The University Of Queensland AUSTRALIA | CREATE CHANGE

Why we need a Reference Architecture for Research Data

David Abramson

Research Computing Centre

University of Queensland, Brisbane, Australia

Adjunct Professor,

Monash University, Melbourne, Australia

Abramson, D., Betbeder-Matibet, L., Bird, S., Francis, R., Goscinski, W., Soo, A-L., Walsh, C., Wightwick, G and Wilkinson, J. M.

Background

- Advanced data infrastructure
  - high performance
  - scale
  - rich access models
- Implementations
  - Specific domains, or
  - Advanced in an ad-hoc way,
  - Often driven by the urgent need to deliver infrastructure
  - Sometimes driven by what is available in the market
- Without an underpinning model,
  - May only meet a subset of the requirements
  - Build systems that don’t interoperate
  - Serve limited domains or
  - Scale badly.

A Reference Architecture

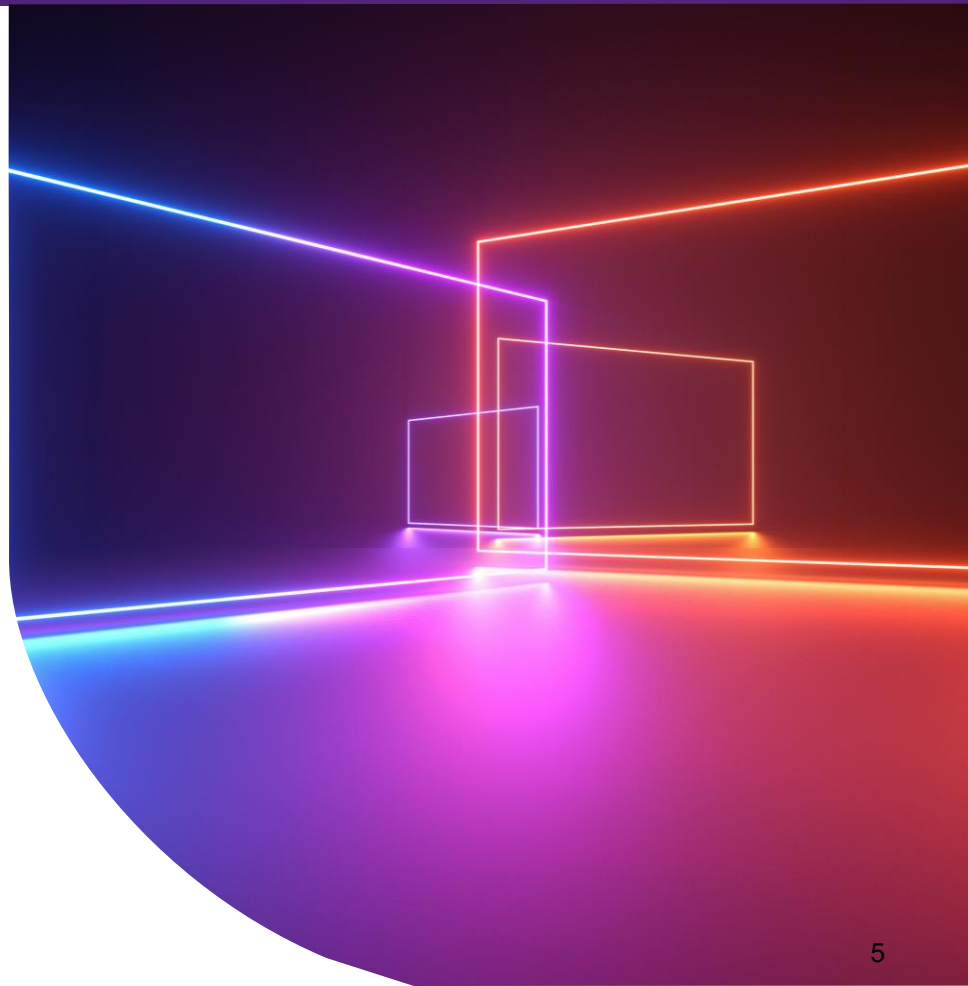
- Our goal is to specify a set of features that should be supported in all implementations, irrespective of technologies and products.
- Not our intention to comment on how information should be used or managed, nor address issues such as data format preservation, curation and data quality.
- It is also not our intention to select implementation technologies or particular methodologies.

Is this really new?

- After all research and development work over a long period, often under the auspices of international activities (e.g. Global Grid Forum, Research Data Alliance) we, might have developed such a specification.
  - Specific platforms
    - Often been domain specific, and tied to implementation technologies
    - Have not provided high level implementation agnostic attributes
    - Don’t necessarily drive implementation choices
  - Generic platforms
    - Not sufficiently abstract to guide other implementations.
    - Often specify products or services that are useful but limit the implementation choices to the ones that have been taken.

RDRA high-level abstract properties

- Resilience
- Discoverability
- Manageability
- Accessibility
- Governable
- Scalability
- Versatility
- Security



Resilience

- Reproducibility is a foundational tenet of good research practice and therefore requires a resilient data store.
  - Codes of Conduct are increasingly pointing to the need for data to be available and open for independent validation.
  - as data citations increase, it is important that a dataset is still available (unmodified) if an external reference has been created.
- Increased expectation that some data will be available for reprocessing
- High level of assurance that data will persist for an agreed retention period according to legislative obligations and academic conventions

Discoverability

- ***A significant number of research organisations have absolutely no idea what data they hold***
  - ***partly because it is held in multiple distributed storage systems***
  - ***mainly because there is a lack of catalogue services.***
- A discoverable system is one where data can be found easily, even if it is not openly available.
- Data often forms the intellectual base of research and are important assets in an organisation.
- Just as companies maintain asset registers for their capital equipment and intellectual property, catalogues of data need to be created and maintained.

Manageability

- Given the increasing volume and scale of research data, data needs to be structured to make it manageable.
- Storage can be allocated in arbitrary units with varying **granularity**.
  - Traditionally, file systems allocate storage in single file units, but these can be clustered into folders or directories when multiple files share similar properties.
  - With research data, it is common to want to store multiple objects, or files, for a single project.
- Data is more manageable if “collections” are the unit of allocation.
  - Collections can contain multiple files (or objects) and have shared meta-data at the collection level.
  - this level of granularity is sufficient and flexible enough to apply to any research domain

Accessibility

- Research is increasingly collaborative and so research data has increasingly complex accessibility requirements across multiple individuals, groups and organisations.
- Such access control features, applied on at least the collection level are not trivial and require maintenance
- Different users might have different access rights.
  - E.g, the lead Chief Investigator (CI) of a project might have the ability to read, write and delete files in a collection, but collaborators may only have the rights to read the files.

Governable

- Research organisations need to build and follow procedures and processes to decide whether data allocations can be made, whether access rights can be changed and ultimately when data can be deleted.
- Roles, such as data custodian, data owner have different rights and responsibilities and these need to be reflected.
  - For example, a data owner may be allowed to change the access rights to a collection, but a data custodian may be the only person who can authorise deletion.
  - It is important that the governance function is integrated into the research data storage architecture from the beginning rather than being an afterthought, to enable traceability, auditability and reporting.

Scalability

- Small numbers of large-scale collections through to large numbers of small-scale collections.
- Storage infrastructures will often start out small but grow over time.
  - Generate data at higher resolutions
  - The number of new projects and researchers growing, and as more traditional research areas become digitised, so does the number of collections.
- Data stores can grow to many peta-bytes
  - At this scale it is usually not cost effective to store data using a single technology.
  - Almost all real-world research stores of significant scale, multiple tiers are used to keep active data on faster, but more expensive, technology platforms.
  - Less active, or archived data can be held in slower, cheaper and more energy efficient technologies.

Versatility

- Research pipelines are extremely heterogenous, and no single processing platform will suffice to support all domains.
  - E.g. hypersonic engineers might make extensive use of high-performance supercomputers,
  - Humanities researchers may choose cloud platforms or even desktops.
- Platforms have different access requirements:
  - Supercomputers usually need high performance parallel file systems
  - Cloud platforms either use POSIX or object-based storage systems.
  - Mobile systems also use cloud-based synch-and-share platforms.
- Must deliver data using a variety of protocols and techniques
  - minimise efficiencyunnecessary data movement and
  - maximises storage

Security

- In research, collaboration is the norm, unlike enterprise data
- Avoid inadvertent leakage of information that could compromise ethical or privacy obligations, primary investigator privileges or impede accepted academic conventions of independent validation.
- Security levels vary across different organisations, domains and projects
  - From completely open to completely closed.
- Several security levels have been proposed for research data
  - open, protected, sensitive, etc)
  - these must be enforced at the most basic levels of an architecture.
- Security classifications must be enforced through protocols and access pathways, often at an individual user level and by the actual storage platforms.

Why not just use existing enterprise systems?

Feature	Enterprise	Research
Resilience	Mostly focused on enterprise needs	Same as enterprise but also covers user errors.
Discoverability	Mostly inward	From inward to global
Manageability	Warehouses, Databases and Files	Collections, Repositories.
Accessibility	Mostly role based	Network based
Governable	Enterprise controls	Research controls, research ethics.
Scalability	Relatively small, database focused	Exponential growth, collection focus
Versatility	Relatively few protocols	Large number of protocols and pathways
Security	Enterprise focussed on restricting access	Collaborator focussed with demanding edge cases, multiple categories of access, etc

Conclusions

- Identified 8 high level abstract properties for research data systems
- Many real-world systems conform, but some don't
- Have developed a capability maturity model
- Also developed a high-level functional specification
- Goal is to allow implementers to measure their systems against these high-level criteria
  - AeRO Workshop at SCAsia 2024 where the RDRA will be tested against products and systems

AeRO
Australasian eResearch Organisations

SCA 2024
Supercomputing Asia
Gathering the Best of HPC in Asia

Thank you and Questions