Connecting the dots to unify research data and metadata

Salim Taleb¹, Peter Hicks²
¹Curtin University, Perth, Australia, s.taleb@curtin.edu.au; ²Curtin University, Perth, Australia, p.hicks@curtin.edu.au

INTRODUCTION

Curtin University is developing a research data management system that aims to provide researchers the necessary tools to plan, create, store, access, share, describe, archive and curate their research data[1]. These tools include a data management planning tool, a data management tool, data stores and metadata management tool. The objective is to capture research data for the majority of activities undertaken at the University, linked to high quality metadata to maximise reusability. Most institutions within Australia have (or are developing) a similar set of tools. Almost all institutions are developing data management planning tools or guides for researchers. Some are in document formats [2] and others are in a web friendly formats[3]. The information captured in these data management planning tools are designed for human consumption and do not necessarily play any further active role in the life of the research project. Data management tools are usually connected to a set of underlying storage solutions. Data storage solutions vary in purpose, design and implementation. Existing solutions include large federated data stores such as LaRDS[4] and discipline specific data stores such as ASSDA [5]. These solutions provide default data storage spaces that does not necessarily take researcher’s specific data requirements into consideration. Additionally, the data management solutions might not encompass a metadata management solution that aggregates object/collection level metadata and associated information. Some institutions incorporate the metadata management solution into their insitutional repositories and others adopted standalone solutions, such as Vitro, for managing the metadata[6]. These tools are not necessarily connected, which may lead to a disjointed data management experience. Curtin University’s approach is unique in that the tools are developed in a connected manner to provide researchers a convergent approach to data management.

DEFINING THE DOTS

The components of the research data management system are built on the core principle of information reusability. Each component will create, manage and propagate information that can be utilised or reused by other components of the system. This interconnection provides value that is higher than that of the individual tools. Additionally, the data management system will connect to existing institutional systems to leverage relevant information. The three major components forming the data management system are the data management planning tool (DMP), the data management layer, notionally labelled Research Data Portal (RDP) and a metadata management system, named the Metadata Hub. The DMP will assist researchers determine their requirements for data capture, storage, access, reuse, ownership, archival and preservation. The requirements will assist in recommending appropriate data storage solutions, customize the information management structure for the research data and provide the institution a basis for storage capacity planning.

The RDP is a data management layer overseeing a number of data storage solutions. To researchers, the RDP will be the primary portal through which researchers can view their research data. Collection level metadata is important for research data discoverability through national initiatives such as the Australian Data National Service (ANDS).

The Metadata Hub will be an aggregator of information about research data (from RDP), the researchers and their respective projects (from other institutional systems). The Metadata Hub will, among other functions, transform the collated metadata to RIF-CS, an ANDS friendly metadata format, and publish it to the Australia Research Data Commons (ARDC). The components of the research data management system are depicted in Figure 1.

![Figure 1: Conceptual design](image)

LINKING

The DMP tool is developed as an online web form. It will assist researchers identify their data requirements through and beyond a project lifecycle. It will also gather contextual information that will be used to define the collection level metadata. The DMP tool will draw relevant activity information from the research management system for associated grant proposal or project. Researchers will also be able to create plans for activities that are not captured by the research management system, such as creative works. The information collected in the DMP will be the utilized by the researcher as an attachment to the grant proposal, if applicable, and it will also be used by RDP for data storage customization.

Melbourne, Australia 5th eResearch Australasia Conference 6-10 Nov 2011
The RDP will primarily focus on the management of data across a number of existing storage locations. The type and nature of storage for a research project might be determined by factors gathered in the DMP including confidentiality, ethical, funding and data requirements. The RDP contains an engine that utilizes the DMP information to recommend or create default connection to storage solutions. Once the connection is established, access and security policies are applied, default files and folders organization structure is applied using Curtin’s Common Vocabulary (institutional taxonomy), milestones transform to notifications and collection level descriptions gets populated. Additionally, it contains a workflow management layer that controls collection level metadata check and object level metadata extraction. Figure 2 depicts the connection between DMP and the RDP. The metadata hub will receive collection level descriptions after a successful metadata check in the RDP. These will be linked with descriptions of associated elements, such as the project or the researchers, from other institutional systems to provide relevant contextual information. The hub is distinct from RDP to enable other sources of collection description in the future. This decoupling will allow the metadata hub to be the central institutional repository of all collections descriptions. These descriptions will be harvested by ARDC for external discoverability.

Researchers will be directed to the DMP during the grant proposal process, via the research management system. Completed data management plans will be available as downloadable files to researchers, who will be encouraged to include the plans in grant applications. Upon approval or acceptance of a project, the DMP information will be used to create default RDP with storage solutions. Researchers will utilise the RDP to manage the data and metadata. The collection level descriptions will remain editable in the RDP during the lifetime of the research project metadata. It will be checked in a workflow and culminate in the description being discoverable via the Metadata Hub and ARDC.

DISCUSSION

At the time of writing the project is in delivery phase and due for completion in November 2011. Long term sustainability is a significant influence in the choices of platforms was by leveraging University and sector resources. The use of SharePoint for the DMP and RDP generated interest from several universities. This design enables domain-specific data capture workflows and systems to be integrated with institutional metadata and data capture channels. Projects underway or planned include integrated workflow for bioinformatics, code versioning for computational sciences, online survey archiving and a local instance of MyTARDIS. The technical elements of the project are progressing smoothly. The major difficulties are associated with quality and business rules. Quality assurance and release for automatically compiled metadata requires a combined technical, process and training solution.

ACKNOWLEDGMENT

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

Griffith University and University of Melbourne provided invaluable assistance to implement Vitro as the Metadata Hub.

REFERENCES

ABOUT THE AUTHOR(S)

Salim Taleb is currently a member of the eResearch Support Team at Curtin University in Bentley, Western Australia. He is the lead business analyst working on the Australian National Data Services (ANDS) projects in Curtin University. Salim has extensive requirements gathering and communication skills, has been involved in the analysis and documentation of large-scale projects, has a demonstrated ability to manage multi-faceted programs and effectively interface with all levels within an organization. His extensive knowledge and experience stem from many years as a business analyst in the higher education, telecommunication and engineering industries.

Peter Hicks is currently responsible for the eResearch Support portfolio in Curtin IT Services, focused on helping researchers enhance their research outcomes through the use of ICT. For the past 10 years he has worked in various ICT service roles at Curtin University including the dial-up helpdesk, network systems engineer and team coordinator, and leading School/Faculty teams through consolidation and restructure.