FAIR Simple Scalable Static Research Data Repository Demonstrator

***Peter Sefton*[1]**

Peter Sefton[1], Michael Lynch [2]

[1]University of Technology Sydney, Sydney, Australia, peter.sefton@uts.edu.au
[2] University of Technology Sydney, Sydney, Australia

BACKGROUND

The University of Technology Sydney has had a centralized Research Data Management service since 2014, using the open source ReDBox application to manage data across the research data process [1] [2]. The original data repository component was a bespoke system using a Fedora [3] version 3 back-end, which was replaced by a repository is based on an emerging standard known as the Oxford Common File Layout (OCFL)[4] – with other repository functions (metadata entry, ingest workflows and discovery being performed by a set of discrete services).

The UTS Research Data Repository uses standards-based metadata using linked-data principles, and the widely-used Schema.org vocabulary, as described in the DataCrate specification. Note that DataCrate has now been folded into an international standardization effort known as RO-Crate (Research Object Crate) – we are submitting a proposal to launch a draft of the RO-Crate spec at this conference as well.

We have also been awarded a grant under the ARDC Data and Services Discovery Activities (GFA-182 : D&S/IR11) to demonstrate the scalability of the OCFL static-file approach and to show how access control can be achieved at scale via the use of group-based access licenses with access control via standard web-server infrastructure (Ngnix).

A NEW RESEARCH DATA REPOSITORY

We will present the UTS Research Data Repository system which uses OCFL for use as a general-purpose data repository, for both open and sensitive data with a discovery portal. OCFL is chosen because it is based on established technology and can be used at computing facilities and on shared infrastructure without requiring server-based repository software or expensive and slow migration of large data collections via APIs.

The OCFL website summarises its benefits:

> This Oxford Common File Layout (OCFL) specification describes an application-independent approach to the storage of digital information in a structured, transparent, and predictable manner. It is designed to promote long-term object management best practices within digital repositories.
> Specifically, the benefits of the OCFL include:
>
> - **Completeness**, so that a repository can be rebuilt from the files it stores
> - **Parsability**, both by humans and machines, to ensure content can be understood in the absence of original software
> - **Robustness** against errors, corruption, and migration between storage technologies
> - **Versioning**, so repositories can make changes to objects allowing their history to persist
> - **Storage diversity**, to ensure content can be stored on diverse storage infrastructures including conventional filesystems and cloud object stores

Source: https://ocfl.io/

We will demonstrate:

- specific datasets from the UTS data repository from a wide variety of disciplines including microbiology, history, computer science & speleology all with DataCrate/RO-Crate metadata.
- varying scale from single collections to an entire university research data repository, building on DataCrate/ RO-Crate for describing and packaging data, and
- the scalability of our approach by automatically generating a large number of plausibly-linked simulated test datasets and contextual entities (people, organizations, equipment, software describing data provenance) with group-based access permissions and demonstrate how a search portal can be used to ensure **Findability** and appropriate **Access** for the sensitive data by using an automated test suite to check the visibility of objects in a portal.
- How a high-performance web server (nginx) can be configured to efficiently serve versioned access, controlled OCFL datasets without having to use a heavyweight repository application.
- How this approach can be used in data-capture context where machine-produced data can be streamed to a simple static repository and then indexed for discovery and analysis into appropriate text retrieval and database software.
- We will also (4) demonstrate how individual data collections can be indexed in detail to produce collection-level discovery services, using two projects that were funded by the ANDS Major Open Data Collections: Farms to Freeways[i] and Dharmae[ii] (UTS).

REFERENCES

[1] L. Wheeler, S. Wise, and P. Sefton, "End-to-End Research Data Management for the Responsible Conduct of Research at the University of Technology Sydney," presented at the Asia Pacific Research Integrity 2018, Taipei, 2018 <https://eresearch.uts.edu.au/2018/07/04/APRI_2018_provisioner.htm>.

[2] G. Kennedy and P. Sefton, "Open Repositories 2018 Presentation: ReDBox 2.0 / Provisioner - ptsefton.com," presented at the Open Repositories 2018, Bozeman Montana, 2018 <http://ptsefton.com/2018/07/06/RedBoX-Provisioner-OR2018.htm>.

[3] Lagoze, Payette, Shin, and Wilper, "Fedora: an architecture for complex objects and their relationships," *Int. J. Digit. Libr.*, vol. 6, no. 2, pp. 124–138, Apr. 2006 <http://dx.doi.org/10.1007/s00799-005-0130-3>.

[4] A. Hankinson, N. Jeffries, R. Metz, J. Morley, S. Warner, and A. Woods, "Oxford Common File Layout Specification 0.2" <https://ocfl.io/0.2/spec/>.

[i] http://omeka.uws.edu.au/farmstofreeways/
[ii] https://dharmae.research.uts.edu.au/