supports publications and needs to be maintained and available for set periods of time.
- Support for multidisciplinary or multi-institutional projects to collaborate around data.

9.3. Operating Plan

Once the infrastructure starts to be deployed, without infinite capacity and resourcing up front, there will be operational pressures, including managing income streams, managing service levels and managing allocation of resources to certain activities. These include staffing for support services. The operating plan must include the budget and resource allocation processes, metrics and auditing mechanisms, implementation and prioritisation processes for the service offerings, and how all of these will be able to evolve within the wider Strategic Plan.

10. Investment approach

As noted earlier, there are a diverse range of funding sources, each with their own funding priorities and particular constraints. The technical design process together with the governance and planning mechanisms will identify the best ways to seek access to those funds and target them effectively and optimally.

There must be clear indications of:

- what level of co-investment (subscriptions, fee-for-service) participating institutions will commit to, and what purposes they may be tied to.
- that the expertise and infrastructure required to comply with various policy frameworks, such as recommended retention periods, has significant cost.
- The timeframes, and timing, for expenditure on infrastructure, service development and operations.

The generally preferred model is one where the initial funding is used to create the infrastructure, but subsequently the services become "just another service", capable of sustaining themselves.

There is clearly a need to refresh the infrastructure over time. A massive up-front deployment may lead to a massive refresh being needed in 3-5 years, and may deliver more capacity than can be initially used. A phased introduction will allow for a steady uptake and capability building, and will stagger the refresh pressure as well.

11. Next steps and time line

This paper will evolve to a broadly agreed position. From that point there appear to be three major themes of activity that can be undertaken in parallel, and started almost immediately: the use-case gathering, the infrastructure design and the governance design; each of them within a solid project management framework and their own timelines. The development of the planning documents will follow naturally from these.

To ensure a timely process, feedback should be provided to the author by **Friday 20 November 2009** (3 weeks in total, 1 week after the e-Research Australasia 2009 conference), to allow for an updated (and near final) version to be released before the end of November. Earlier feedback is strongly encouraged, and the mailing list is available for discussions. Comments are sought from institutions and organisational perspectives as well as the wider user community.

Subject to the feedback provided, as well as the ARCS/EIF processes, the three themes of activity above could then start immediately following and be given some specific timelines of their own. Some preparation can happen in advance, and direct feedback on each of them would also be very helpful.

All three must be carefully aligned with the processes being managed by DIISR for the ARCS/EIF planning, as well as the efforts being undertaken by ANDS. Once those processes become clearer the activities outlined here can be integrated with them and/or inherit from them as appropriate. The ANDS business plan for its combined NCRIS and EIF investments has already been released. The ARCS business plan for EIF is being developed and must be finalised before March 2010. The 'stimulus' nature of the EIF investments provides another indication of urgency, but this is set against the 4-year timeframe for ARCS.

12. Appendix 1: Glossary

The benefit of a (carefully-scoped) national e-research acronym directory is becoming obvious. As an aside the community could collaboratively establish a single online location and include it by reference. Obvious locations could be at www.pfc.org.au or www.eresearch.edu.au. Intersect can assist this process.

13. Appendix 2: References and other useful sites

The references here come directly from the feedback provided to-date, and will be structured and referenced more appropriately in future drafts.

13.1. Background for the purposes and processes

- Towards an Australian Data Commons: http://www.pfc.org.au/pub/Main/Data/TowardstheAustralianDataCommons.pdf
- The Australian Social Science Data Archive (ASSDA – assda.anu.edu.au).
- The US National Science Foundation (NSF)and their review of the needs for "cyberinfrastructure" http://www.communitytechnology.org/nsf_ci_report/report.pdf
- the International Federation of Data Organisations for the Social Sciences (http://www.ifdo.org/)
- EDINA - the UK national academic data centre based at the University of Edinburgh. http://edina.ac.uk/
- The findings of the UK's Office of Science and Innovation (OSI) e-Infrastructure Working Group (http://www.nesc.ac.uk/documents/OSI/)
- Oxford EIDSCR http://eidcsr.blogspot.com/ ;
- NSF DataNet http://www.nsf.gov/pubs/2007/nsf07601/nsf07601.htm
- APSR http://www.apsr.edu.au/publications/index.htm
- ANDS EIF: http://www.ands.org.au/infrastructure.html
- ARROW http://www.arrow.edu.au/docs/
- Robin Rice Data Sharing Continuum http://www.disc-uk.org/docs/data_sharing_continuum.pdf
- "Culture is not a Department: The Role of Governance in National Cultural Institutions". http://www.nla.gov.au/openpublish/index.php/nlasp/article/viewFile/1393/1682
- http://www.naa.gov.au/about-us/director-general/2009-05-21.aspx Last accessed: 21/10/2009 See also the scope for the National Digital Heritage Archive: http://www.natlib.govt.nz/about-us/current-initiatives/ndha Last accessed: 21/10/2009

13.2. Use cases

Researcher perspective:

- The JISC SCARP has done full-immersion analysis of researcher's use of data and released four of the ten reports so far: http://www.dcc.ac.uk/scarp/
- Polydoratou, Panayiota. (2007). Use of digital repositories by chemistry researchers: results of a survey. Program: electronic library and information systems, Vol. 41 (4), pp386-399.
- Coggins, John (2009). A Researcher's Perspective: the Value and Challenge of Data. Paper presented at UKRDS Conference 26 February 2009

- Dinkelmann, Karl; Edwards, Michelle; Fry, Jane; Humphrey, Chuck; Nakao, Ron; & Thomas, Wendy. (2009). Work flows - Data Discovery and Dissemination: User Perspective. DDI
- Van de Sompel, Herbert; Payette, Sandy; Erickson, John; Lagoze, Carl & Warner, Simeon. (2004). Rethinking Scholarly Communication, Building the System that Scholars Deserve. D-Lib Magazine, Volume 10 Number 9, September 2004
- McKay, Dana. (2007) Researcher data practices at Swinburne: results of a survey. Melbourne, Australia: Swinburne University of Technology, Information Resources

Data perspective:

- DCC Curation Lifecycle Model (2009). http://www.dcc.ac.uk/lifecycle-model/
- The Interagency Working Group for Digital Data (IWGDD) data lifecycle
- Dinkelmann, Karl; Edwards, Michelle; Fry, Jane; Humphrey, Chuck; Nakao, Ron; & Thomas, Wendy. (2009). Work flows - Data Discovery and Dissemination: User Perspective. DDI
- Lyon L. (2007). Dealing with data: roles, responsibilities and relationships, Consultancy Report. June,
- 2007, Bath: UKOLN. http://www.jisc.ac.uk/publications/publications/dealingwithdatareportfinal.aspx
- Green, Ann, Macdonald, Stuart, & Rice, Robin. (2009). Policy-making for Research Data in Repositories: A Guide. Version 1.2. JISC.
- Research Information Network. (2007). Stewardship of digital research data: a framework of principles and guidelines. Available at http://www.rin.ac.uk/dataprinciples.

Information gathering

- UNSW MemRE http://membranes.edu.au/wiki/index.php/Main_Page
- Intersect survey
- EUAsiaGrid survey http://www.surveymonkey.com/s.aspx?sm=ZtoA5sG3n0RuLVQ5ggBsdQ_3d_3d

Document Identifier			
Document Author	Markus Buchorn		
Version No.	2	**Version Date**	23 October 2009
Document Name and File Location			

Revision History

Version No.	Revision Date	Summary of Changes	Revised By
1	**30/9/09**	**First version. Asks the questions.**	**MB/IG**
2	**30/10/09**	**Second version. Aggregates feedback from multiple institutions and sets out models**	**MB**

Approvals

This document requires the following approvals.
Signed approval forms are filed in the Quality directory.

Name	Signature	Title	Date of Issue	Version