

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

## Summary of Workshop 2, November 27, 2006

This summary from notes taken at the workshop with amendments following feedback – Paul, 10 Dec 2006

### Introduction – Rhys Francis

<http://pfc.org.au:80/twiki/pub/Main/DataWorkshop2/Intro.pdf>

We are going to look at what might be a national data scheme and what NCRIS might spend money on. A lot is happening in DM in the next few months that we will need to align with, e.g. PMSEIC. Some areas we might examine today:

- Boundaries between service providers and clients;
- Current experiences – successes & failures;
- What access and DM might fall to institutions – moving DM out of its natural environment is a bad idea, e.g. for development of tools;
- Public or private data – complicated, e.g. as universities move to commercialisation;
- Who decides on DM access and ownership?
- What is the lifecycle of data and different missions driving?

The NCRIS capabilities indicate current needs where no current funding. At this stage, we're "on the surfboard" but still on our knees...

- Perhaps we have a "brand problem" with the term 'e-Research'. This is about the future of the knowledge industry and research is the most advanced knowledge industry - it's really about the people, not e-Government or e-Industry;
- Looking at an e-Research trusted federation, national data collection service, national grid and computing facilities;
- Issues with Aarnet and what they focus on as a business – perhaps members have to pay more.
- There are many repositories and organisations, most not under the control of 5.16. NCRIS is only a participant - it's not a simple problem.

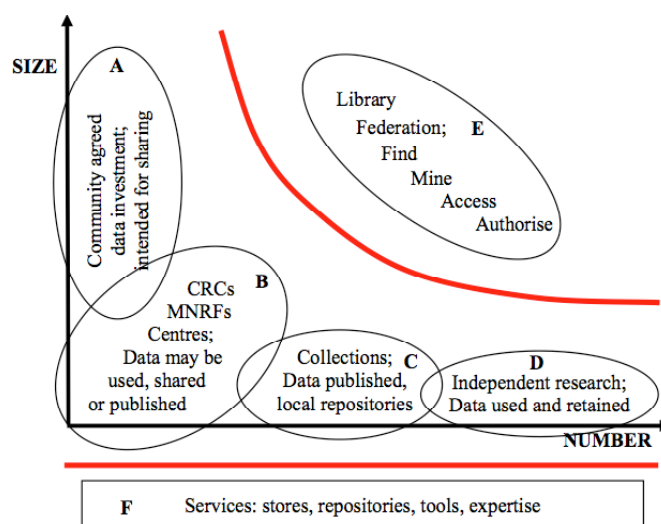
PMSEIC suggests the need for a national strategic approach, some sort of federated solution, desirability of open access and incentives, cultural and privacy issues, etc.

Researchers determine what data is kept, but it depends on discipline, legal issues, commercial requirements, etc. - no single answer. Suggests that institutions need to step in to help, e.g. if researchers move, the institution needs to take over. Institutions will need policies, records, guidelines and archiving systems, etc. The preservation process might be detailed in the research plan, but the institution will probably struggle with this and the researcher wouldn't comply.

Requirement to keep research data for 'not less than 5 years'... Some communities have requirements for community collections, although it is hard for them to get money to support them.

Different data management 'missions' indicate how NCRIS might spend its money (see Fig):

- A. the 'big guys' - these communities know what they're doing, although what happens after the investment period expires?  
B. Bodies collect and manage data. Generally don't publish - business depends primarily on products developed related to investment e.g. BOM. Not our



problem, but a possibility of pushing some data out to an outside DM service provided by 5.16?

C. Purpose is to publish – worth collecting data sensibly.

D. Researchers’ desktops –for field workers this is the only source of data so we can’t ignore. A publication process could use DM services.

F. Services, data stores.

E. Find, mine and access services - not the domain of any single institution and a possibility for 5.16.

- For A & B, who decides about retention?
- What support for C & D?
- What services could assist all of these?  
e.g. if you provide DM tools, data will be used more.
- How will NCRIS use Toolkits and Grid services?
- If AAA funding goes ahead, what are the applications that are actually ready for it?  
– the information providers, Learning Management Systems, others.

Information Providers:	Other Systems:
<ul style="list-style-type: none"> <li>• <a href="#">ArtSTOR</a></li> <li>• <a href="#">CSA</a></li> <li>• <a href="#">Digitalbrain PLC</a></li> <li>• <a href="#">EBSCO Publishing</a></li> <li>• <a href="#">Elsevier ScienceDirect</a></li> <li>• <a href="#">ExLibris - SFX</a></li> <li>• <a href="#">ILIAS</a></li> <li>• <a href="#">JSTOR</a></li> <li>• <a href="#">NSDL</a></li> <li>• <a href="#">OCLC</a></li> <li>• <a href="#">Ovid Technologies Inc.</a></li> <li>• <a href="#">Proquest Information and Learning</a></li> <li>• <a href="#">Serials Solutions</a></li> <li>• <a href="#">Thomson Gale</a></li> <li>• <a href="#">Useful Utilities - EZproxy</a></li> </ul>	<ul style="list-style-type: none"> <li>• <a href="#">Bodington.org</a></li> <li>• <a href="#">Condor</a></li> <li>• <a href="#">Confluence Wiki</a></li> <li>• <a href="#">Darwin Streaming Server</a></li> <li>• <a href="#">DSpace</a></li> <li>• <a href="#">eAcademy</a></li> <li>• <a href="#">Fedora</a></li> <li>• <a href="#">GridSphere</a></li> <li>• <a href="#">GridShib</a></li> <li>• <a href="#">Higher Markets</a></li> <li>• <a href="#">Horde</a></li> <li>• <a href="#">Hupnet</a></li> <li>• <a href="#">LionShare</a></li> <li>• <a href="#">Media Wiki</a></li> <li>• <a href="#">MyProxy</a></li> <li>• <a href="#">Napster</a></li> <li>• <a href="#">PHEAA</a></li> <li>• <a href="#">Sharepoint® from Microsoft</a></li> <li>• <a href="#">SYMPA</a></li> <li>• <a href="#">Symplcity</a></li> <li>• <a href="#">TurnItIn</a></li> </ul>
Learning Management Systems:	
<ul style="list-style-type: none"> <li>• <a href="#">Blackboard</a></li> <li>• <a href="#">Moodle</a></li> <li>• <a href="#">OLAT</a></li> <li>• <a href="#">WebAssign</a></li> <li>• <a href="#">WebCT</a></li> </ul>	

In the middle are the NCRIS investments that the Government has indicated are priorities. Finally, we don’t have a lot to spend...

**Discussion:**

BOM is required to provide data and advice. It has traditional ideas about its mission but is “moving into the information age”. Data released could be used by other businesses, although volume is large. Reluctance to share may relate to a culture of “personal attachment to data” and possibility of litigation over decisions made based on corrupted data. Perhaps need to see data as ‘input’ rather than output in life cycle diagram. Another view is that the level of criticism is a measure of success and actually adds to the collection.

Data needs to be in an open format – vendors lock up data within software licences.

Incentives are needed to encourage researchers to manage data well. They need to see benefits – more than just a code of practice. In the open source model, peer visibility is the driver!

The question is: is the data worth accessing? – garbage in, garbage out.

While researchers decide what to keep, a degree of partnering with data managers would help determine the broader value of data, e.g. medical data aggregation. The volume of data is not related to its value.

IT and Information Management people use the same words in different ways – need to be clear on quality and value of data. People are mixing meaning of ‘data’ and ‘information’ (data with context). There are different perspectives on data management from researchers and library people. Maybe the idea of ‘collections’ makes more sense at this level. We need to decide where to put our effort in the spectrum from raw data, e.g. from satellite, ... to refined information in published papers.

**Paul Fritze – survey update**

<http://pfc.org.au:80/twiki/pub/Main/DataWorkshop2/PaulFPfCdata2.ppt>

The survey of Data Management practices and issues now updated with 24 responses. The basic DM Lifecycle model still fits and is useful to map the range of policies drivers and standards. It reveals 9 broad DM communities characterised by mission/policies, culture and levels of maturity.

1. Major scientific facility collaborations - well established around instrument facilities and of a scale that necessitates national and international coordination of data infrastructure, management and research effort.
2. Professional computing services facilities - project driven HPC and data services. They provide consultancy, training; seed change in research culture; and enable connections between projects and infrastructure solutions.

3. Government research organisations - driven by legislative mandate. They are implementing DM strategies and infrastructure relevant at a national and international level – a possible bridgehead role linking industry, government agencies and the Universities. Data is owned by the organisation.

Four academic research community cultures can be identified:

4. The not-for-profit research organisation - has established a marketplace for data products to maintain a professional skill base and reduce dependence on funding cycles on behalf of research community.

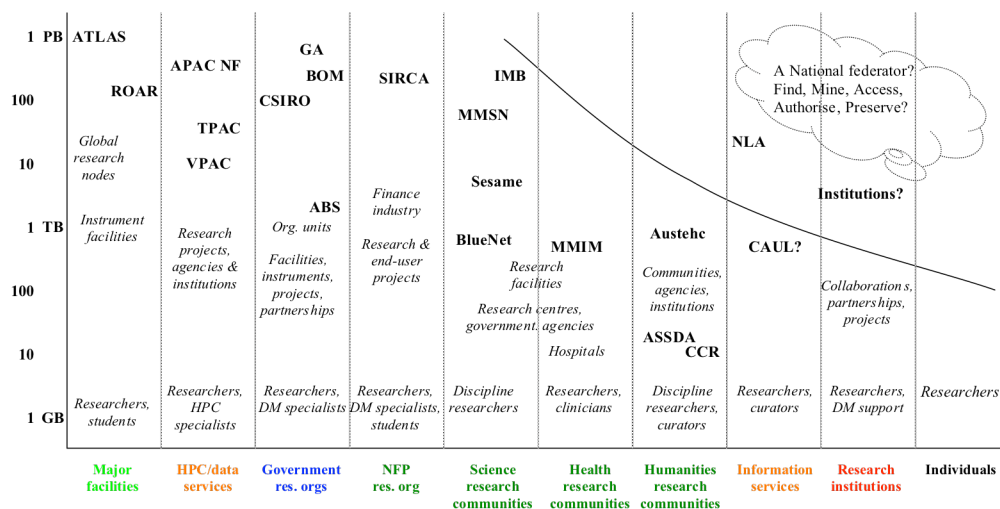
5. Scientific research community - collaborations in various stages of sharing diverse other data from large instrument facilities and individual laboratories. Data ownership is with researcher.

6. A Health research community - characterised by patient related data, clinical settings and multiple end use. Federated access and discovery through metadata - would benefit from National frameworks.

7. SS/Humanities research communities - building qualitative data resources. Shaped by importance of context, non-digital data, broad interests and academic focus; challenged by funding and researcher engagement.

8. Professional information services facilities - support information management and access for broad cultural and research purposes at a national and institutional level. Focus on consultancy, PD and DM policy.

9. Research institutions are establishing DM strategies, services and infrastructure – individual academic research culture and ownership but institutions are beginning to assume some responsibility. Pressures of research performance are key drivers.



## Short presentations

Deborah Mitchell and Sophie Holloway – ASSDA<sup>1</sup>

ASSDA has been running since 1981 at ANU as a data archive for community access. It has now moved to a distributed data model. Other changes over that time include an explosion of data and inclusion of qualitative and government data. Ethics questions about secondary uses of datasets have arisen – i.e. data life after the research cycle. People are tending to hoard data, although ARC has a clause in discovery and linkage grants that all social sciences projects must deposit digital data in ASSDA within 2 years after end of research.

ASSDA is now working with different institutions: UQ specialises in qualitative data, UWA will join in 2007, UM in 2008. Each institute has its own specialty research area, but all are accessed through the same interface. ASSDA needs to accommodate diverse forms of data and it allows local communities to have control. There is a need to differentiate these different nodes, but you can't artificially lay down central data rules. Data is now used by researchers and students in the Group of 8 and regional universities.

<sup>1</sup> Australian Social Science Data Archive <http://assda.anu.edu.au/>

ASSDA has to be careful of generic solutions, e.g. mixing open access and sensitive data will “shut down” researchers. The Social Science community is “very tight”, e.g. the same metadata is understood in UK – so NCRIS can’t override this.

The lack of continuity caused by LIEF funding schemes is a gap NCRIS could step into.

The ASSDA site sets access restrictions down to variable and dataset level. Users can search for data through the central metadata and once found a legal process is followed for access. Users have access to Nesstar<sup>2</sup> analysis tools and metadata usage guides.

### **Nathan Bindoff - TPAC**

<http://pfc.org.au:80/twiki/pub/Main/DataWorkshop2/TPACdata2.ppt>

TPAC<sup>3</sup> is concerned with earth systems data – all about climate change. This brings together a range of disciplines to provide useful a picture for government.

The WOCE<sup>4</sup> project (World Oceans Circulation Experiment) is an example of a legacy dataset that ran from 1988 – 2001. It originated as a collation of data from number of experiments, with 22 nodes and a developing discipline community. Publication of data has shifted from CD-ROMS, DVDs, to online delivery.

90% of data originating from 12 source sites has been collected. Knowing the amount of data made it possible to propose the project to find and rescue these data. It required a team to chase scientists and develop the community. The users (researchers) helped develop the standards and variable definitions for data and metadata. Once established, these made it possible to search and retrieve data across data centres.

- Even though experiments were initially specifically identified, new data streams emerged;
- Over time, data was lost and recovered;
- Data ‘orphans’ developed that had to be re-attributed;
- There is an attempt to capture “self-describing data” – aiming for knowledge discovery;
- OPeNDAP<sup>5</sup> made it possible to make data available online, giving it longevity;
- Although data was “archived”, in the form sent by PIs, it was difficult to “unravel” and recover.

The key to longevity is making data discoverable so it is taken into other datasets and re-used constantly (currently by over 200). Capturing metadata is key.

NCRIS should not do anything that requires research – otherwise it will fail. The technologies must already exist that can be used. The strategy must scale and the scope of the project must be controlled.

Currently TPAC has OPeNDAP servers at multiple partner sites providing a range of products ~ 20-30TB. They have developed a crawler to collect metadata and support both grid and traditional users.

The WOCE business model involved 35 staff in quality control, updating and interacting with providers. The cost of making data accessible is actually very small in comparison with this human interaction.

It is important to develop a continuing relationship with data providers – it is necessary to go back continuously. The focus is on dataset delivery – here you can do “good things” with just 2 EFTs per year. The bottom line is that the cost of delivery services is much less than the cost of community engagement, changing habits, etc.

---

<sup>2</sup> Nesstar online data publishing and analysis solutions <http://www.nesstar.com/>

<sup>3</sup> Tasmanian Partnership for Advanced Computing <http://www.tpac.org.au/>

<sup>4</sup> World Ocean Circulation Experiment Global Data Resource <http://woce.nodc.noaa.gov/wdiu/>

<sup>5</sup> OPeNDAP: Open-source Project for a Network Data Access Protocol <http://www.opendap.org/>

### **Graeme Dudgeon - Department of Primary Industries**

The e-Science initiative is about dealing with data gathered in ~1000 researcher projects/field trials in different disciplines, e.g. fisheries, climate studies. About 95% of these data are not accessible, yet these tiny projects capture hugely valuable datasets – valuable inside and outside the organisation. There is not much in the way of primary datasets.

The aim is to look at patterns and trends across the datasets. Temporal trends are really important, e.g. across different disciplines, such as water, soils, plant data.

The project has \$500k to help link such project information. It starts with incentives to researchers who may well lodge data but won't engage. Researchers are not funded to do this and it would make them non competitive in their research if they did. Perhaps the organisation might subsidise this? It suggests these data need to be managed as collections, i.e. a library equivalent role.

Currently samples may be sent to a lab for analysis, but the possibility of doing analysis across this primary data at this time hasn't been exploited. It is hard to determine the organisational or national boundaries of discipline data, e.g. of plant nutrition. A data management service needs to think about how to create a DM environment where you don't know the full potential use of data. Different projects use different tools in their initial analyses, but we need some sort of classification form to facilitate cross disciplinary use.

Data sharing is not just about researcher data. For example, fisheries 'catch' data – where, when, what, etc. as part of regulatory requirements. These need to be put together with researcher data.

We are not just putting away data for the sake of it – we need some idea of its value and have to start now! We can't go back – it's a big enough task just gathering the new data.

### **Jane Hunter – GRANI project**

The GRANI<sup>6</sup> initiative enables collaborative management and analysis of images and data within the NANO community. The trend is to higher resolution data, more images, use of robots (requiring quicker data collection), remote access and control of experiments, more modelling, 3D reconstructions, more time series data, video recordings of experiments. Data is being integrated from the synchrotron, electron microscopes, etc. to verify and validate models.

Requirements include middleware for data management, security, education and training, data processing, collaboration, publishing and preservation.

Users are not just universities but also corporations, etc. The GRANI-ARC-NANO initiative provides a federated system linking microscopy centres around Australia.

- Developing a portal using shibboleth;
- The different platforms at facilities requires different charging schemes;
- Provide access to services and expertise for instruments;
- Different instruments have different access requirements – need to work with APIs. Using MIT iLabs<sup>7</sup>.

This is heterogeneous, multidisciplinary research, with different communities using the same material. There is a need for:

- High speed processing;
- Ways of incorporating legacy data;
- Common services, e.g. Matlab<sup>8</sup>;
- Ways of accessing and using any instrument on network to process and save data.

APAC is providing a central repository for data management and federated database with each instrument. It is necessary to capture the full context with each image, e.g. the processing steps across multiple users and

---

<sup>6</sup> Grid-enabled Archive of Nanostructural Imagery <http://www.itee.uq.edu.au/~eresearch/projects/grani/>

<sup>7</sup> MIT 'iLabs' - internet access to real labs - anywhere, anytime <http://icampus.mit.edu/ilabs/>

<sup>8</sup> MATLAB <http://www.mathworks.com/products/matlab/>

processes - detailed metadata is necessary. People also need to integrate data from public databases – metadata on ontology schemas and links to publications are required.

Telepresence provides remote access to and possibly control of experiments via iLabs and Cima. Users are able to communicate with the technician during a session and collaborative drawing tools are being developed to assist in communications. Annotation of video and 3D images by multiple groups is possible.

Publishing tools for researchers to package and put data in the repository are being developed.

Materials science is ~3 years behind bioinformatics –

- There is a need for cultural change;
- Most researchers want a proprietary period of use of data;
- Need persistent identifiers for people, experiments and data objects;
- Need ICT expertise and support.

### Ben Evans – ANU and APAC NF

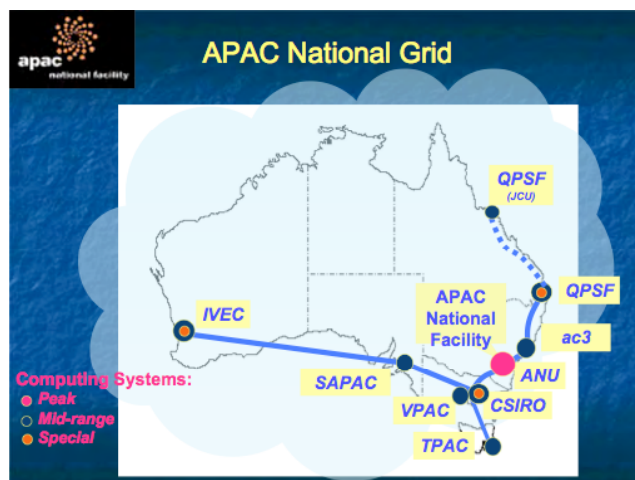
<http://pfc.org.au:80/twiki/pub/Main/DataWorkshop2/ANUdata2.ppt>

ANUSF case study – first university mass data storage system started in 1995 to store data from computation and other sources beyond capability of individual areas. The focus has now expanded to include complex datasets, data analysis and expanded services and expertise. ANUSF manages in the order of 400TB of raw storage. New projects starting in 2007 will substantially accelerate the data uptake.

ANU established an e-Research task force to investigate various movements and methods of supporting e-Research activities at the ANU. They found increased data-enabled research in all areas - each group saw data as critical. There is a view towards central integration of management of digital assets and need to establish the value of e-Research and build in processes.

ANU continues support for the APAC National Facility<sup>9</sup> which includes a major national computational facility and data service. This data service approves project levels through merit allocation processes by leading academics. Assessment is done on a yearly basis to see how researchers are using it. Research project plans need to show value of data projects. At present APAC National Facility may require funding arrangements to be discussed for very large projects.

The responsibilities for principal researchers are for overall project and outcomes, establishing collaborations. The role of APAC National Facility is to provide infrastructure, connectivity, provide a cohesive and extensible software environment, and provide high-end expertise and researcher support. There is no single software system that covers the needs of all projects any many such systems are currently supported by the facility.



The overall goal is best overall management of data within lifecycle as negotiated. We are moving from a large data archive strategy to infrastructure that supports different styles of access, including large on-line data and managing relational databases in real time. Having computer facilities next to large datasets is critical as is high data bandwidth to national and international networks. Software management is also critical - a toolbox/registry of packages is available on the National Facility web site.

Management through generations of hardware, software and changing data standards in research projects is common in the data lifecycle management. The trend is currently that large projects are to becoming larger as bandwidths and technology improves. Historically this has been led by scientific datasets but now other

<sup>9</sup> APAC National Facility <http://nf.apac.edu.au/>

projects such as humanities with video data are becoming larger. However there is a large trend for all projects to have complex data management aspects, which covers all disciplines using the facility. There is now much more interest in real time access, and data access arrangements. Data transfer requirements require high bandwidth transfer mechanisms and currently exceeds the network bandwidth available in Australia. APAC has also developed a data transfer fabric to enable large datagrids and transfers to be managed across the country or internationally. Many areas are increasingly outside researchers' competency or core interest areas and have found more interest in working in teams.

APAC NF has provided data management support for three of the four presentations today, and looks to work with researchers as part of the national data management service.

**Roze Frost – CSIRO**

<http://pfc.org.au:80/twiki/pub/Main/DataWorkshop2/CSIROdata2.ppt>

CSIRO needs to work out an Information Management framework to take advantage across all communities, e.g. APAC and CSIRO. A matter of 'information' rather than 'data' management. Currently datasets are not connected. CSIRO's biggest asset is its science data and we need to add value to this.

There are disparate technology platforms – currently spending \$50M to fix this. Infrastructure needs to be seamless to researchers across enterprise centres and data storage services. An Aarnet 3 backbone and fibre network is being set up. The desktop becomes the key portal to repositories, administration, etc.

There is a need to look at people, processes and technologies holistically – in the Strategy Plan 2007-11.

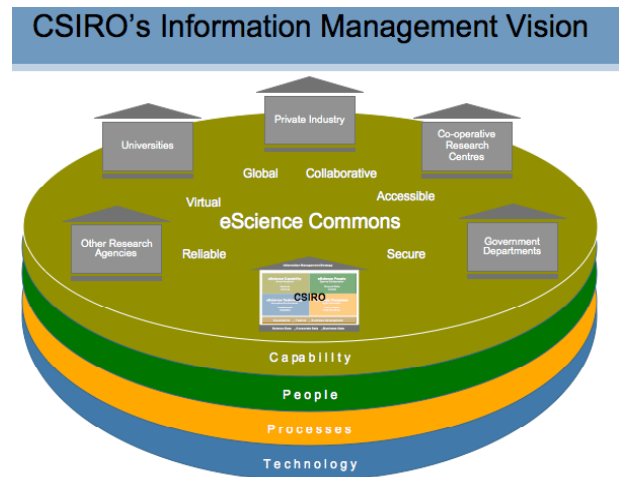
- technology alone never works;
- the capabilities for doing (e-) research is best left to researchers;
- importance of people side – requiring cultural change;
- The CSIRO workforce plan identifies an information services role for librarians;
- How do we grow APAC, HPC? e.g. cluster computing;
- Need to understand privacy issues and metadata across disciplines.

All this requires underpinning of a governance plan.

The idea of a scientific commons might be to provide an information management framework across all organisations in CSIRO: the key is collaborations, common processes, technologies, capable people and building on what is already existing.

There is a spectrum of involvement from research support to scientific research. Activities you can define can be pushed more into support services end.

As CSIRO moves to a more collaborative involvement, it's likely that IP will be shifted to the point of discovery. Ownership needs to be clearly defined as you share data with other institutions.



[In TPAC cross disciplinary data and requirements for real time access is leading to removal of embargos on data access – these are actually a burden.]

Ownership is an issue that hits a nerve - something NCRIS could build into its report to unpack to semantics and legal issues. This is something that could be done at 'no cost'.

**Linda O'Brien – University of Melbourne**

There is a window of opportunity as MU implements its 'Triple Helix' vision<sup>10</sup>. This is providing opportunities to leverage e-Research to inform and link across the strands of research, teaching and learning,

<sup>10</sup> Melbourne Uni. Growing Esteem vision <http://growingesteem.unimelb.edu.au/strategicplan/vision.html>

and knowledge transfer. Promotion criteria need to be reviewed with an emphasis on academic ‘leadership’ in research, teaching and learning, and knowledge transfer.

Melbourne is a very devolved institution. A major project is underway to revise infrastructure - to identify what’s common, what’s unique, common standards, etc. Information Services providers are also responsible for corporate management and archival collection. In the area of publication, the ePrints<sup>11</sup> repository is being linked to the research management system. This will provide researcher profiles. The University is trying to get a small critical mass of people to move things forward, e.g. in Information Management, HPC, library. This requires creation of new roles.

### Gaven McCarthy – Austehc

Austehc has been a science archival project running since 1985, starting with only a small amount of money. Transformation by the Web made control of publication possible and revealed latent demand for materials.

One example of data is a collection of photos of plant life in Victoria taken immediately before Black Friday in 1989. This environmental snapshot is of immense value to researchers.

The professional archival process is not just about creating a product, but of “telling the story of the records”. It requires the original materials, locations, people, publications and digital objects within international standards. Austehc is working with the NLA on this. Datasets need to be linked in a contextually meaningful framework.

The NLA have made their database open to Google. This public access and being able to cite every work of scholarship completely changes the nature of scholarship – this is really important. The NLA has a legislative mandate to do this, although they have resource limits. [perhaps the NLA is a potential receiver of NCRIS funds – a national home for collections?].

Picture Australia<sup>12</sup> and the new People Australia are datasets using people identifiers.

### Adrian Burton – APSR

<http://pfc.org.au:80/twiki/pub/Main/DataWorkshop2/APSRdata2.ppt>

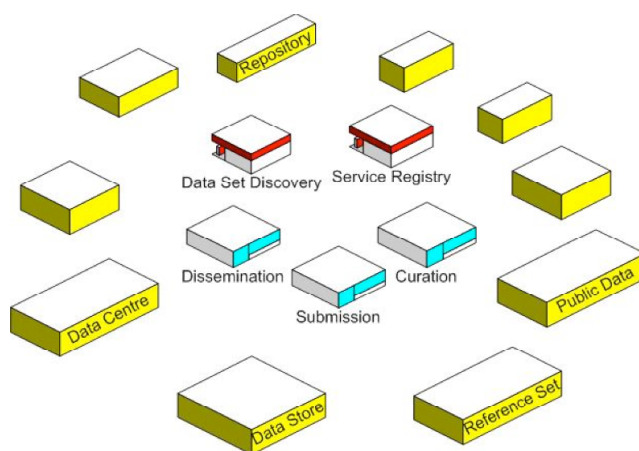
APSR<sup>13</sup> is involved in many partnerships, e.g. with NLA, institutions. It is funded by DEST as a Systemic Infrastructure Initiative. Its aims:

- to raise awareness of data and direct people to services;
- provide a registry of datasets (collections) and services on how to use them;
- enable cross dataset workflows within the NCRIS projects – item and collection discovery, access and third party use.

Integration requires interoperability across discipline data.

The Collections Service Registry Project aims to engage different research areas; to gather existing collections/repositories/data centres. It provides both alerts services and methods for discovery, e.g. of institutional repositories, both national and international. A pilot will run in 2007 using APAC as the basis and extending to DART<sup>14</sup>.

If the CSR was applied in the NCRIS environment it would be possible to build middleware tools and



<sup>11</sup> University of Melbourne ePrints Repository <http://eprints.unimelb.edu.au/>

<sup>12</sup> Picture Australia <http://www.pictureaustralia.org/>

<sup>13</sup> Australian Partnership for Sustainable Repositories <http://www.apsr.edu.au/>

<sup>14</sup> DART project <http://dart.edu.au/>

scale up the registry. Additional services could be provided by other providers, e.g. services for dataset discovery, curation, submission (e.g. from research tools), service registry – connecting to repositories, data centres, repositories and public data.

Supporting the “human network” is important, e.g. to facilitate communication between data collection managers, support people. This is the key – and comes first! APSR’s role is to broker sector-wide coalitions. The set of services needs to be pushed by the participants – the people who run the collections.

[Comments: we need to identify who are the key institutions, layers of government, state organizations. The ‘Peoplepicker’ service<sup>15</sup> is coming online.]

### **Ah Chung Tsoi – Reflection on broader issues**

In trying to understand Rhys’s diagram the conclusion is that data management is complex – we need to make assumptions and set principles.

Dimensions of data management that make decisions difficult:

- Ownership – this is complex and difficult e.g. is it researchers, ARC after a period, students?
- Data format – raw, processed, published – often not specified;
- Accessibility – who, over what time frame?
- Metadata;
- Intellectual property – who owns the IP?;
- Privacy – e.g. it is not simple to merge clinical and other data in MMIM. Suggests need for trusted authority.

Assumptions that can be made:

- Public funded data should be available to others;
- Need availability of the raw data, rather than just the processed data;
- Data should be immediately available if required – and close;
- Life cycle issues – data is available across the lifecycle;
- Meta data is available – compatible and interoperable;
- Assume no IP or privacy issues.

Guiding principles are:

- Public datasets are supported publicly;
- Private or limited access datasets are not publicly supported, but
- Metadata is publicly available.

Option 1: Datasets stored in institutional repositories, research websites. A possibility is for institutions to be federated. There are logistical issues and a need to justify to institutions. If data requirements are not known at start of a project, it is difficult to arrange funding. Understanding of value is different at the end.

Option 2 (less desirable?): Datasets are held in centrally located and supported in public repositories, e.g. NLA. The institution submits and leaves management to repository.

What do we store? Perhaps it is like archaeology – leave it to chance and future generations to work out?

### **Paul Davies – VerSi**

<http://pfc.org.au:80/twiki/pub/Main/DataWorkshop2/VerSiData2.ppt>

The e-Research coordination committee: e-Research is about data, information and knowledge. Infrastructure should be planned as a national resource.

E-Research provides links across the Science value chain:

Basic (Universities) ... Strategic & Applied (CSIRO, DSTO, etc) ... Adaptive Science (Private).

VerSi<sup>16</sup> is a coordinated approach to adoption of e-Research. A 5 year/\$10M initiative:

- mainly about awareness and support at this stage;

---

<sup>15</sup> MAMS PeoplePicker service <http://mams.melcoe.mq.edu.au/wiki/display/PPKR/>

<sup>16</sup> Victorian eResearch Strategic Initiative <http://versi.edu.au/>

- building security and storage solutions;
- demonstration projects, e.g. VBL Virtual Beam Line collaborations for the Australian Synchrotron;
- integrated with national strategy.

It uses multi-site SRB federation, Shibboleth and MAMS, and interoperates with other federations, e.g. DART/Archer.

VerSi is strongly about ‘user pull’, rather than ‘technical push’ – there is a strong need to engage with users. The technology is already here – we just need to fill in the gaps.

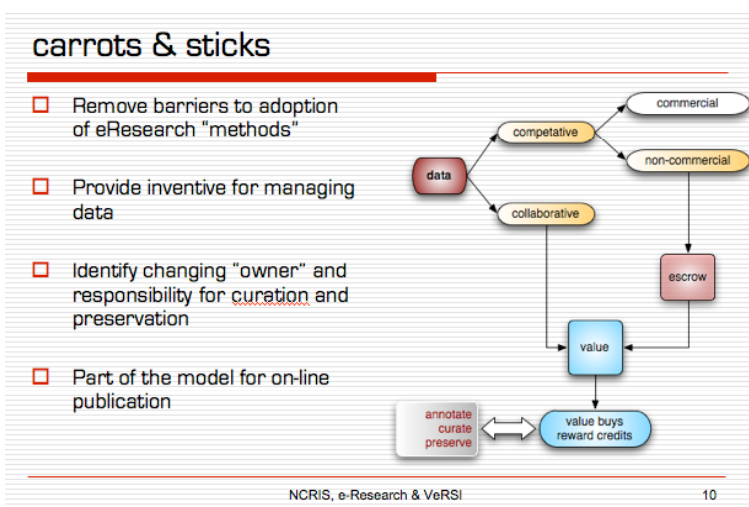
In the Synchrotron project, for example, users wanted to bring data together from multiple experiments. They wanted data security, audit trails to ensure data integrity; to store data for 5 years; metadata automated as far as possible; and to store data at mirrored site. A national data architecture is a result. Raw data is collected at the different instruments, housed at local eResearch centres and institutions, but all accessed by the researcher.

Data from instruments is of national importance and the collections are of strategic value. NCRIS could provide funds for data storage and services for experiments and lead national debate.

Carrots and sticks are needed to remove barriers to the adoption of e-Research.

Data can be categorised as either ‘competitive’ (commercial/non commercial) or ‘collaborative’. It is necessary to assign a value to data, e.g. for promotions, to establish obligations for archiving, curating and preserving.

Data can also be described as private and public (there is a lot of public data - very simple to deal with.) Data may also change in role over time - the difference between data and metadata is also noted.



[Who makes these decisions about value?? Maybe NCRIS interest should be in collaborative data. The problem is too big for individuals – what can NCRIS do about it? NCRIS can connect services, but can’t buy storage].

OECD incentives include tax deductions for companies to put data out to public.

## Concluding thoughts – Rhys

Data	Tools
Competitive	Federation
Collaborative	Interface tools
Public	Repository
Raw data	Metadata
Research (curated)	
Published data	

The tools are “incredibly important”.

A lot of people are in the repository space but the rules are hard to figure out. There appears to be one missing in the Social Sciences area – how is it that a requirement to use is written into ARC funding but no permanent funding exists?

If data is not publicly funded, why would 5.16 publish? e.g. there is already a capability for medical data. The state of data will affect what NCRIS spends on and any collection has multiple possibilities. We could use existing services, e.g. CSIRO.

We need a rationale to go into the repositories area – which datasets? e.g. difference between raw, collated and published data. One principle is to not transfer funding.

How do you value data? What additional services are needed to get nationally significant value out of data? You can't just look at value in isolation – need to be in terms of collaborative value and use. For 5.16, a merit process is needed to prioritise which datasets should be focused on.

If only public data is funded, we could highlight others. An access policy could be defined up front - CCLRC are trying to do this. The role of research organizations is the key.

NCRIS funds are on making data accessible nationally and accessible to all researchers - in 5 years a whole lot of data will be available from NCRIS. A lot of money is being put into specialist areas – this should be done in a way to consolidate expertise, build skills, consultants, but not to create a cartel.

There is a shift from small researchers to teams – NCRIS needs to disseminate skills into research teams. We need to build on, rather than create, e.g. add new people to existing skilled groups to change directions and scale operations.

ARC should require a DM publishing plan, audit process and an accredited repository to be specified. Everyone has to adjust to the idea of spending on data differently.

The best places for repositories are at the intersections, e.g. institutions sharing, global collaborations. Market forces will eventually provide pressure to rationalise. Eventually we might expect only a few large data centres across Australia.

NCRIS 5.16 process from here:

1. Find out what's in PMSEIC report on Dec 8, also RQF policy;
2. Reference group to work out next steps;
3. Work out what federated services might be required.

Summary by Paul Fritze, University of Melbourne