Workflows, processes and technical solutions for seeding the research data commons

Craig Milne*, Ellen Thompson, Yi Sun, Lance De Vine

Queensland University of Technology, Brisbane, Australia

** corresponding author – c2.milne@qut.edu.au*

Background

Queensland University of Technology (QUT) has recently completed an Australian National Data Service (ANDS) funded 'Seeding the Commons Project'. The project employed two Research Data Librarians and was provided with technical support from the High Performance Computing and Research Support Specialists.

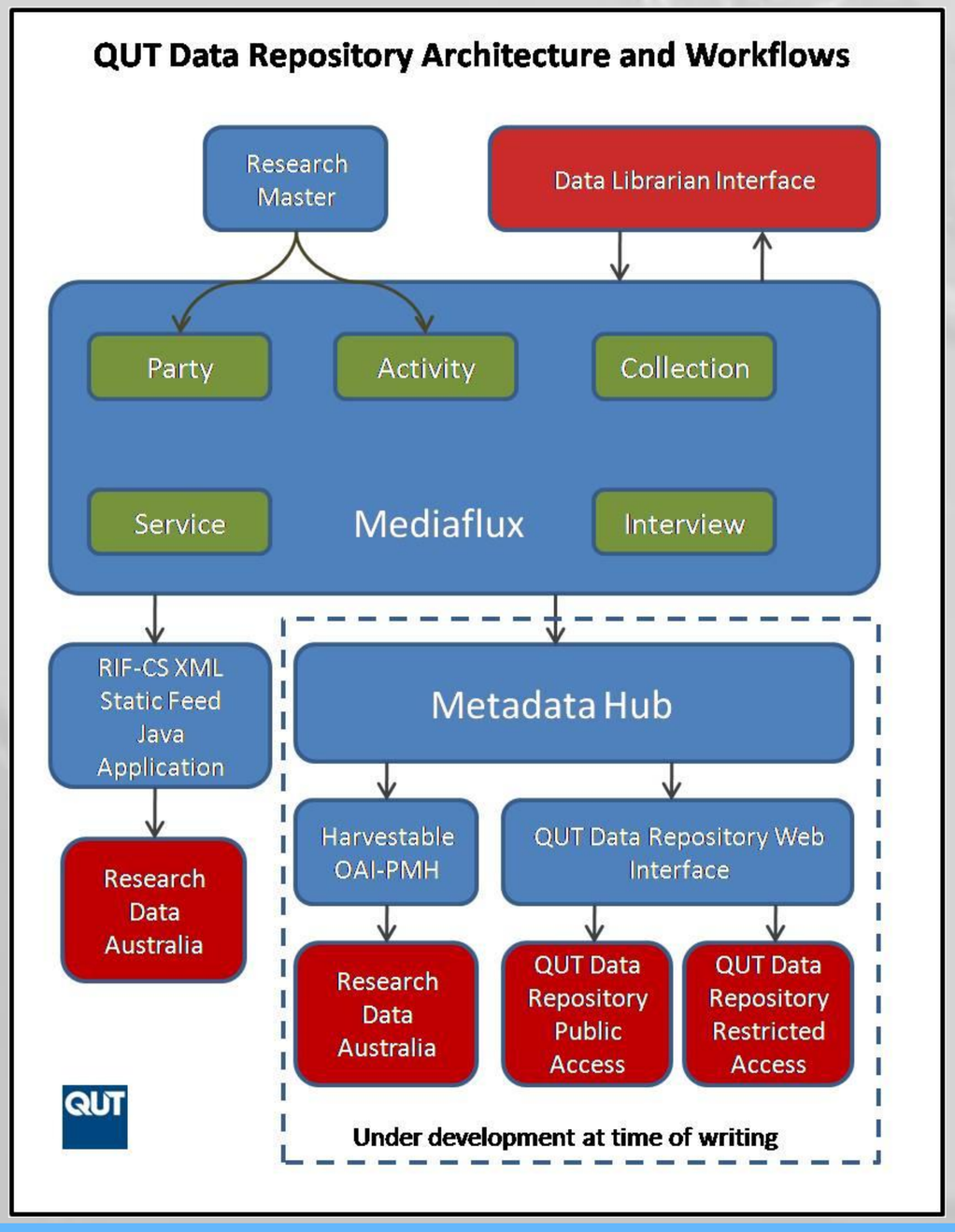
The project first identified QUT's Category 1 (ARC / NHMRC) research activities and then metadata for the datasets was obtained by data interview and from existing university records. This metadata was stored in the QUT Research Data Repository through a combination of data ingest and data entry via a specially developed repository interface. Datasets which were suitable for display in Research Data Australia were provided to the Australian Research Data Commons in RIF-CS format.

Data Repository Architecture and Workflow

The QUT Data Repository is built on the XML Object Database Mediaflux. The database was created to contain all of the objects required for RIF-CS (Collection, Party, Activity, Service) along with an additional 'Interview' object. The data repository used a local schema which provided the ability to store metadata at a more granular level than RIF-CS. The local schema was mapped to RIF-CS to create a feed for Research Data Australia.

An interface to the Data Repository was built by the High Performance Computing and Research Support (HPC & RS) Specialists in collaboration with the Research Data Librarians. The Research Support Specialists ingested existing metadata for Party and Activity objects from the Research Master system and the Research Data Librarians obtained metadata for the Collection objects by undertaking 'data interviews' with researchers.

Future developments will include the implementation of an OAI-PMH compliant feed to Research Data Australia and a QUT Research Data Repository web interface using an implementation of the ANDS funded "Metadata Hub Exchange" project.



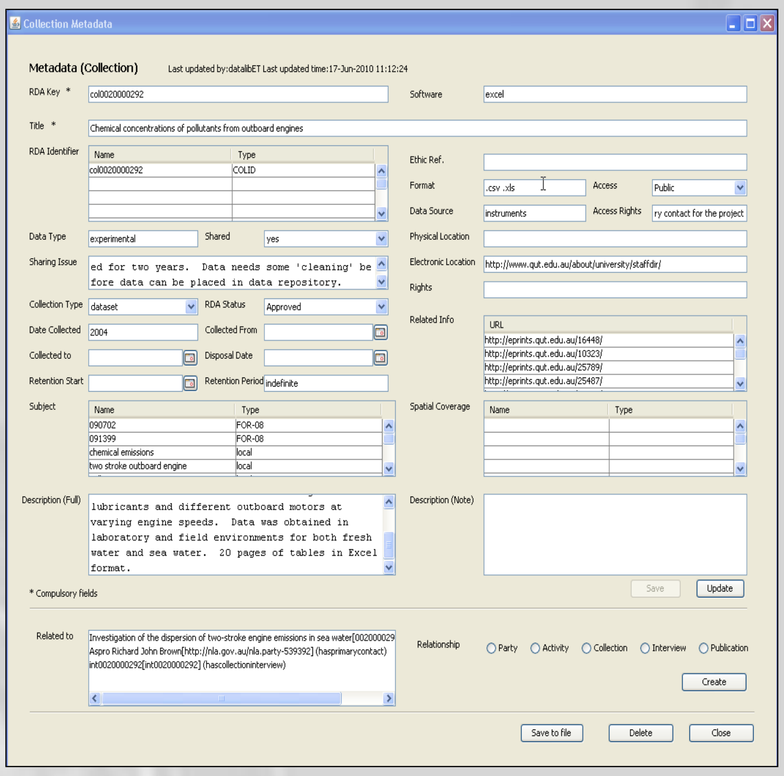
Data Interview Process

Research Data Librarians undertook 195 interviews with researchers in relation to 424 research activities discovering 492 datasets. The basic interview workflow was:

1. Identify research activity using Research Master information
2. Contact the Chief Investigator by telephone, explain the project, book a data interview
3. Carry out some background research on the research activity to help inform the data interview
4. Carry out data interview, verify existing metadata and record any metadata for data on the data interview template
5. Complete any follow up research after data interview
6. Record metadata in data repository (a temporary solution using spreadsheets was developed as the interface and repository were under development while data interviews were being undertaken).

Data Repository Interface

The data repository interface for use by the Research Data Librarians enables entry of metadata at a granular level and stores metadata which can be mapped to RIF-CS or be stored for local use. The interface also enables the creation of relationships between repository objects. The fields for the entry forms were selected based on the RIF-CS Schema and local requirements.



Mapping QUT metadata to RIF-CS

Mapping the QUT repository data to RIF-CS was a straight forward process. A user manual for the interface was written which mapped each element to a RIF-CS element or indicated that it was a QUT only element. The mapping was based on the ANDS Content Providers Guides. As data in the interface was much more granular than RIF-CS some of the elements in the interface were combined into one RIF-CS element for example several elements were combined to create the *Description (Full)* element in RIF-CS. An application using the Java API's and libraries on the ANDS website was created to produce a RIF-CS XML file which could be uploaded to Research Data Australia. The application creates a RIF-CS file based on 'Collection' objects within the data repository which have been approved for Research Data Australia. It uses the relationships established with Party and Activity objects to include this information as part of the XML file.

Other Project Outcomes

Significant project outcomes included:

- Providing an initial feed of 12 QUT data collections and associated records to Research Data Australia
- Development of the QUT Data Repository and initial Research Data Librarian Interface
- Raising awareness of good data management practices and data sharing at QUT
- Identification of barriers to sharing research data at QUT which included ethics, contracts, data condition, publication cycle, data quality, competitive advantage
- Identification of issues to be addressed in progressing a Research Data Management Service including workflows for identifying and capturing research data, more integration of systems to automatically obtain metadata, stronger emphasis and support for research data management.

This project is supported by the Australian National Data Service (ANDS). ANDS is supported by the Australian Government through the National Collaborative Research Infrastructure Strategy Program and the Education Investment Fund (EIF) Super Science Initiative.