Metadata Aggregation – A Critical Component of Research Infrastructure for the Future

Metadata Aggregation – A Critical Component of Research Infrastructure for the Future

Malcolm Wolski[1], Joe Young[2], Joanne Morris[3], Lance DeVine[4], Robyn Rebollo[5]
[1]Griffith University, Brisbane, Australia, m.wolski@griffith.edu.au
[2]Queensland University of Technology, Brisbane, Australia, j.young@qut.edu.au
[3]Griffith University, Brisbane, Australia, j.morris@griffith.edu.au
[4]Queensland University of Technology, Brisbane, Australia, l.devine@qut.edu.au
[5]Griffith University, Brisbane, Australia, r.rebollo@griffith.edu.au

INTRODUCTION

The development of research data management infrastructure and services and making research data more discoverable and accessible to the research community is a key priority at the national, state and individual university level. This presentation will discuss and reflect upon a collaborative project between Griffith University and the Queensland University of Technology to commission a Metadata Hub or Metadata Aggregation service based upon open source software components. It will describe the role that metadata aggregation services play in modern research infrastructure and argue that this role is a critical one.

BACKGROUND

It is widely recognised that research has moved into a data-centric era where the generation, collection, federation, analysis, mining and visualisation of data sets is critical and that research data is much more than just a by-product of research – research data is now a valuable product of research. Thus the storage, processing, transmission, curation, discoverability, sharing and re-use of research datasets (which are increasing in size and number) are prominent issues today and that the appropriate sharing of data for re-use to increase research efficiency and speed the research development cycle is critical. Making research data available increases the verifiability of research, and there is evidence it increases the impact and citation count of related published journal articles and conference papers.

In this new research environment it is critical that research activity be exposed at the University level in a managed way that creates a rich discovery environment. Some of the drivers for this are:

- it enables data to be discovered and explored in the context of the research group, the institute, the collection, the research project;
- it improves the economics of doing research by locating and re-using existing data;
- it increases productivity and turnaround time by leveraging existing research;
- it provides environments for researchers to connect to other researchers and projects to facilitate collaborate, and
- it lifts the research profile of the researcher, the research group and the University.

A metadata aggregator will be a core component of the enterprise infrastructure in this new discovery environment. Griffith University and Queensland University of Technology developed such an aggregator through a project funded by the Australian National Data Service (ANDS). The product will be available as an open source solution.

This presentation will also discuss the project's implementation of a metadata aggregator from a University perspective. It will cover why this has become a core piece of infrastructure, explain where it sits in the enterprise architecture, present an overview of the product selected (Vitro), discuss knowledge gained and the lessons learnt and explain how other universities can take advantage of the open source solution.

UNIVERSITY RESEARCH DATA MANAGEMENT

Griffith University and the Queensland University of Technology, both have major initiatives underway to improve the preservation and re-use of research data and also to develop the skills of the University community in relation to improving data management used for research. Both Universities already have research collections held in centrally managed digital repositories. However, in a university environment with an active, broad research program, it is a fact

that research data collections will reside in a range of different repositories (e.g., specialised discipline specific repositories for stem cell research, art images). Hence, universities will need to generate and collate a consistent metadata feed in order to populate Research Data Australia or other local discovery environments. From a university perspective, the question arises should individual feeds to these environments be maintained centrally or via a range of individual research groups? This presentation argues that a centrally managed solution using a metadata aggregator is the appropriate approach because it will draw data from a number of research repositories and enterprise systems (such as Research Administration Systems and HR systems) to provide a centrally managed service (i.e. a single feed from each University). This will ensure that metadata from university repositories are established and maintained, ensuring that university, national (ANDS) and other discovery tools have the most up-to-date data, in a consistent, reliable manner. It will also provide a single university contact point for national services (e.g., ANDS) regarding national or international metadata stores (e.g., Research Data Australia) and an audit trail back to source data.

The project described in this paper commissioned a metadata aggregator solution that is seen as the first phase of an ongoing initiative. The goal of this project was to produce an open source solution that can be freely used by Australian universities to populate Research Data Australia (RDA) and open the door to a myriad of the possible services such as feeding into other discovery environments (e.g. internal to the University, other federated discovery sites) and exploring research relationships visually etc.

The project outcomes were targeted to:

- Enable QUT and Griffith to populate and update the RDA on an ongoing basis using a system that would accommodate changes in university collections (e.g. addition of new collections, modifications to metadata within collections) through automated data feeds
- Capture relevant metadata from research repositories and corporate systems through standard feeds
- Develop a suitable system architecture that will promote re-use of research data through a standard approach
- Identify and address legal, ethical, security or other constraints over the lifecycle of data in the solution developed
- Identify and utilize current technologies and approaches (e.g. Persistent Identifier Service, Semantic Web)
- Incorporate current good practices and standards (e.g. Codes of Research Practice, ISO2146, standard schemas and vocabularies)
- Leverage the national platforms being commissioned by ARCS, ANDS and the AAF
- Capture usage data for later analysis (e.g. assist with decisions about retention and storage).

As the project involved building a system from scratch with little knowledge of requirements or suitable technologies it involved developing a requirements specification from scratch in consultation with ANDS (e.g., technical requirements for metadata synchronisation with RDA), the university Office of Research sections (e.g., sources of metadata for people and projects), the Library and corporate systems/ infrastructure support sections (e.g., sources of data and infrastructure facilities).

Preliminary work on the project resulted in the identification of a number of possible open source technical solutions, tools and architectures. As the result of evaluating and trialing a shortlist, Vitro was selected along with some other tools such as Kepler for a workflow engine, and OAI-CAT for providing OAI-PMH feeds. The selection of Vitro also resulted in collaboration with Melbourne University and Cornell University in the development of the solution and provided a level of assurance as to the future of the product.

The project has also resulted in another significant outcome - the development of a standard metadata schema based upon RIF-CS and other appropriate international standards and the identification of data sources to feed the aggregator. The development of a standard ontology is a core requirement of any aggregator solution and this had to be developed as a national standard with ANDS and other key players nationally.

Finally this presentation will show how other research institutions can obtain this open source product. Guidance on project managing the deployment of the system will also be addressed.