



The Preservation and Sustainability of Research Data in Australia

*Dr Markus Buchhorn,
Director, ICT Environments
Australian National University;
Also in www.APSR.edu.au*

*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*

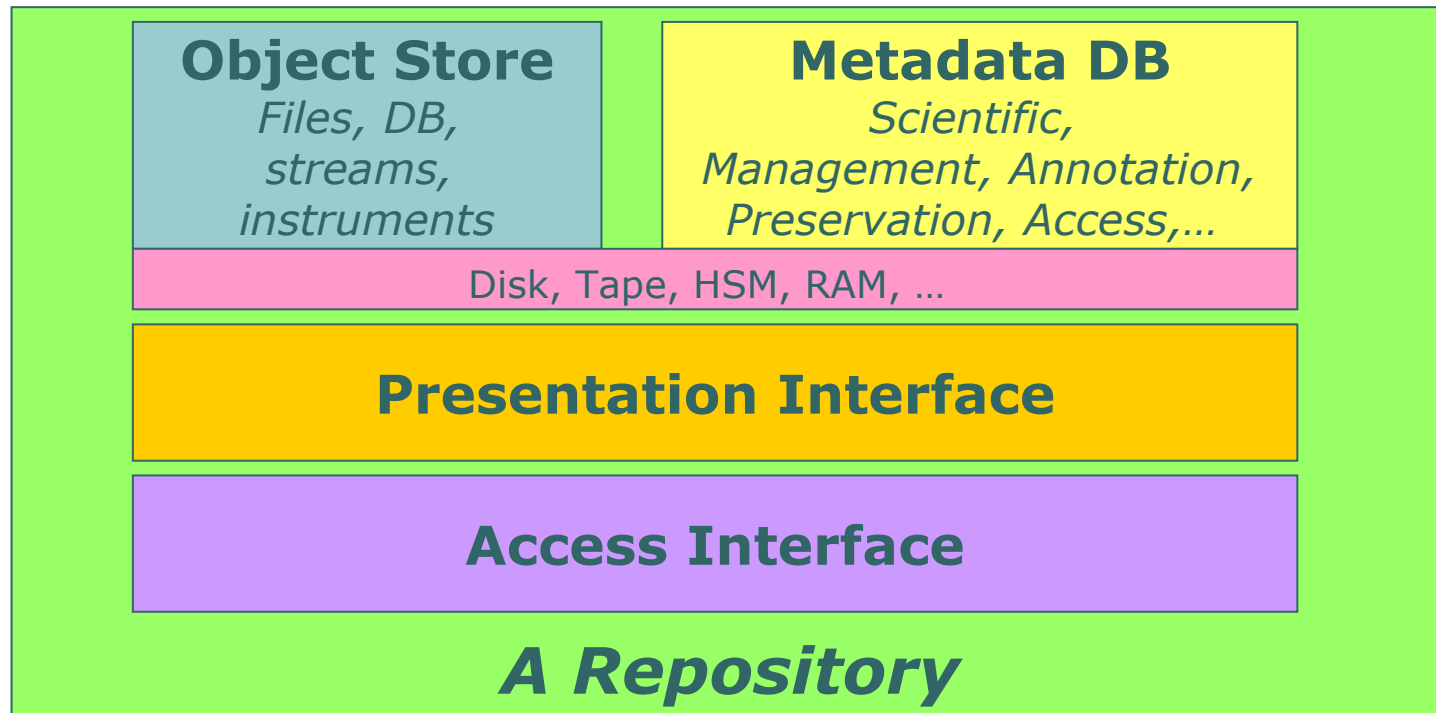




Research Data

- 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
- Just “in Australia?”
 - It’s a global problem!
- E-Research infrastructure?
 - Transparent and appropriate access to all resources,
 - to enhance research processes and build greater knowledge

We sort of **know** this...



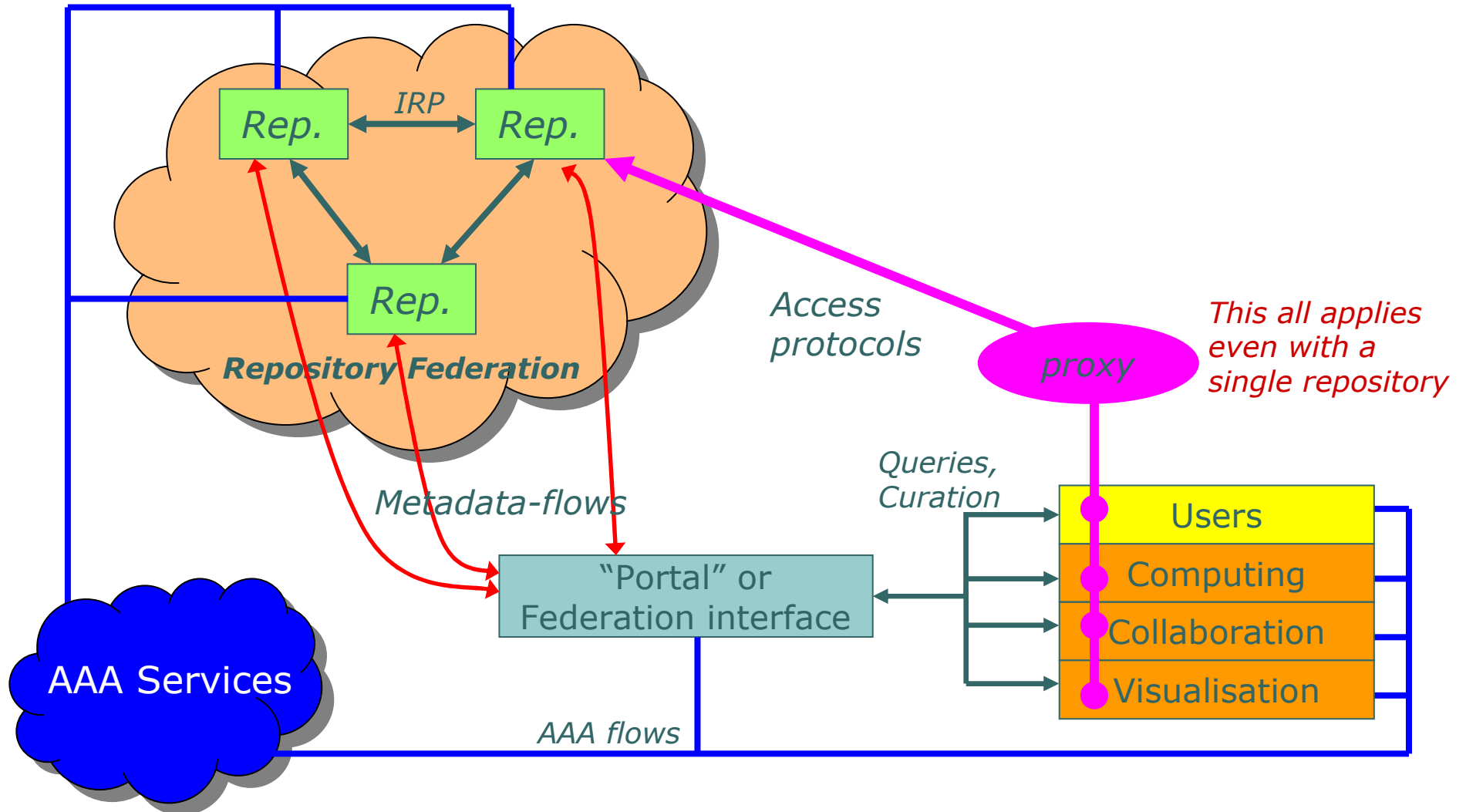
- **A (good) Repository**

- is the sum of these things, and more...

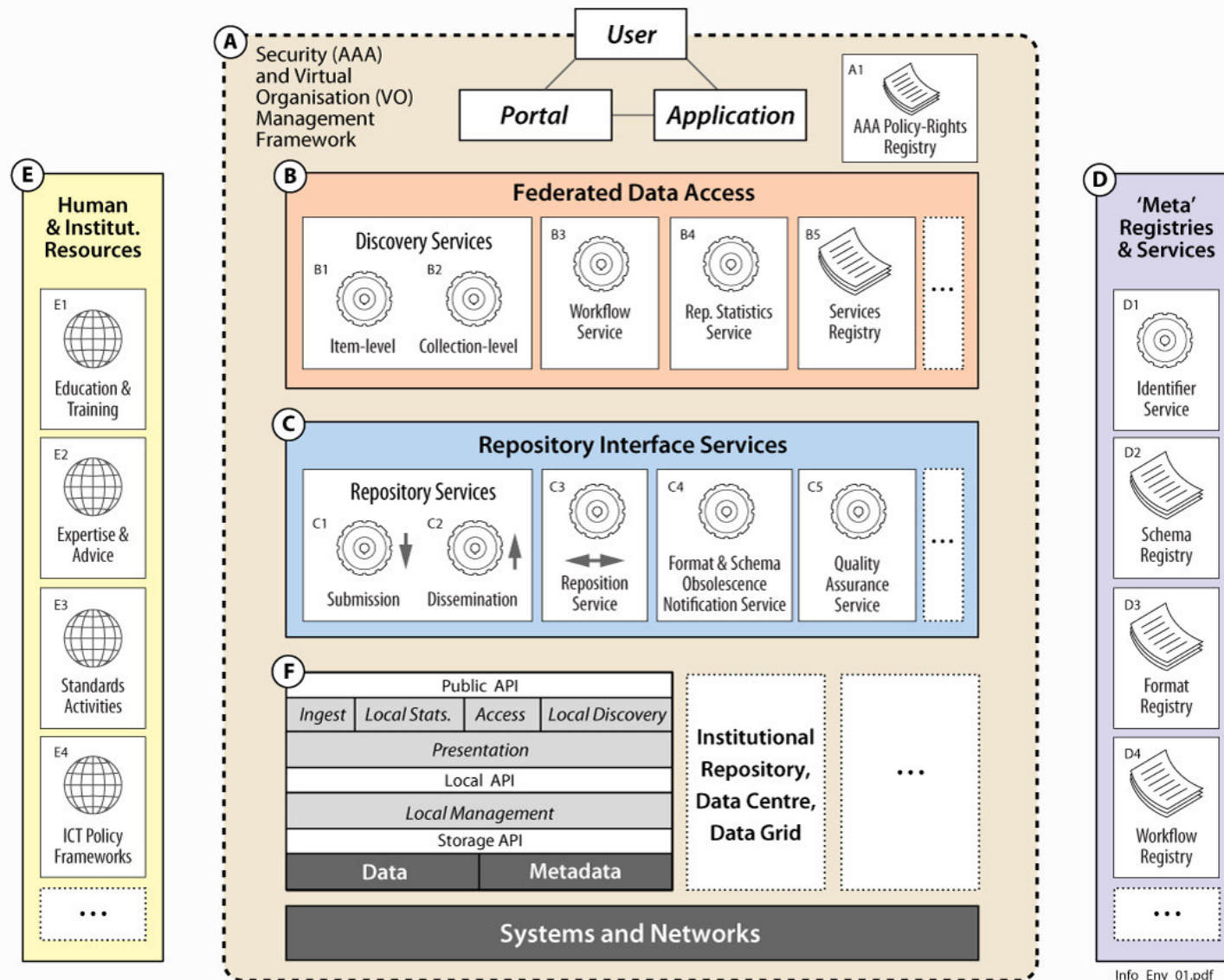
- Interfaces and services for management and curation, processes, security, standards, support, etc.



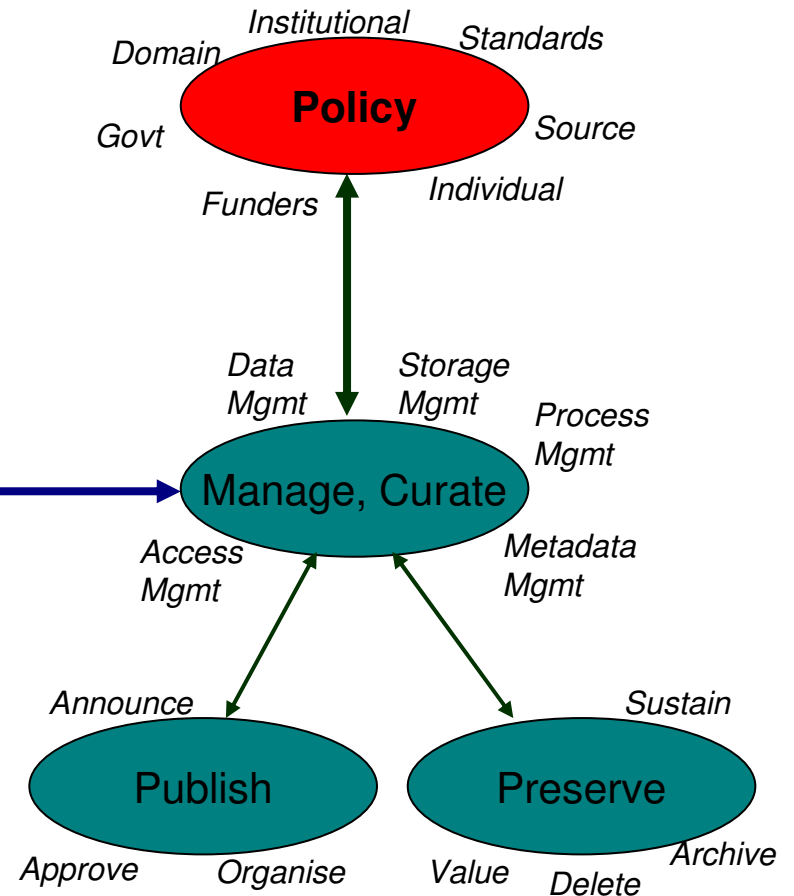
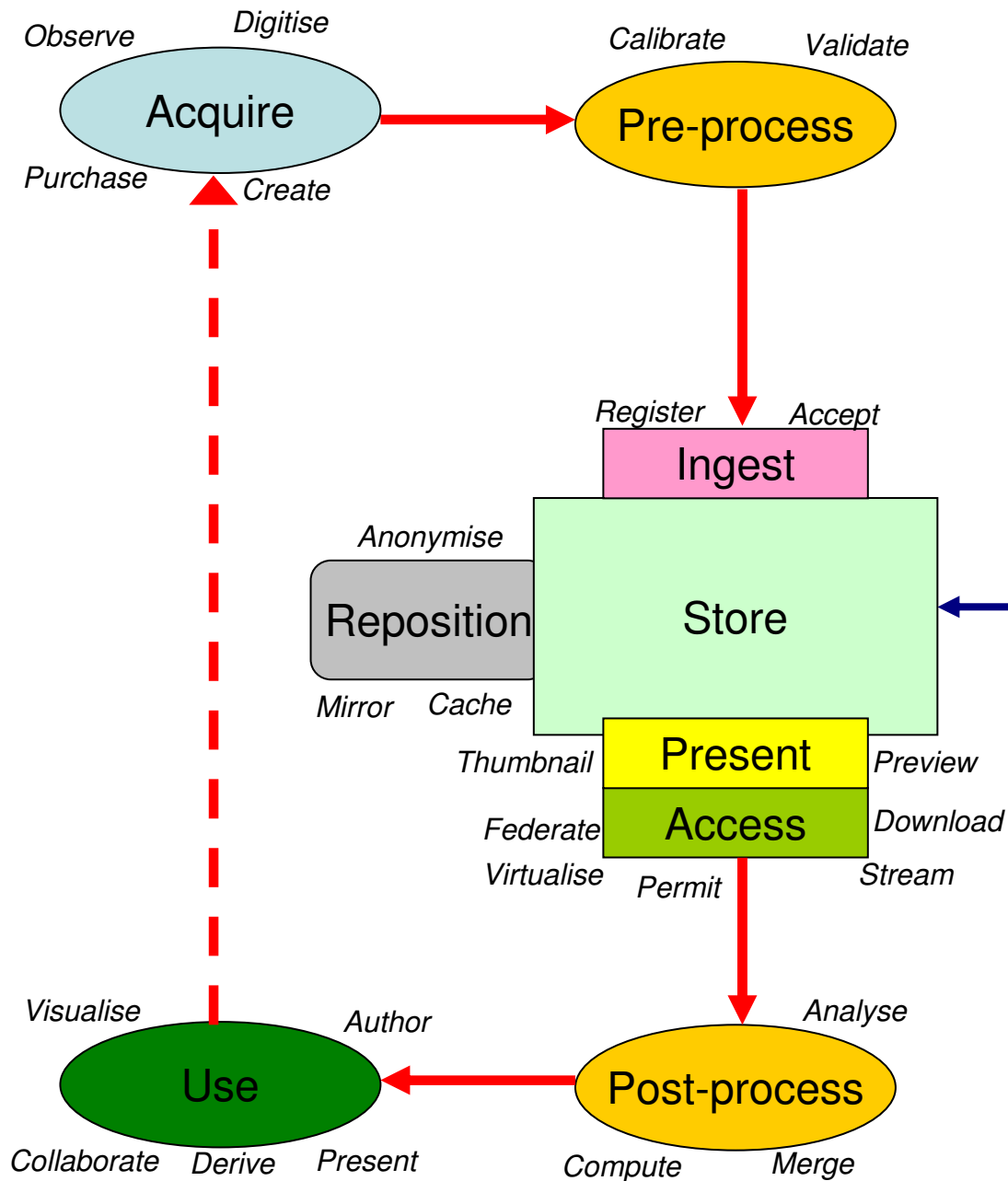
...and we can **architect** things around it...

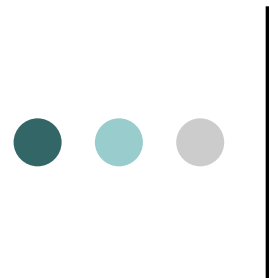


...and we can identify the services



We can classify the processes





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



Application Areas - 2

- Earth Systems Sciences
 - Massive remote sensing data sets, large and complex simulations
- Astronomy
 - Big data, complex reduction process, big simulations, long-term research
- Financial
 - Many sources, SX, FX, news, ...
 - Timeliness and long time scales are important
- Music, Arts, Sports
 - Performance, formal and practice
 - Education focus



Application Areas - 3

- 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* Issues - Longevity

- Sustainability
 - Data formats
 - Descriptions, Compression, lifetimes
 - Simplex vs Complex (compound) objects
 - Software
 - Algorithms, implementations, OS
 - Versioning
 - Recalculation, interpretation, validation, derivatives
- Underlying infrastructure, technologies
 - Storage Facilities
 - Mirroring for protection – policy and technical issues

Geo, Bio, ESS,
Astro, Ling, SS,
Arch, Fin, Mus.

Geo, Bio, ESS,
Astro, Ling, SS,
Arch, Fin, Mus.



Issues- Metadata

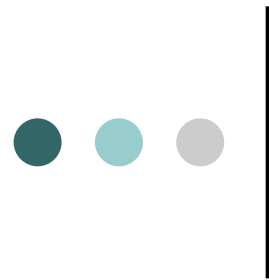
- Varied research schemas
 - 1 is nice, but most have zero or five...
- Baseline DC is almost non-existent..
- Scientific description
 - Itself contentious...
- Provenance and processing
- Preservation, curation and valuation
- Subjective metadata, annotations



Issues - Rights

Geo, Bio, HEP,
ESS, Astro, Ling,
SS, Arch, Fin,
Mus.

- Needs AAA to be working, to scale
 - Authentication, Authorisation and Accounting
 - Requires identities and roles to be understood
- Privacy, Security
 - Personal information leakage
 - Anonymised data, needs to stay usable
- Ownership
 - Not always (almost never!) with the researcher
- Time-varying
 - Data sourced under old agreements
 - Rights vary by status of source
 - people die, agreements expire, ...



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 (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 (and old) fears by users (see later)
 - New data is easier to cope with than old data
 - Introduce new workflows and processes starting now
 - Recover old data as/when needed



Some of the players: Government and funders

- Strengths:

- Control \$\$,
- Control Policy
 - Define requirements, enforceability, and encouragement!
 - Set frameworks for ethics
 - Can of worms in its own right (c'tees getting involved in technical elements; too many c'tees at different layers, contradictory rulings)
- Control some data (ABS, BoM, GA, RTA, AADC. ...)
 - And can be data triggers (tobacco, regulators, ...)



Government and funders

- Weaknesses:
 - Policy politely suggests publically-funded data should be well managed and accessible
 - No teeth
 - No infrastructure to back it up
 - No recognition of good effort
 - Funding is project oriented, infrastructure has to be systemic
 - One-off grant for lifetime support?



Government and funders

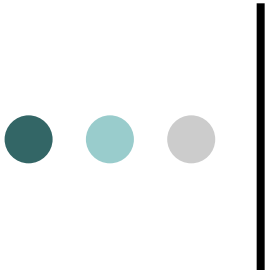
- Opportunities:
 - Effective policy, with \$\$ to back it up
 - Build a coordinated and sustainable infrastructure
 - Build skills, expertise
 - Save money
 - Increase research effectiveness
 - Increase leverage of investment



Government and funders

- 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,
 - especially long-term (beyond staff tenure)
 - Legal and moral (is research data a ‘record’?)
 - Probably “own” the data
- Certainly have opportunities
- Have existing funding models
 - Shuffling between buckets...



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 unknown custodians/stewards
 - Can't do as good a job as my PhD students
- 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



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



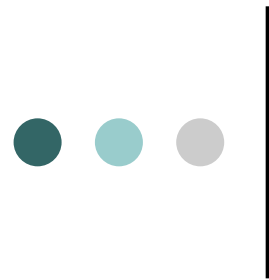
Valuation

- What to keep?
 - Ideal model keeps everything, for ever
 - Pragmatism dictates some data deletion
 - Who has the right to make that decision,
 - and takes on the responsibility
 - Especially if later proven wrong
 - Cost is going down
 - Storage (physical media) is getting cheaper
 - Processes for management are starting to scale
 - Especially for the basic storage/access services
 - Keeping everything is becoming reasonable
 - Keeping it for ever is becoming manageable



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 as well



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
 - E.g. Scalable AAA mechanisms are now coming along nicely
- Users provide domain specific skills and domain policies
 - Coordination role within a domain – required!
 - But need technical backing when it crosses some boundary



Implementation - 2

- All repositories don't need to do everything
 - Some can be more equal than others.
 - By domain, by technology, by fundamental services...
 - As long as the sum of them exceeds the sum of needs
 - Most technical problems can be solved today.
 - Policy is the main hurdle.
- Achieving the goal
 - What are the carrots and sticks that actually work?
 - Who are best placed to wield them?
- Sustaining the goal
 - The answer is money, but what is the question?



Is anybody *thinking* about this?

- Universities and partnerships
 - APSR and other groups
- Federal and State Govt
 - DEST, PMSEIC, NCRIS (SII), eResearch-CC, Productivity Commission, ...
- Funders and managers
 - ARC, NHMRC, AVCC
- Here's hoping...