# Towards a Step Change in Managing Research Data at Western Sydney University

## Assessment of the active data management ecosystem

**Western Sydney University**

**January 2023**

**Juan Cooper, Andrew Leahy, Mike Kmiec and Tracy Donelly**

**URL: https://doi.org/10.26183/4atv-3s04**

## Executive Summary

This paper describes the approach taken by Western Sydney University in the assessment of a turnkey research data management platform. Key recommendations from the Active Research Data Management Element of the Research Data Management Framework developed by the Australian Research Data Commons were implemented. WSU-derived research data are currently stored across many different silos and there is consideration of management of data through research data management plans. However, the use of metadata is not consistent and often non-existent. WSU trialled the use of the Idea package provided by Cloud Services provider Intersect, which is based on Mediaflux, for the management of research data throughout the full data lifecycle from collection to archiving. The Active Research Data Management Element provided useful insights into the assessment of Idea in the areas of data management, use of metadata, automated data workflows, research data governance and the criteria for platform selection.

## The Challenge of Managing Research Data at WSU

The Australian Code for the Responsible Conduct of Research[1] (The Code) places responsibility upon an institution for “providing infrastructure, systems and services that facilitate good active research data management”. The Code (year and page) states that:

> *Research data controlled by the institution and/or its researchers should be stored in facilities provided by or approved by the institution. These facilities, including information technology, must comply with privacy requirements and other relevant laws, regulations and guidelines, and research discipline-specific practices and standards related to safe and secure storage of data and information.*

Best practice in research data management has the prime objective of managing the full life cycle of the data including planning and designing, data collection and analysis, write up and dissemination of findings, and storage and/or disposal of all items related to the research project and all related data, upon completion.

Western Sydney University (WSU) has appropriate policy surrounding Research Data Management (RDM). However, there are several challenges in achieving research data management nirvana:

- Storing your data: In the age of the data explosion, with increased resolution from research instrumentation, the storage of captured data requires increased resources at increasing costs

- Finding your data: Most datasets lack meaningful primary information about the data. This leads to not knowing what data we have, where it is and how it got there. And with this, the invisibility of these data to future researchers.

- Workflow automation. With complex datasets comes the requirement to automate data workflows ensuring that sufficient informational content is captured during this process.

[1] Australian Code for the Responsible Conduct of Research, 2018

- Data governance: With the collection of data comes the requirement to manage the security of the data in compliance with institutional and regulatory requirements.

The need for a well-integrated RDM system was considered especially important for providing a seamless user experience. Researchers should ‘know what they know’, be aware of what’s captured in their own research datasets and where they are stored in physical and digital form. After completing a research project, this information may be poorly documented or not documented at all, making research data difficult to find, access and reuse. This poses challenges for other researchers when it comes to reproducibility and building advances in research.

## Current Approaches to Managing Research Data at WSU

The WSU Research Data Management Policy mandates:

- *All data collected and/or created by researchers must be maintained securely to prevent unauthorised access, alteration, removal, destruction, accidental or intentional damage.*

- *Research Data, research records and primary materials must be deposited in Working Data Storage as soon as possible after collection or creation.*

- *Researchers and HDR Candidates, in consultation with supervisors, must develop a research data management plan at the beginning of their research project to ensure a robust data management framework is in place prior to starting research at the University*

WSU provides recommended options for working data storage including the University network share drive, OneDrive, AARNET CloudStor and eResearch support partners access to managed cloud storage. However, the general approach is that researchers are given their own file storage directory and perhaps access to shared collaborative research group storage with researchers taking their own responsibility for how data is organised, managed, and shared. Legacy issues around data storage abound, especially the challenge of “shadow IT” which is the use of non-approved methods of storing or managing research data.

WSU currently uses ReDBox[2] from Queensland Cyber Infrastructure Foundation (QCIF) to collect Research Data Management Plans (RDMP) from researchers. This RDMP tool details how the data will be managed, analysed and stored, as well as what mechanisms may be in place for licensing and preservation. However, a RDMP does not provide a robust mechanism for the management of data and compliance with Policy is difficult to audit.

## Towards a Step Change in Managing Research Data at WSU

As part of the Australian Research Data Commons (ARDC) National Data Assets Initiative, the ARDC funded a collaborative multi-university Institutional Underpinnings[3] project to

[2] ReDBox Research Data Box

[3] ARDC Institutional Underpinnings Project

develop a national Institutional Research Data Management Framework (the Framework). Several essential Elements to the Framework were developed, one of which is the Active Research Data Management Element[4]. The intention of this Element is to provide "*institutions with guidance to ensure that research practice is efficient and impactful, and that research data is managed according to requirements such as those outlined in The Australian Code for the Responsible Conduct of Research*".

The Element defines active data management as "*the management of research data at any stage during the life of a research project, including selection of what to collect/acquire, collection/acquisition, storage, analysis, visualisation, and collaboration. Active data management ends when the data is either disposed of or moved to long-term storage after project reporting is complete, and does not include sharing data for re-use after the life of the original project"*.

Key highlights from the Active Research Data Management Element to enable effective RDM are:

- Consideration of infrastructure requirements.
- The importance of user focused design in ensuring adoption by researchers
- Integration with existing research administration platforms
- The importance of soft infrastructure elements such as governance and communication

To explore future directions and options for management of research data at WSU, an active research data management approach was applied to evaluation of research data management platforms. Several recommendations from the Active Research Data Management Element were found to be very useful in this process.

## Pilot Study with Intersect Idea

Western Sydney University has engaged Cloud Services provider Intersect to provide access to Mediaflux[5] on an Intersect research data management platform called Idea[6]. Idea has been designed with the intention of managing research data throughout the full data lifecycle from collection to archiving. WSU-derived research data is currently stored across many different silos and although some thought has gone into management of data through research data management plans, the use of metadata is not consistent.

### Approach

Our approach in implementation of the Active Research Data Management Element of the Framework was to take the turnkey Idea platform, configure using default metadata schema and engage with researchers in ingesting live research project datasets. A key consideration was to have researchers involved at all stages of the Active Research Data Management solution planning process (**User-Focused Design**)[7] and foster compliance by design to

[4] ARDC Institutional Underpinnings Element: Active Research Data Management

[5] Arcitecta Mediaflux

[6] Intersect.data.edu.au (Idea)

[7] ARDC Institutional Underpinnings Element: Active Research Data Management – Recommendation 2

reduce the burden of managing working data at scale.

Taking this turnkey approach obviously brought researchers in much later than if they had been involved in the full process of planning, acquisition and implementation of a data management platform. However, lessons were still learned in how to engage with researchers and data administrators in trialling a new platform.

The type and quality of user engagement with the Idea system proved to be very important. Although demonstrations of the platform were useful in showcasing the functionality, intensive, and tailored, small group or one-on-one training was required to quickly train users to competency. The training and support requirements should be factored in and heavily weighted in the evaluation of a RDM platform.

## Data Management

A small number of datasets were onboarded to the new system during the training phase of implementing Idea as an active data storage solution. This involved provisioning of data collections in Idea, data cleaning and standardising file naming conventions for datasets. The pilot study was used to assess the usefulness (and hence level of incentive) of Idea and associated utilities to participating researchers in Active Research Data Management in the organisation of data and moving data in and/or out of their working storage, with a key consideration being **Compliance by Design**[8]. Access to a centrally provided research data management platform should make research easier and reduce the burden of meeting compliance obligations.

The Idea platform provided two portals for access to Mediaflux; one was an application run on a local machine (Mediaflux Explorer) and the other was an application in the user’s web browser (Mediaflux Desktop) hosted by Intersect.

The focus of Mediaflux Explorer is uploads, downloads, queries and sharing of data. Mediaflux Explorer proved to be a very simple to use graphical interface for fast and reliable upload of data to the Idea server and retrieval of data from the server. In the main screen, the representation of data to desktop file explorers such as Windows file explorer and Mac Finder provided some familiarity and confidence to users tasked with organising their data files and moving them in and out of their working data storage in Idea.

Mediaflux Desktop is a fully functional general purpose graphical user environment that offers a portal into the data, metadata, and services available in Idea. Users can import/export data files, search and discover data files, create, and update metadata documents (schemas) and manage user account access. Comprehensive user training and a clear definition and understanding of user roles was critical in user acceptance of the Mediaflux Desktop application. As with any IT-based platform, the availability of high-quality instructional material proved crucial, especially for provisioning of project data collections and completion of the required metadata in the provided schema formats.

[8] ARDC Institutional Underpinnings Element: Active Research Data Management – Recommendation 3

## Metadata

Implementation of the Active Research Data Management Element of the Framework also considered the importance of metadata when it comes to research datasets. Metadata (those elements which describe the research data) allows discoverability, interoperability and reusability of research data and is therefore a key consideration to any institution's successful adoption of a research data framework and platform. The identification of a set of key metadata about data that supports good management and automation, determination of the points where this metadata will be collected, and a common source of truth for this metadata.

In legacy file storage systems commonly used at WSU, the metadata is typically limited to path, owner, creation, modification, and last access dates, as well as read/write permissions. In Idea, much more detailed contextual information is provided at both the institutional and researcher level. This includes:

- Research and project information associated with the data
- Personnel (researchers, data management, collaborators)
- Funding source
- Description of the data in the file
- Security and access protocols
- Governance requirements
- Protection and rights requirements
- Retention requirements (expiry/destroy review date)

Idea enables metadata based on Dublin Core[9] schema to be associated at the Project data file level. And was shown to overcome one of the biggest (at least perceived) oppositions to data management through the use of metadata, that being the time and effort required by researchers to curate their data files. The Idea program structures research data by creating a Project directory with all data files placed in a subfolder of that Project. Workflows are established to ensure that each subfolder and file inherits the Project's metadata tags. Researchers only then need to add metadata specific for a particular data file as it is being imported into a Project subfolder.

The Idea platform was also able to provision the actual storage location, and allocated resource size, for the Project and data files. This was especially important in empowering data managers to ensure that resource data was stored in approved storage locations reducing the impact of "shadow IT".

Investigation of WSU's Research Data Management Plan (RDMP) repository, ReDBox, and Idea found that whilst there were some shared metadata attributes, there were other attributes used in both systems that did not have equivalents. Since the Idea asset structure features a complex series of nested nodes whereas the RDMP tool captures metadata at the project level, this was not unexpected. A comprehensive case study evaluating Idea metadata schema against Registry Interchange Format – Collections and Services[10]

[9] The Dublin Core Metadata Initiative

[10] Registry Interchange Format – Collections and Services

(RIF-CS) format, DataCite[11] and WSU's RDMP fields determined that RIF-CS format based on ISO 2146:2010 is the preferred metadata schema for active research data at WSU.

## Automated Data Workflows

Idea and its collection-based system of storing and managing active research data also offers the potential for automation of data collection through interfacing with analytical instruments that produce both raw and processed datasets. Western Sydney University have embarked on a Proof of Concept with Intersect to validate a data movement tool called DataBolt[12]. DataBolt provides an automated process to move raw data from the data acquisition computer of an instrument to a defined collection in Idea. Local data collection agents are installed on instrument computers and remote data agents are configured to deliver the data to Idea. The file structure for the data and associated metadata are designed to be consistent and aligned with informative file system structure of the Idea collection.

## Research Data Governance

The Active Research Data Management Element recommends: *Clear responsibility within the Institution for decision making for (governance of) active research data management*. With this comes the need to define the institutional roles and responsibilities related to research data management and the complementary **Research Data Governance**[13] framework. Researchers also need to be provided with clear direction on their responsibilities for decisions about data and its management. The WSU Research Data Management Policy provides some guidance on who has the responsibility to make decisions about active (working) data and its management but does not establish a support framework that incorporates compliance, operational policies; procedures guidelines and training and reporting.

Western Sydney University is currently developing a Data Classification and Handling Guide to define a framework for assessing the sensitivity of data in electronic systems, apply a classification with minimum handling requirements needed to protect the University's data. This will be applied to all data that are created, collected, stored or processed by the University, in electronic formats. This Guide will define the key roles of:

- Data Custodian (responsible for the business use of the data)
- Data Owner (with administrative and/or operational responsibility for data)
- Data User (An individual who has been granted explicit authorisation by the relevant Data Owner to access, use, alter or destroy data within an information system)

We have focused on these key roles and responsibilities for institutional data governance at WSU. These data handling roles were easily mapped to the Idea data management access

[11] DataCite

[12] Intersect DataBolt: Securing remote data into managed storage

[13] ARDC Institutional Underpinnings Element: Active Research Data Management – Recommendation 1

roles of system administrator, project leader and researcher, respectively.

Cross matching of roles across research data management policy and associated guidelines and procedures is underway.

### Platform Selection

The Active Research Data Management Element provides a set of key considerations in selecting a platform to underpin active research data management[14].

- Disaster recovery
- Ability to support data discovery
- Interface with instruments that collect data
- Ease of local collaboration
- Ease of external collaboration (particularly the management of authorisation and authentication)
- Ease of use with required research tools, analysis environment, or analysis workflows
- Ability to maintain central control/oversight
- Location of storage
- Ability to manage access

The Idea platform was assessed against these criteria. Whilst the outcome of the assessment of a platform is important in the decision whether to purchase, or not, this exercise concentrated on whether these were the right considerations and whether additional considerations should be included.

Appendix 1 contains a set of key criteria that were developed for assessment of vendor specifications when going to market for an Active Research Data Management platform. This set of platform selection criteria covers and extends those described in the Active Research Data Management Element with additional criteria developed in the important areas of data security and metadata schemas.

## Conclusion

Implementation of the Active Research Data Management Element of the Framework in the pilot of Idea highlighted several important soft considerations which are critical to acquisition of an enterprise-wide platform for research data management.

- Early engagement of researchers and data administrators is critical to planning, acquisition, and implementation of a data management platform.
- A research data management platform should be intuitive and easy to use especially in ensuring compliance with data policies and regulations.
- Metadata schemas are essential to ensuring data is tracked and discoverable
- Modern analytical instrumentation can provide massive dataflows which are best handled with automated harvesting and transfer to data storage
- Key roles and responsibilities for institutional data governance must be established

[14] ARDC Institutional Underpinnings Element: Active Research Data Management – Platform Selection

and understood by researchers

The Active Research Data Management Element also assisted in development of a meaningful set of requirements and recommendations essential for inclusion in the scope of works for selection of research data management infrastructure platform for WSU. This set of functional requirements covered areas such as:

- Data access. How will data be controlled?
- Data organisation. How will data be discovered?
- Metadata schemas. How will data be described?
- Data integration. How will data interact with other platforms?
- Data collaboration. How will data be shared?
- Data security. How will data be saved?

During the process of trialling Idea, we have established a unique working relationship between Research Services, the IT division and the Library, lead by a solutions-focused working group to promote best practice in data management. Research data flow from research projects, through IT systems and is then managed by the Library. Thus, there is a natural involvement in active research data management from these three key support areas in our university.

We have gained valuable insights which have been integrated the update of WSU Research Data Management Policy, with key aspects of the Active Research Data Management Element incorporated into the Policy.

## Appendix 1.

| | **Functional Requirements- Data Access** |
|---|---|
| **FRDA 01** | Provide self-service functionality in managing data files including importing, exporting files and folders including an intuitive user-interface for the end-user and in line with similar formats available to the public such as Windows File Manager. |
| **FRDA 02** | Be able to manage different levels of access, for data users, data owners, data custodians and systems administrators within different areas of WSU and different approved external collaborators |
| **FRDA 03** | Sufficient controls and protocols to ensure central control/oversight of platform is maintained |
| | **Functional Requirements- Data Organisation** |
| **FRDO 01** | Software must provide an intuitive folder structure/layout that will be easy for users to navigate and understand. |
| **FRDO 02** | Data must be easily searchable and able to be filtered according to pre-set and customizable criteria that fits user needs, such as owner, project, etc, as well as free text search. |
| **FRDO 03** | The ability to set standardised formats of data recording must be provided to make it easy to exchange and integrate with other data, as well as understand the information within different contexts. |
| | **Functional Requirements- Metadata Schema** |
| **FRMS 01** | Ability to customise metadata schemas based on a variety of standard schemas including Dublin Core, DataCite, etc. |
| **FRMS 02** | Ability to prompt attachment of different metadata schemas to project collection, folders and datasets |
| | **Functional Requirements- Data Integration** |
| **FRDI 01** | Potential for integration with instrumentation for automatic gathering of data files and transfer to project data collections using an API or equivalent application |
| **FRDI 02** | Potential for data transfer to and from research tools and through analysis workflows including external data analysis software systems |
| **FRDI 03** | Ability to integrate with HPC scratch storage adding appropriate metadata during transit through the data processing pipeline. |
| | **Functional Requirements- Data Collaboration** |
| **FRDC 01** | Ability to create a shared data repository with multiple user access to function as a central database for a project team, streamlining teamwork and ensuring all have access to the relevant data. |
| **FRDC 02** | Provide ability to export data in universally usable formats (such as PDF, Excel etc.) |
| **FRDC 03** | Provide ability to sharing data across disparate working data storage sites including presenting data via a “single global namespace” |
| | **Functional Requirements – Data Security** |
| **FRDS 01** | Ability to automatically backup data with full disaster recovery and the option to recover historical versions. |
| **FRDS** 02 | Provide encrypted communication between computers and servers. SSL encryption methods, such as AES-256, are a good example of high-security encryptions. |
| **FRDS** 03 | The location of the primary data centre should be in Australia |
| **FRDS** 04 | All data and systems must be held within data centres of at least Tier III classification |