



# SUPPORTING THE DATA LIFECYCLE: WORKSHOP SUMMARY

Andrew Treloar  
Deputy Director, ANDS

## Notes on the following

- Following slides are one attempt to condense the main messages into a single slide
  - they should be read in conjunction with the richer detail in the individual presentations

## Ross Wilkinson: data-driven research world

- Need to match services to researchers' needs
- Australian Code says data is a first-class output
- Potential conflict between system-level advantages of data-sharing and individual drivers
- Conceive, Acquire, Explore, Share, Publish, Re-use
- ANDS aiming for early exemplar discoverable data spaces
- More researchers re-using more data more often

## Jim Richardson: data management reqts

- eResearch development turns out to be harder than one might expect
- Building it doesn't mean they'll come
- We have the opportunity to benefit from lessons hard learned elsewhere (i.e. Grid)
- Vital to co-develop infrastructure and services together with researchers and in accordance with their actual requirements
- Long-term eResearch infrastructure will grow organically, not by design – we are not so much building it as fostering it
- Before we look at answers we need to be asking the right questions!

## Leon Sterling: Unimelb experiences

- Research data management is a compliance issue
- APSR survey a useful way to start conversations
- Attempt to identify uni-wide data centre requirements was a failure and abandoned
- Address long-term storage needs first
- Hard to identify right services and financial model
- Under pressure, people revert to old behaviour
- Easy to be diverted by other priorities

## Andrew Cheetham: UWS experiences

- Trying to bring HASS sector along ('e-research' is a turnoff for them)
- Useful responses to APSR survey
- Identifying e-research needs
- Data mgt needs to be embedded in workflows

## Questions and Discussion

- In some areas quicker/better to regenerate data
- Many small projects may be harder than some large projects
- ‘Petabyte envy’ in the health sciences
- Importance of sharing methods as well as/instead of their data
- Need to look for common patterns of use

## Mark Wilkins: bioscience researcher req'ts

- New bio technologies producing data avalanches
- Needs:
  - raw storage (short and long term)
  - analysis pipelines (raw to processed)
  - data comparison tools
  - data visualisation tools
  - databases (high and low quality data)

## Mark Turnbull: HASS perspective

- *Our Cultural Commonwealth* a useful guide
- <http://www.historycooperative.org/journals/ahr/105.1/ah000001.html>
- Web 2.0 critical for humanities because they have always connected to the public
- Cultural challenges around data production in long-lived formats
- Data re-use in non-conventional ways and new media
- Shift towards complex cultural data objects with attendant curation challenges

## AM Discussion, Table 1

- Important to capture structure and relationships between some data types
- IT can't make selection decisions
- Services required:
  - 'benevolent' IT support
  - information clearing house
  - visualisation tools for collaboration networks
  - digitisation
  - IT infrastructure needs to be funded as other infrastructure
  - tame BAs to bridge gap between researchers and IT

## AM Discussion, Table 2

- Accessing own data vs getting access to others
- Things like lab notebooks can help standardisation
- Data gift economy can leverage change
- Concept of data ownership is poorly understood
- Just adding a dept server would improve things
- Public funding should mean data sharing post-pub'n
- Making data accessible is both a process and social issue
- Relationship between institutional and global stores

## AM Discussion, Table 3

- 'Data' is something very diff. for humanities
- Need to aggregate health data at many levels
- Commonalities across health and humanities
  - digitisation
  - data mining
- Requirements
  - sharing of methodologies, workflows, data models within disciplines
  - need for IT support to be embedded with researchers
    - possibly postdocs 'gone bad'
  - anonymisation of health data

## AM Discussion, Table 4

- Think about research lifecycle, not data lifecycle
- What is needed to bootstrap better data mgt and sharing practices
  - data dictionary and standardisation
  - support for cross-disciplinary practice

## Overall AM discussion

- Need rock-solid storage solution with flexible APIs to build on top of it
- Difficult for one institution to handle all kinds of content
  - better to go for sector-wide approach
- Hard to get a single researcher voice to argue for a particular approach (unlike, say, HR)
- Need for more things like bio-informaticians

## Data store discussion

- JR: benefits in working collaboratively towards data storage solutions through ARCS and ANDS
- RW: there may not always be common solutions, but there are often common 'business' needs
- NC: bilateral arrangements are easier than multilateral
- TC: need for services for users on top of raw storage
- RW: relationships are the best metadata
- LS: sometimes hard to align researcher and uni priorities

## PM Discussion: Table 1

- Required services:
  - collaborative editing and annotation
  - environments for collaborating around data
- but collaboration (in different ways) around data for different disciplines

## PM Discussion: Table 2

- Definitional discussions around archiving, curation
- Need clear understanding of terms
- Distance to service provider reduces trust
- Need citeable objects for reuse
- Better bindings between publications and data when submitting

## PM Discussion: Table 3

- Need drivers for desired behaviour
- Assign identifier as early as possible
- Need identifiers for collections and people
- Need flexible data naming structures because understandings evolve
- What is publishing?
- Timestamp service for data
- Flexible authentication/authorisation services

## PM Discussion: Table 3

- Re-use requires ability to
  - discover
  - access
  - manipulate

## PM Discussion: Table 4

- Sharing and Re-Use: Available
  - good in some disciplines only
- Sharing and Re-Use: Desirable
  - citing data
- Minimal metadata
  - what is needed for understanding, reproduction
- Need to be able to search globally, not just in Australia

## Overarching issues

- Data is no longer just a by-product
- Need to better support researchers
- Who does what (user, group, dept, institution, PfC, global)
  - but see distance problem
- Challenge of co-ordinating solutions
- Chris Rusbridge approach:  
<http://digitalcuration.blogspot.com/2009/02/national-research-data-infrastructure.html>

## Other stuff

- Pathways work on dis-aggregating scholarly communication lifecycle
  - **Registration**, which allows claims of precedence for a scholarly finding.
  - **Certification**, which establishes the validity of a registered scholarly claim.
  - **Awareness**, which allows actors in the scholarly system to remain aware of new claims and findings.
  - **Archiving**, which preserves the scholarly record over time.
  - **Rewarding**, which rewards actors for their performance in the communication system based on metrics derived from that system.
- <http://www.dlib.org/dlib/september04/vandesompe/09vandesompe.html>

## General discussion

- Need for a two-page description of each of ANDS and ARCS
  - RW: we will do this and send it to everyone here
- How about a group-editable wiki page for people to type up their requirements?
- Need a shared language of discourse
- Not a lot of scenario work today – what will ARCS/ ANDS be doing
  - RW: working with the ready, willing and able

## General Discussion

- Need a 10-year plan
  - RW: we have one through to mid-2011, and we are building on a lot of very good existing work
- ANDS/ARCS will undertake to provide a space to continue the discussion