



Supporting the Data Lifecycle

#DLCsyd09

How to ensure co-development?

Jim Richardson

Relationship Manager for eResearch

Information and Communications Technology, University of Sydney

11 February 2009

<http://www.and.s.org.au/datalifecycle-symposium.html>



Licensed under Creative Commons Attribution-Noncommercial

<http://creativecommons.org/licenses/by-nc/2.5/au/>



Summary

- **Part 1: Overview**

- 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.
- It's 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!

- **Part 2: Materials for discussion**

- How to achieve seamless interplay between data stores for active research, and a Data Commons for preservation, appropriate access and re-use?
- Levels and layers
- Different repositories for different domains
- Questions about researcher requirements
- Meta-questions – how we proceed at and after the symposium



eResearch: background

- **"eResearch" is the application to research practice of advanced information and communications technologies**, including:
 - data storage, management and sharing
 - high-performance computing
 - on-line collaborative tools.
- **eResearch approaches can facilitate research at a range of scales**, from large and inter-disciplinary research projects which would be difficult to organise otherwise, to the “long tail” of smaller projects with a focus on data management or analysis.
- Systematic **management of data via eResearch methodologies allows and encourages re-use** of the data by subsequent research groups or projects (subject to intellectual property and ethical constraints), broadening the social benefits.



The benefit of hindsight 1: the Grid

- **David De Roure: *How [Data] Repositories can avoid the Failings of the Grid***
 - Indianapolis IEEE e-Science 2008 repository keynote
<http://www.semanticgrid.org/presentations/DEROURERepo3.pptx>
 - **If the Grid is successful, why are there three UK JISC uptake projects and a theme in the e-Science Institute?**
 - **Goal: not “heroics” but everyday researchers doing research they couldn’t do before**
 - Ease of use, without taking away from researcher autonomy
 - Understand researchers’ needs by journeying with them
 - Be open-minded about what problems have to be solved.

Dave De Roure's 8 Data Repository points

1. Not just a specialist few doing heroic science with heroic infrastructure – repositories for all!
2. There is new value in data, through new digital artefacts and through metadata e.g. context, provenance, workflows
3. e-Science now focuses on publishing as well as consuming
4. Usage informing recommendation [community intelligence]
5. Researchers work with collections - Object Reuse & Exchange
6. [Technologies that] are easy to use
7. Anything that takes away autonomy will be resisted
8. e-Science is about the intersection of the digital and physical worlds (not 1970s library catalogue interfaces)

How Repositories can avoid Failing like the Grid

1. Understand what the users will need by going on the journey together
2. Be open-minded: are we solving the right problem?
3. Don't create artificial distinctions from Web
4. Beware standards as a barrier to adoption
5. Think cloud, outside the institutional box
6. Think of a new name for repositories!



The benefit of hindsight 2: IRs

- **Dorothea Salo: *Innkeeper at the Roach Motel***
 - <http://minds.wisconsin.edu/handle/1793/22088>
 - On lack of uptake in institutional repositories for publications:
 - **“Trapped by faculty apathy and library uncertainty, institutional repositories face a crossroads: adapt or die.”**
 - **The ‘build it and they will come’ proposition has been decisively proven wrong.”**



The benefit of hindsight 2: IRs (detail)

- **Dorothea Salo: *Innkeeper at the Roach Motel***
 - “Trapped by faculty apathy and library uncertainty, institutional repositories face a crossroads: adapt or die. The “build it and they will come” proposition has been decisively proven wrong. Citation advantages and preservation have not attracted faculty participants, though current-generation software and services offer faculty little else. Academic librarianship has not supported repositories or their managers. Most libraries consistently under-resource and understaff repositories, further worsening the participation gap. Software and services are wildly out of touch with faculty needs and the realities of repository management. These problems are not insoluble, but they demand serious reconsideration of repository missions, goals, and means.”
 - “Institutional repositories have not fulfilled their early promise of increased access to the scholarly journal literature through faculty initiative. ...
 - Thus far, at least in the United States, doubts about the viability of institutional repositories have been kept quiet or denied altogether. As long as libraries and repository managers remain silent about the current deplorable situation, however, no one can rectify it.”
 - Salo quotes Graham Pryor (2007): “[Cultural and organisational barriers prevail in all disciplines, which serve to deter the deposit of research data in repositories, and] an inherent culture of self-sufficiency in the generation and organisation of data militates against what might be viewed as prescriptive intervention by knowledge management professionals”
 - <http://minds.wisconsin.edu/handle/1793/22088> ; <http://www.ijdc.net/index.php/ijdc/article/view/32/35>



Co-development in partnership with researchers

- **Alex Voss** on e-Science Institute *Adoption of e-Research Technologies* theme:
 - “Should e-Infrastructure creators see themselves, not as architects, but as gardeners?”
 - A few early adopters in a research community are not enough: this theme looked at ways of spreading the uptake of e-Research beyond a minority of enthusiasts and into the mainstream of research.
 - ... **'fostering' rather than 'building' e-Infrastructure ...**
 - Technical problems are not usually the main barriers to uptake of e-Infrastructure.
 - **What is vital is that technical people should understand the needs of particular disciplines and the social environment in which the technology will be used.”**
 - *Tending the Garden*, Iain Coleman reporting on Voss October 2008 lecture in NeSC Newsletter issue 65
<http://www.nesc.ac.uk/news/newsletter/December08.pdf>



Part 2: What are we looking at today?

- **Big picture: Can we achieve seamless interplay between linked “collaborative research data stores” for active research, and a Data Commons, or “cohesive collection of research resources”, for preservation, appropriate access and re-use?**
- More detailed requirements, from the researcher participants thinking broadly for their research communities
- Some sample particular questions, and meta-questions about the questions
- From the symposium webpage: “What are the key features of these environments? Are there any natural cluster points along this spectrum? What are the fundamental requirements? What infrastructure models might support these? How do institutional, national, and global interests coalesce? Who is responsible for retention and disposal policies, what should these be, and what are the resource implications of these decisions?”
- See also: Lorcan Dempsey <http://orweblog.oclc.org/archives/001875.html> (on data curation infrastructure)
 - Chris Rusbridge <http://digitalcuration.blogspot.com/2008/07/negative-click-positive-value-research.html>
 - Luis Martinez Uribe <http://oxdrrc.blogspot.com/2008/12/research-data-management-services.html>
 - Randall et al <https://urresearch.rochester.edu/handle/1802/6053> (study of grad student authoring practices)



Categorisation of levels for research data management

<i>Level</i>	<i>Name</i>	<i>Description</i>	<i>Examples</i>
Level 4	Active	Research data structured via custom metadata, with content-specific ingestion, search, presentation, display, conversion and/or “value-adding” tools.	Google Earth; Godiva2 ; Online Corpora ; ... data commons?
Level 3	Repository	Publications and research data, indexed and searchable through standard metadata, for downloading “as is”.	dSpace, Fedora
Level 2	“Collaborative research data store”	Organised storage specifically for research data-sets, with authorised shared access for collaborators. Encouragement towards structure, metadata, and data management plans. Multiple protocols for convenient access from anywhere. Backed up.	Monash LaRDS ; ARCS Data Fabric VerSI data store
Level 1	File store	Unstructured research data kept on a general-purpose file store or server, and backed up. Access for local researcher or research group.	Departmental file server
Level 0	Chaotic	Research data held by individual researchers. Little or no structure or shared access. Backup non-existent or <i>ad hoc</i> .	PC C: drive; USB memory stick

See also Robin Rice’s http://www.disc-uk.org/docs/data_sharing_continuum.pdf

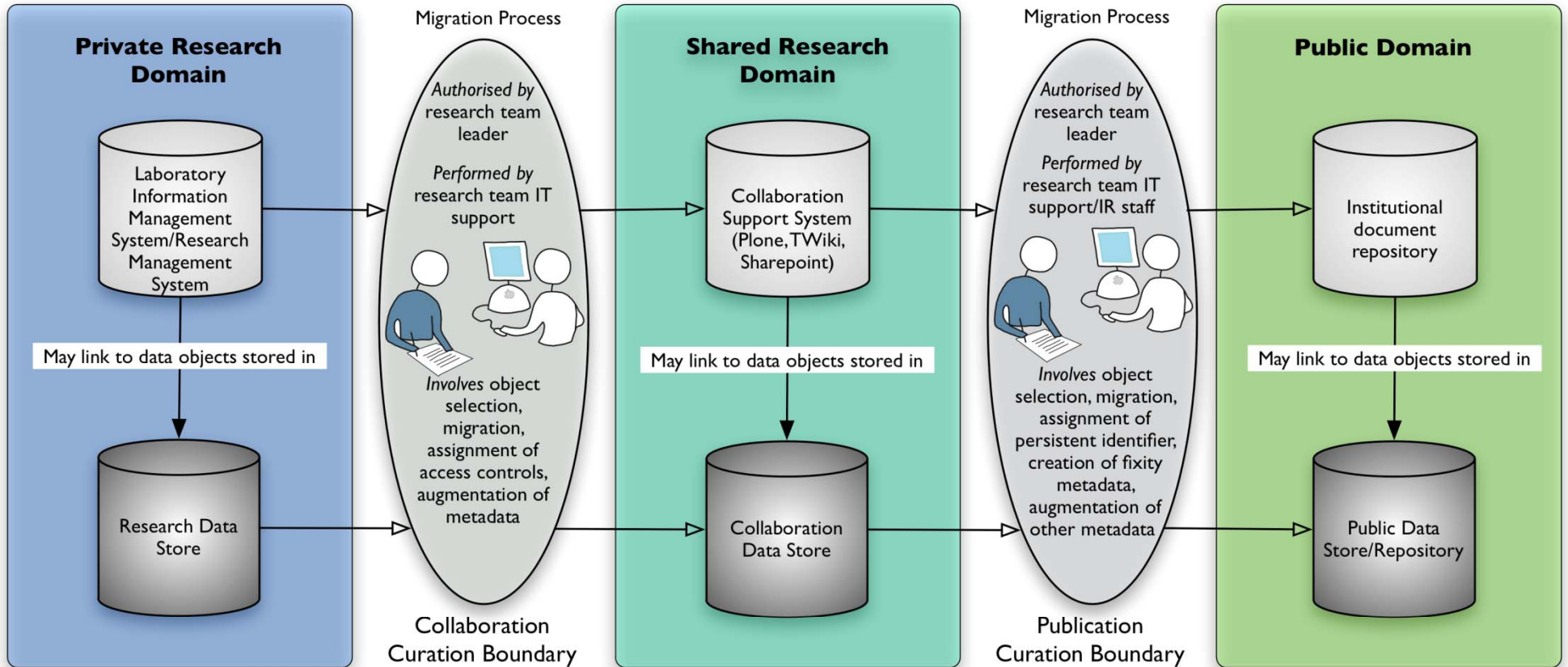


Layers of requirements for Level 2 “Collaborative Research Data Store”

<i>Layer</i>	<i>Name</i>	<i>Description</i>	<i>Notes</i>
Layer 5	Enhancement to research practice	Encouragement, advice and support towards structure, metadata, and explicit data management plans.	How to move up the continuum to re-use as well as use ...
Layer 4	Governance	Institutional data management policy. Funding model. Policy on kind of usage for store (<i>research only?</i>). Constraints e.g. on copyright material. Principles for allocation of space to groups.	See NCRIS Roadmap page 21. Need to agree on ongoing responsibility for funding and provision of data storage.
Layer 3	Presentation	Access methods, e.g. local, web. File type/size limitations if any. Reliability levels. Persistent identifiers. “How to use” support/documentation.	What are the researchers’ own requirements?
Layer 2	System administration	Access permissions for individuals and groups. Quota allocation for file space. Allocation of files to storage tiers (fast/slow disks, tape).	Delegation of control to different researcher and/or support authorities is required.
Layer 1	Infrastructure	Interoperability between nodes in a national grid. AAF. Network bandwidth: include multi-campus, hospitals etc. File systems. SRB. Disks, tape, servers. Backup and disaster recovery. In-house or outsourced.	Capability to provide an institutional node may depend on size of institution.

The different layers will require different discussions and communications with different stakeholders and audiences.

Different Repositories for Different Domains



This domain involves the core research team as they undertake the research, usually within a single institution. Access is often tightly controlled as hypotheses and analyses are developed.

This domain involves researchers outside the core team as they collaborate with colleagues, often across institutions. Access is more open, but not everything is shared.

This domain involves the public sphere (publication in the sense of making public). Access will usually be open to all.



Sample questions on researcher requirements

1. How do research project data stores relate to preservation environments? How can we achieve a seamless data lifecycle, incorporating curation from inception? How does lifecycle differ with discipline?
2. What is the balance of responsibility between different scales of provision (lab/ department/ institution/ state/ discipline-based/ national)? Consider convenience, efficiency, expertise, funding.
3. How can the data store reproduce the convenience of files on the desktop?
4. How to ease or automate the recording of overall and fine-grained metadata in standard forms?
5. What data workflows need to be catered for? Is version or revision control required? Snapshots? Check-in/out? Audit trails? Granular cross references e.g. to database elements?
6. Does the location of data matter, e.g. because of network bandwidth, or for co-location with HPC? How to provide for smaller and more remote research sites?
7. How early in a project can persistent identifiers for data objects be allocated?
8. How can researchers transfer their work when they move? Where do researchers who work across or outside institutions store their data?
9. What data management arrangements are appropriate to support ethics approvals?
10. What is the role of national (e.g. [ACRCR](#) §2; [ARC](#) 1.4.5.3) and institutional policy?
11. What is the role of research project data management plans?
12. What support resources are required?
13. What tools or services can help researchers find and assess existing data for possible re-use?
14. What mechanisms might we adopt to change the scholarly culture?

With thanks to Margaret Henty, Ann Borda, Tim Churches and Clare Sloggett for contributions



Meta-questions

- Is each question: unimportant, easy, hard, unanswerable?
- If easy, what is the consensus answer at the symposium?
- If hard, what process will answer it beyond the symposium?
- If unanswerable, what is the risk?
- **What important questions are missing from the list?**
- **What should be our working approach from here on, after the symposium?**



Acronyms

- AAF Australian Access Federation
- ACRCR Australian Code for the Responsible Conduct of Research
- AeRIC Australian eResearch Infrastructure Council (top PfC body)
- ANDS Australian National Data Service (PfC component)
- ARC Australian Research Council
- ARCS Australian Research Collaboration Service (PfC component)
- AREN Australian Research and Education Network (from AARNet)
- DIISR Department of Innovation, Industry, Science and Research
- HPC high-performance computing
- ICT Information and Communications Technology (concept; University unit)
- ICT Information and Communications Technology (concept; University unit)
- Intersect NSW Institute for Transdisciplinary eResearch Services and Technology
- IR Institutional repository, initially for textual research outputs (ePrints/dSpace/Fedora)
- LaRDS Monash Large Research Data Store
- NCI National Computational Infrastructure (PfC component)
- NCRIS National Collaborative Research Infrastructure Strategy
- NeAT National e-Research Architecture Taskforce (PfC project guidance)
- PfC Platforms for Collaboration (NCRIS-funded)
- SRB Storage Resource Broker
- TFRC Tools and Frameworks for Research Collaboration (University ICT+Library project)