The Preservation and Sustainability of Research Data

Dr Markus Buchhorn,
Director, ICT Environments
Australian National University;

Formerly:
Head, ANU Internet Futures
Grid Services Architect, APAC
Grid Services Coordinator, Grangenet

This talk is based in parts on the “AERES” survey and report for APSR with Paul McNamara: www.apsr.edu.au
Research Data

- This is not about publications but primary, derived or simulated data,
  - Which (may) lead to publication
  - Scholarly inputs and outputs

- Why is it different?
  - Data has a very different lifestyle

- Why is it hard?
  - Data has very different, and more complex, problems

- E-Research infrastructure?
  - Transparent and appropriate access to all resources,
  - to enhance research processes and build greater knowledge
We sort of **know** this…

- **A (good) Repository**
  - is the sum of these things, and more…
    - Interfaces and services for management and curation, processes, security, standards, support, etc.
...and we can **architect** things around it...

Repository Federation

- IRP
- Metadata-flows
- "Portal" or Federation interface

AAA Services

- AAA flows

AAA flows

Users
- Computing
- Collaboration
- Visualisation

Access protocols

Queries, Curation

This all applies even with a single repository
...and we can identify the services
We can classify the processes.

Version 5.1
Markus Buchhorn
Let’s look at Application Areas

- Geosciences
  - Minerals, oils and gases, tectonics, Govt, Surveys, Industry
  - Many data sources (spatial and physical) and simulations

- Bioinformatics
  - Genomics, proteomics, …
  - Public datasets, private queries, private annotations

- Chemistry
  - Simulation, need data services mainly

- High Energy Physics
  - Large expensive instruments, projects
  - Massive data, computation and simulation

- Earth Systems Sciences
  - Massive remote sensing data sets, large and complex simulations

- Astronomy
  - Big data, complex reduction process, big simulations, long-term research
Application Areas - 2

- **Financial**
  - Many sources, Stock/Financial exchanges, news, ...
  - Timeliness and also long time scales are both important

- **Music, Arts, Sports**
  - Performance and creation, formal and practice
  - Education focus

- **Linguistics, Musicology**
  - Archives of digitised cultural material
  - Complex analyses

- **Social Science Data**
  - Census, health, surveys, ...
  - Complex data structures, qualitative data

- **Archaeology**
  - Digitised physical materials, spatial and chronological data
Consider just *some* of the issues…

- **Sustainability of data**
  - Data formats, Simplex vs Complex (compound) objects
  - Software (Algorithms, implementations, OS)
  - Versioning (Recalculation, interpretation, validation, derivatives)
  - Underlying infrastructure (hard and soft)

- **Describing data: Metadata**
  - Varied research schemas (1 is nice, but most have zero or five…)
  - Scientific description can be itself contentious…
  - Many types: (Provenance and processing, Preservation, curation and valuation, Subjective metadata, annotations)

- **Rights around data:**
  - Needs Authentication and Authorisation to be working, and to scale
    - Requires *identities* and *roles* to be understood
  - Privacy, Security
  - Ownership - Not always (almost never!) with the researcher
  - Time-varying (Data sourced under old agreements, people die, agreements expire, …)
So why do this anyway?

- Create opportunities
  - For re-analysis, re-use; expected or otherwise

- Solve problems
  - Waste of $$, people and collection effort
  - Loss of irretrievable data
  - Inability to verify research

- Requirements (have to do it)
  - National good, cultural heritage, input to policy
  - Reference materials
    - Atlas, catalogues, …

- Value not just in collection but in accessibility
Is it happening already?

- Data re-use/re-analysis
  - Ever more examples, some very good, some horror stories…
  - Policy conflicts
    - Data must be kept
    - Data must be deleted (ethics; anything involving people)

- But…
  - New culture
    - This data has value outside of my domain, or after my project?
  - New capabilities, provided by the Internet
    - Discovery of who has useful data
    - Accessibility of useful data
  - New data is easier to cope with than old data
    - Introduce new workflows and processes starting now
    - Recover old data as/when needed
  - New (and old) fears by users (see later)
Some of the players: Government and funders

- **Strengths:** Control $$ and Policy,
  - and some data (ABS, BoM, GA, RTA, AADC. …)

- **Weaknesses:**
  - Policy politely suggests publicly-funded data should be well managed and appropriately accessible
    - No *teeth*, no *infrastructure*, no *recognition* if done
  - Funding is *project oriented*, infrastructure is *systemic*?
    - One-off grant for lifetime support?
Opportunities:
- Effective and coordinated policy, with $$ to back it up
- Build coordinated, sustainable infrastructure; skills, expertise
- Increase research effectiveness and leverage of $$ investment

Threats:
- Loss of irretrievable data
- Waste of $$ and effort in collecting the same data
- Insufficient data for policy input
  - Environment, healthcare, education, security, …
- Loss of research effectiveness; other countries are doing this
  - UK, US, Asia (Taiwan, Korea, …)
Another key player: Organisations, Institutions

- Not just Universities
- Employ the staff that collect the data
- Manage the funds acquired by staff
- May have obligations,
  - Probably “own” the data
  - Long-term support (beyond staff tenure)
  - Moral and legal (is research data a ‘record’?)
- But have pressures to leverage their work…
And Users, who are *human*...

- **Fear of missed “nuggets” in their data**
  - Milk it for everything, for ever and ever
- **Fear of missed errors**
  - Probably varies by domain and career-stage
- **Fear inappropriate leaks**
  - Privacy/ethics,
  - first-to-market,
  - relationship to data providers (drug users, fishermen, …)
- **Fear the cost of effort**
  - Takes time (and money) away from what they’re good at
- **Fear lack of recognition**
  - I’ve done it for the national good, how about some accolades?
- **Fear of trusting somebody else’s data**
  - That person, or their repository may have done something wrong
- **Fear unknown custodians/stewards**
  - Can’t do as good a job as my PhD students
Recognition

- “We” require data to be effectively deposited
  - But don’t have anything to back up this requirement

- Implies an effective *place to* deposit
  - Recognition (certification) of repositories
    - How good, and how sustainable? What are the metrics?

- Implies an effective *process of* deposit
  - Recognition of the deposit effort
    - How well is it deposited? 1 star deposit into a 5 star repository?
  - Recognition of the deposit content
    - Depositor gets recognition, somewhat like a paper
      - Which requires a sufficiently good effort, and a citable repository
      - Interesting question of who “owns” the data, and hence accrues recognition

- Who carries out recognition, certification?
  - Domain-specific skills, technology-specific skills
  - Curation, preservation skills
Skills – and the lack of them

• In Australia, and globally
  • Skills around discipline-specific data management
    • Need to learn from researchers what their issues are
  • Skills around generic data management
    • Need to learn library/archive skills for non-publication materials

• Need more “translators”
  • Who seem to come from disciplines and not from IT or IM
    • And more from Humanities/Social Sciences than Physical Sciences
  • Who seem to have given up academic careers
  • How can we create more of them?
    • And the generic research data curators?
Valuation – what to keep?

- Ideal model keeps everything, for ever
- Pragmatism dictates some data deletion
  - Who has, or wants, that responsibility?
- Cost is going down
  - Storage (physical media) is getting cheaper
  - Processes for management are starting to scale
  - Keeping everything is becoming reasonable
  - Keeping it for ever is becoming manageable

- **BUT:** May not be able to manufacture fast enough…
  - *Peak Storage Capability, like Peak Oil?*
Sustainability

- Follow the $$$

- Govt top-slice, or top-up to institution/user
  - Fund fewer people to do more things?
  - Fund the same number to do more with less?
  - Create a whole new funding stream?

- Institutional top-slice, or top-up
  - Same questions.

- Leave it to users/communities
  - Where there’s a will, …
    - But we need to support areas where there isn’t a will yet
Implementation?

- Get users out of data management at some level
  - Scale costs on infrastructures, services and skills that are sufficiently common

- Deal with user fears
  - Some of it needs education, some of it needs trust to be established

- Users provide domain specific skills and domain policies
  - Coordination role within a domain – required!
  - But need technical backing when it crosses some boundary
Who is thinking about this?

- Institutions and partnerships: APSR and other groups
- Govt: DEST, PMSEIC, NCRIS (SII), eResearch-CC, Productivity Commission, ...
- Funders and managers: ARC, NHMRC, AVCC/UA, ...

NCRIS:
- Emerging Australian National Data Service (ANDS)
  - Programs around: Policies, key services, repository management, and research practices
  - Foundation set of key services enabling the creation of a national data commons

Here’s hoping…