

AUSTRALIAN NATIONAL DATA SERVICE: OVERVIEW AND UPDATE¹

Andrew Treloar
Director, ANDS Establishment Project

Outline

2

- Context
- Platforms for Collaboration
- Rationale
- Structure
- Establishment Project
- Progress

eResearch Co-ordinating Committee Strategy

Thematic Issues

- Continuing Need for a Focus
 - through national coordination
- Human Capabilities
 - People, skills and understanding
- Linkage of eResearch Resources
 - seamless access to resources
- Access to Data
 - best practice data management and curation
- Structural and Cultural Change
 - evolution of organisational structures and cultures
- Awareness and Support
 - develop researchers' ability to adopt eResearch

Service Clusters

- Data
 - outreach, curation, data management
 - meta-services, location, access, movement
 - practice, providers and users
- Computing
 - capability computing facilities
 - national computing environment
- Interoperation
 - discipline services (tools ((software))
 - user and operations support
 - collaboration services support
- Access
 - the Australian access federation
 - the Australian research and education network

Australian Code for the Responsible Conduct of Research

- The objectives of the Accessibility Framework are also reflected in the new *Australian Code for the Responsible Conduct of Research*
- Published in 2007, the Code replaces the former Statement and Guidelines on Research Practice (1997) from the universities and research funding agencies
- It describes the responsibilities of institutions and researchers in the management of research data and primary materials
- Eg. Institutions are to retain research data, provide secure data storage, identify ownership and ensure security and confidentiality of research data

http://www.nhmrc.gov.au/publications/synopses/_files/r39.pdf

National Collaborative Research Infrastructure Strategy (NCRIS) follows a decade of investment

1997: High Performance Computing Committee

- Established the Australian Partnership for Advanced Computing to provide access to high performance computing capability

2000: Advanced Networks Programme

- Established advanced demonstrator networks

2002: Higher Education Bandwidth Advisory Committee

- Established the Australian Research and Education Network Advisory Committee, and the Australian Research and Education Network

2004: Research Infrastructure Taskforce Report

- Established the National Collaborative Research Infrastructure Strategy Committee to implement a program of strategic investment in research infrastructure

2006: eResearch Coordinating Committee Report

- Outlined an integrated program of skills development and of middleware and computer science research

2007: NCRIS Platforms for Collaboration

- Commitment to an infrastructure program covering computing, data and inter-operation components, and supporting the development of the Australian Access Federation

NCRIS Investments

\$542M** over the five years: 2007-2011

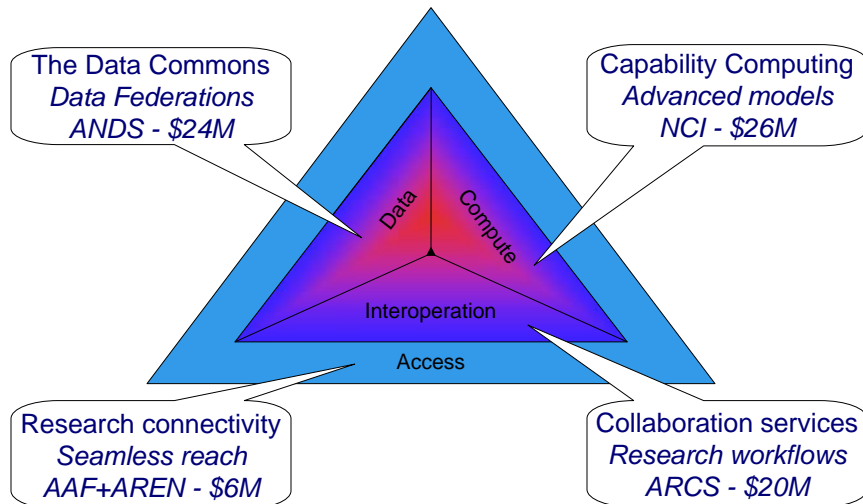
- | | |
|--|--|
| <ul style="list-style-type: none"> • Evolving bio-molecular platforms and informatics • Integrated biological systems • Characterisation • Fabrication • Biotechnology products • Optical and radio astronomy • Integrated marine capability • Structure and evolution of the Australian continent | <ul style="list-style-type: none"> • Networked biosecurity framework • Population health and clinical data linkage • Terrestrial ecosystem research network |
|--|--|

+ Platforms for Collaboration (allocated \$82 M)

**Note: scaled to EU or US economies this is analogous to 1B USD per annum

Platforms for Collaboration: Major Investments 2007-2011

7



Towards the Australian Data Commons

8

- Developed during 2007 by ANDS Technical Working Group
- Mapped out coherent vision of what needs to be done in the data space
- Available at <http://www.pfc.org.au/bin/view/Main/Data>

Why Data? Why Now?

9

- We are in an era of increasing data-intensive research
- Almost all data is now born digital
- Increasing amount of data generated (semi-)automatically
- *“Consequently, increasing effort and therefore funding will necessarily be diverted to data and data management over time” (Towards the Australian Data Commons (TADC), p. 4)*

Need for standardisation

10

- Software and hardware keep getting cheaper, wetware keeps getting more expensive
- Fixing data management problems is enormously labour intensive and costly
- *“Consequently, standardisation within forms of data and simplification in the frameworks around retention, storage, access and use of data, and the elimination of differences whose resolution requires labour, must be made, if the on-going keeping and reuse of data is to remain affordable” (TADC, p. 5)*

Role of data federations

11

- With more data online, more can be done
- Possible now to answer questions unrelated to reasons why data was collected originally
- Increasing focus on cross-disciplinary science
- *“Consequently greater clarity is needed over control and access to community-funded data, and the means of aggregating, federating and accessing such data are increasingly important”* (TADC, p. 5)

Data unlocks potentials

12

- New scientific instruments
 - ▣ Large Hadron Collider at CERN generates 1.5 gigabytes of data per second
 - ▣ the Square Kilometre Array (1 EB/day!)
- New scientific Models
 - ▣ The mapping of the Human Genome: A billion DNA letters in a human sequence
 - ▣ Global climate models
- New teams
 - ▣ 195 scientists mapped the genome of the fruit-fly
- New knowledge from unlocked data
 - ▣ Most research from Hubble telescope data was not “first use”
 - ▣ Common data sets unlocked the power of search technology – TREC

TREC- Text retrieval success story

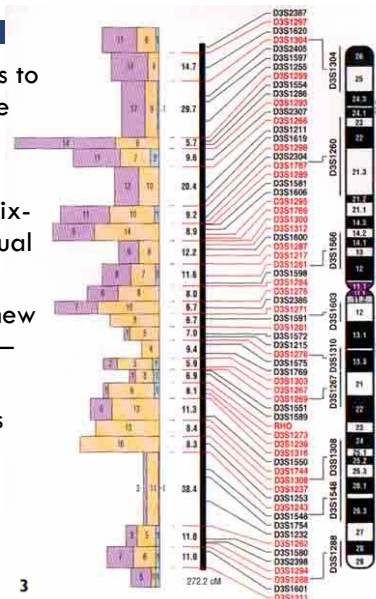
13

- Till 1992 - text retrieval research was done by small teams on small collections – pre AltaVista/Google
- In 1992, a US program was initiated where teams from around the world worked on same problems and reported on success/failure collaboratively
- The success was amazing – text retrieval by best systems was up 25% per year for first four years
- Shared data/problems/solutions were the key
- <http://trec.nist.gov/>

Mapping the Human Genome

14

- Took a large team of scientists 10 years to map the 30,000 genes that describe the human body
- In 2007, Craig Venter, published his complete DNA sequence, unveiling the six-billion-letter genome of a single individual for the first time
- The work required a large team using new instruments to produce a large dataset – indeed 2 competing large teams!
- No single lab could have completed this project with available technology in a reasonable time



The Hubble Telescope

15

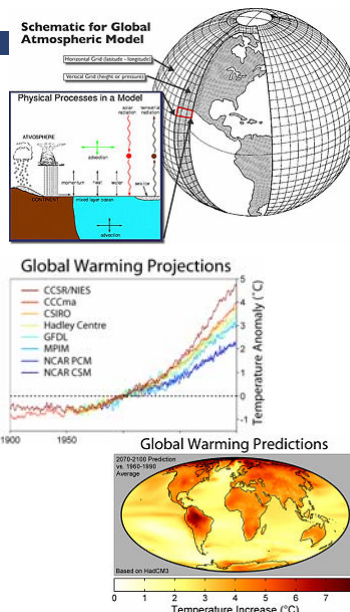
- The Hubble telescope launched in 1990
- Increasing focus on cross-disciplinary science
- Observations are proposed, and if accepted, data is collected and made available to the proposers – who then write a research paper
- Each year around 1,000 proposals are reviewed and approximately 200 are selected, for a total of 20,000 individual observations
- The data is stored at the Space Telescope Science Institute
- *There are more research papers written by “second use” of the research data, than by the use initially proposed*



Global Climate Models

16

- Predictions of climate and global warming supported by global climate models
- Models are large, data very large – typically high levels of co-operation between both institutions and researchers
- Data comes from:
 - ▣ Ice cores
 - ▣ Weather station observations
 - ▣ Satellite observations
 - ▣ Glacier data....
- Multi-disciplinary data for a big problem



The ANDS Vision

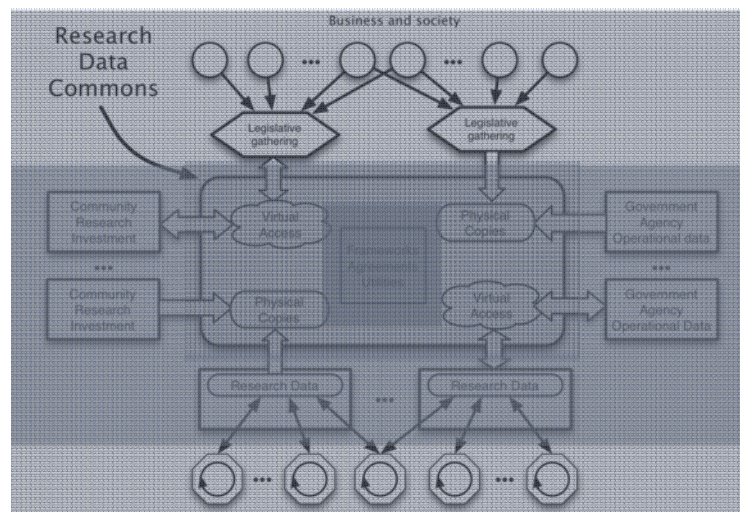
17

- “As a vision, ANDS sets out to transform the disparate collections of research data around Australia into a cohesive corpus of research resources. This transformation would assist the connection of Australian and international data centres, repositories and online collections to enable serendipitous discovery, cross-disciplinary research, and cross-repository workflows.” (TADC, p. 5)

Realising the Vision

18

- Develop the capabilities of research data centres/repositories



ANDS Delivery Structure

19

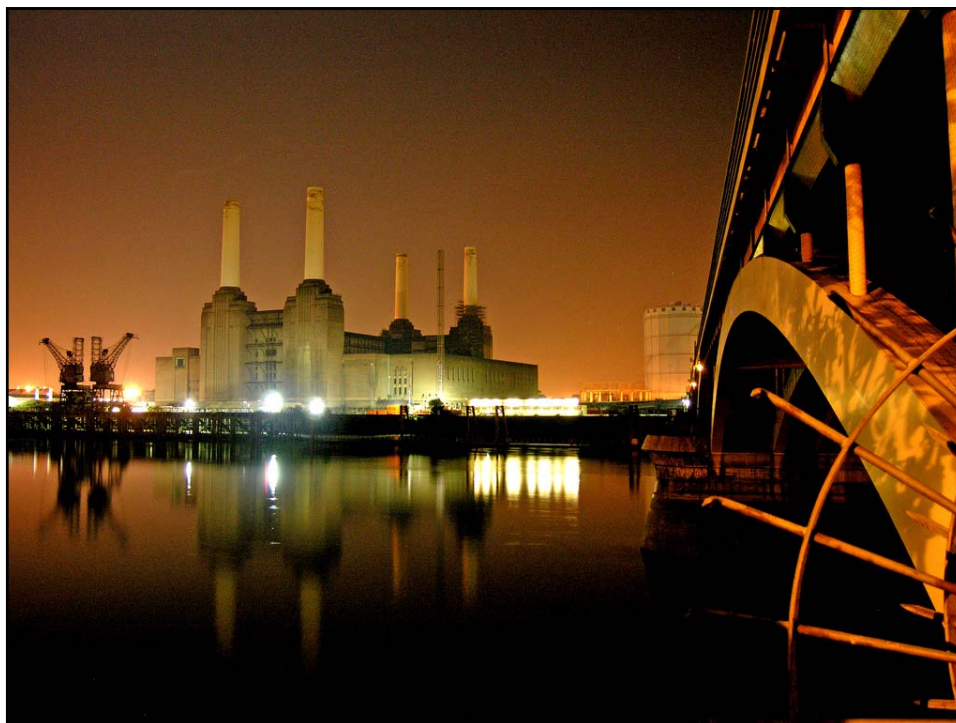
- ANDS has been structured as four inter-related and co-ordinated service delivery programs:
 - ▣ Developing Frameworks
 - ▣ Providing Utilities
 - ▣ Seeding the Commons
 - ▣ Building Capabilities
- Plus development activities funded through National eResearch Architecture Taskforce projects



Developing Frameworks

21

- Influencing relevant national policies
- Building common understanding of data management issues and solutions across government, research funding agencies, and research intensive organizations
- Encouraging moves in favour of discipline-acceptable default sharing practices
- Largely centralised, with some specialised outsourcing



Providing Utilities

23

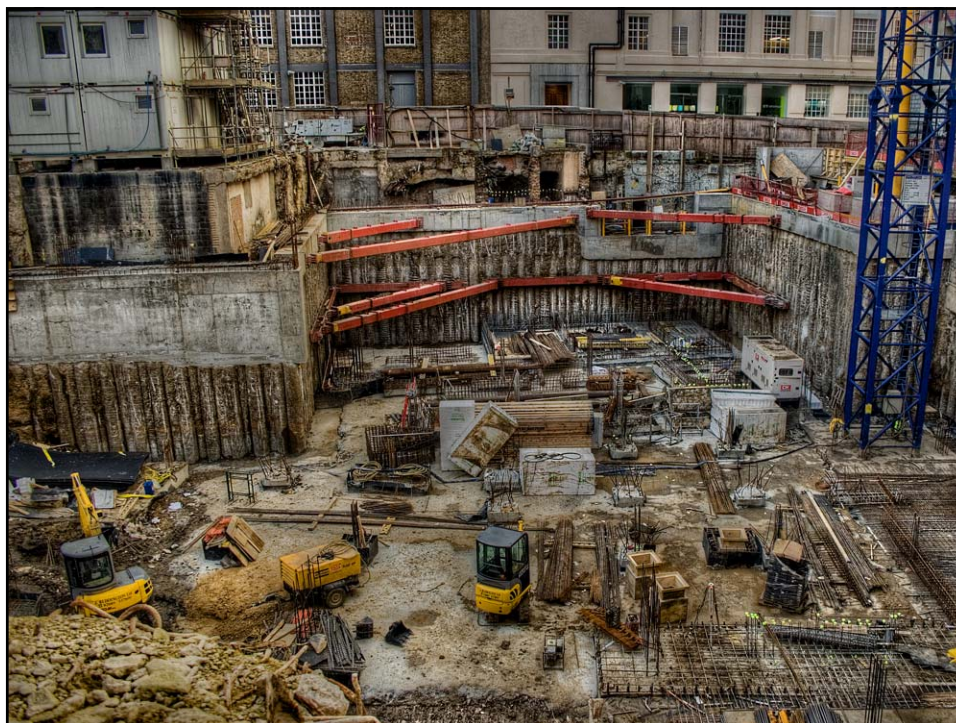
- Building and delivering national technical services to support the data commons
- Examples:
 - Discovery
 - Both “you come to us” and “we come to you” flavours
 - Probably a two-step process for some collections
 - Persistent identifier
 - Collections registry
- Mostly outsourced delivery
- Some insourced technical framework development



Seeding the Commons

25

- In targeted areas (not enough resource to do everything), working to improve:
 - fabric for data management
 - amount of content
 - state of data capture and management
- Plus, opportunistic content recruitment in first year
- Selection process to identify targets
- Placement of ANDS-funded staff, together with co-investment



Building Capabilities

27

- Improving level of capability for research data management and research access to data
 - Train-the-trainer model
- Providing capability within ANDS for integration of existing systems into Australian Data Commons
- Building community around data management concerns
- Largely distributed

ANDS Establishment Project

28

- Monash (lead), ANU, CSIRO
- Four main deliverables:
 1. Formal collaborative agreement to deliver ANDS
 2. Funding agreement with DIISR for ANDS
 3. Selection process leading to an offer of employment for the ANDS Executive Director
 4. ANDS Business Plan for FY 2008/2009
- Originally funded to June 2008; likely extension to December 2008

Progress so far

29

- Formal Collaborative Agreement nearing completion
- Funding Agreement with DIISR under discussion
- ANDS Executive Director recruitment
 - Job ad closes on June 27
 - KPMG undertaking international search process
 - <http://www.adm.monash.edu.au/human-resources/employment/senior/#execdirdata>

Progress so far

30

- Draft (because it is subject to consultation and the ED may want to modify it once appointed) Interim (because the first full version is due in March 2009) Business Plan has been created
- DIBP was subject of successful consultation with ANDS TWG on 5/6/08
- Further consultation with ANDS Forum planned for July
- Currently planning transition from ANDS Establishment Project to ANDS

Implications for IT Directors - II

33

- Increased amounts of data will need to be stored at institutions
 - ▣ ANDS will be providing reference software stack, consulting support, and possible combined ANDS/ARCS helpdesk
 - ▣ ANDS won't be funding/providing storage, but it will be lobbying the people who can...
- Increased access (both open and controlled) from outside institutions to stored data
 - ▣ ANDS will be working with AAF on policies, and AAF-enabling repositories

Implications for Researchers

34

Responsibilities

- Manage data for life of project
- Meet standards for good practice
- Comply with funder/institutional data
- Work up data for use by others

Benefits

- Greater visibility of research
- Verifiable results
- Improved access to data
- Improved collaboration opportunities
- Preservation/Curation
- Citability of data
- Improved quality of data capture

