Chapter 8

Case study 3: Monash University, a strategic approach

Anthony Beitz, David Groenewegen, Cathrine Harboe-Ree, Wilna Macmillan and Sam Searle

This is a preprint of a chapter accepted for publication by Facet Publishing. This extract has been taken from the author's original manuscript and has not been edited. The definitive version of this piece may be found in Graham Pryor, Sarah Jones and Angus Whyte, Delivering Research Data Management Services: Fundamentals of good practice, 2013, Facet Publishing, which can be purchased from www.facetpublishing.co.uk. The author agrees not to update the preprint or replace it with the published version of the chapter.

1. Introduction

Monash University recognises that if research data is better managed, more discoverable, available for reuse and exposed to relevant communities it will contribute to increased research impact, enhanced research practice (including collaboration) and improved education outcomes.

Monash has taken on the challenge of developing research data management (RDM) using a multifaceted, multilevel and strategic approach. This has included leadership and participation in large Australian Federal Government initiatives at the same time as using 'little steps' approaches within the institution. Monash has led national projects to prototype and develop RDM infrastructure, assumed responsibility as the lead agency of the government-funded Australian National Data Service (ANDS), formed an institutional structure for RDM governance, established a Strategy and Strategic Plan for 2012-2015 and a RDM policy with associated procedures and guidelines, delivered programmes for RDM skills development, established a petabyte data store and developed and deployed a range of discipline-specific and versatile solutions for the management of research data and associated metadata.

The University continues to identify RDM as critically significant to its research performance and to the fulfilment of compliance requirements and community expectations. All members of the Monash community share responsibility to improve RDM in a coordinated and integrated way; to support this, the University has made ongoing appointments into research data management roles while also seconding librarians and information technology staff into shorter-term positions to build capability and expertise.

This chapter explores the University's work in the period from 2006 to 2013 and examines the issues and challenges to be resolved when planning and implementing effective RDM. It describes in some detail the characteristics of Monash's organisational approach to RDM, explores both the non-technical and technical components of Monash's RDM infrastructure, looks at what developments are anticipated and outlines Monash's strategy to promote sustainable RDM infrastructure.

2. Background

2.1 Monash University's research environment and history in research data management

Monash University was established in 1958 and comprises 10 faculties on six Australian campuses. It also has campuses in Malaysia and South Africa as well as centres in Italy, India and China. The student body numbers around 63,000 from more than 100 countries (7.1% of these are higher degree by research students). The University has approximately 3,800 academic staff and received over \$AU282 million in externally funded research income in 2011.

Although Monash conducts research across numerous fields of study – over 150 in all – it has a particular focus on a number of specialisations, described internally as “leading capabilities”. These capabilities fall into four categories, aligned closely with the Australian Commonwealth Government's National Research Priorities and its National Collaborative Research Infrastructure Strategy. They are:

- Health and wellbeing (accident, injury and trauma, cancer, health, wellbeing and social change, infection and immunity, neuroscience);
- New therapeutics (public health, stem cells and regenerative medicine and women's, children's and reproductive health);
- New industries and productivity (advanced manufacturing, aerospace, energy-related materials, nanomaterials); and
- Sustainable environments (climate change and weather, energy, green chemistry, sustainability and urban water) and resilient cultures and communities (development economics, economic modelling, education, mental health law).

Monash's journey towards more effective RDM began with the release of its information management strategy in 2006, the establishment of a Research Data Management Subcommittee, and the subsequent creation of a dedicated organisation-wide Data Management Coordinator role based within the Library. Monash's decision to provide centrally funded digital research data storage for all researchers, including Higher Degree Research students (HDRs), was a great incentive for researchers to think about their research data and how it might be better managed. This was a significant practical expression of the university's intention and commitment to improve RDM.

With the release of the *Australian Code for the Responsible Conduct of Research* in 2007 (1), Monash established a Research Governance Implementation Committee (RGIC). This provided the context and a process for the development and endorsement of the RDM Policy and Procedures for staff, adjuncts and visitors as part of a full review of all research related policies and procedures. The iterative process of developing these policies was used as a communication and awareness-raising activity in its own right. At Monash, all policies are required to have associated procedures and guidelines. In order to provide practical guidelines on data management issues for researchers a data management website was launched, developed by the Data Management Coordinator.

Well into the development of the Policy and Procedures for staff, Monash's Research Graduate School (now known as the Monash Institute of Graduate Research) indicated that it wanted to develop procedures targeted specifically for HDRs, to parallel those for staff. The RDM Policy and Staff Procedures were endorsed by the Academic Board in 2010 and the separate but related HDR Procedures in early 2011. The RDM Strategy and Strategic Plan were endorsed in 2012.

These activities have led to new research outcomes, better research practices and facilitated collaborative research. For instance, the Biomedical Data Platform data capture solution, MyTARDIS, is a multi-institutional collaborative venture. It facilitates the archiving and sharing of data and metadata collected at major facilities such as the Australian Synchrotron and Australian Nuclear Science Technology Organisation, and within Monash University and other institutions (see section 5.1. for more detail). Monash is now investing in an extended programme of coordinated RDM activities that holistically address technology, professional development and cultural change. Monash began trialling a faculty-wide approach to improving RDM with a small faculty, Pharmacy and Pharmaceutical Science. The outcomes of this trial are discussed in section 4.1.

2.2 The Australian research environment

One critical aspect of the Australian research landscape is the *Australian Code for the Responsible Conduct of Research*, a guide to good research practice and research integrity, which details areas such as authorship, collaboration, research training and broader research practice. It also covers the management of research data and primary materials, outlining the responsibilities of researchers and their institutions. As a result, it has provided guidance for much of the RDM effort to date in Australia.

In regard to RDM, the *Code* is aspirational. All Australian universities have formally agreed to be guided by it; however, at the time of writing, there are currently no audit processes associated with the RDM sections. Monash has focused on improvements rather than compliance, while using the document to provide context and to raise awareness.

Within the *Code*, RDM is expressed as a joint responsibility, requiring researchers to manage their data well and their institutions to provide tools, advice and processes to enable them to achieve this. As part of the institutional commitment to shared responsibility noted in the *Code*, Monash took a critical first step by offering a large-scale institutionally managed digital research data store with essentially unlimited storage for researchers and HDRs. Monash was the first Australian university to provide this service. The storage is paid for centrally so it can be offered at no direct cost to researchers. The barriers to adopting this solution are thereby reduced, as is the need for researchers to purchase less persistent and lower quality storage from elsewhere, which was a key consideration in its design. Storage is of course only a part of data management, but its ready availability facilitates conversations around other aspects.

Australia has two main government funding councils for publicly funded research, the Australian Research Council (ARC) and the National Health and Medical Research Council (NHMRC). Recent policies introduced by both require publications from funded research projects to be made available on open access within 12 months of publication, but they do not mandate the deposit of data into repositories. Unlike the United Kingdom there is no history of disciplinary archives in Australia, with the exception of the Australian Social Science Data Archive (now the Australian Data Archive), so the funding bodies have been less inclined to mandate the deposit of data. There is, however, a strong institutional repositories community, due to the provision of significant government start-up funding – much of which has been directed to university libraries – and the central role of repositories in Excellence in Research Australia (ERA), a government

research quality assessment exercise. This means that libraries across the country have been well-placed to become involved in institutional RDM developments. It is worth noting that data management plans are not promoted in the *Code* or local research funder guidelines, although the requirement to maintain registries or inventories is; as a result Monash does not actively require researchers to lodge plans, focusing instead on improving practice through the provision of documentation relevant to particular projects.

The Australian government further supported development in this area through the National Collaborative Research Infrastructure Strategy Program (NCRIS) and the Education Investment Fund (EIF) Super Science Initiative. Of particular relevance to data management have been the \$AU75M investment in capability building and infrastructure for data reuse in the form of the Australian National Data Service (ANDS), the \$AU50M investment in national research data storage called Research Data Storage Infrastructure (RDSI) (2) and the \$AU47M investment in collaborative research infrastructure known as National e-Research Collaboration Tools and Resources (NeCTAR) (3). These projects have injected large sums into research institutions across the country to pursue projects in this area.

As Monash is the lead agent for the ANDS project it has been particularly closely involved with ANDS' work. ANDS has established several national services in research data management, has helped to publish data collections that are managed, connected, discoverable and reusable, has partnered with institutions to establish coherent institutional research data infrastructure and has improved the ability of the Australian research system to exploit its research data using tools, policy and human capability.

ANDS partners with research organisations and government agencies to build a cross-connected data environment called the Australian Research Data Commons (4). This environment can be accessed through Research Data Australia (5) which, as the Australian national research data collection, provides a window onto the commons. This portal publishes information on research data collections, the parties that generate them, the projects that used them and the services that enable access to the data or that exploit the data. ANDS has adopted this collections-focused approach to data because it is not possible to make a national cross-discipline portal that is optimised to all of the different types of data; for example, time series data is best accessed in very different ways to spatial data, and is different again to image collections that might be collected from a telescope or a microscope. Research Data Australia focuses rather on the national visibility of data collections within their richest possible research context by including related people, organisations, research projects, journal articles, access services, facilities, instruments, etc. While Research Data Australia enables discovery, access to the data is delivered through local institutional repositories and domain-specific access portals.

Funding from ANDS has had a significant impact on the range of services and projects undertaken at Monash, while also leading to the development of staff capability. Some of these activities are described below. Monash has made a number of collections available through Research Data Australia, and will continue to use its infrastructure to do so.

3. Organisational support for research data management

At Monash RDM is a shared responsibility, as is expected in the *Australian Code for the Responsible Conduct of Research*. The Provost, in association with senior managers, has been a staunch advocate for improving the management of research data, recognising its potential for improved research practice with increased research impact, reuse and communication. Rather than using compliance with regulation to justify improvements in RDM, she advocates improvement as a means of increasing the value of the data, especially by making it more accessible as and when appropriate. Monash is fortunate to have this kind of understanding and commitment expressed at a high level, making it a key university objective.

Governance has changed from time to time, and now the Research Data Management Sub-Committee reports to the Academic Board through the Monash Research Committee. A RDM advisory group with a practitioner focus has also been established, with representatives from all faculties and other key areas.

Leadership of the work to be undertaken is shared between the Library, the Monash e-Research Centre (MeRC) and eSolutions (the University's Information Technology division), who work with the Research Office, the Monash Institute of Graduate Research and, of course, the researchers, research managers, research assistants, faculties, academic units and centres.

Through the efforts of the University Librarian the Library is recognised as the chief steward of RDM at Monash. It coordinated the development of policy and strategy and provides leadership of the governance groups. It is also responsible for the co-ordination, development and implementation of RDM capability building and skills development for both researchers and professional staff, by providing advice to researchers directly and through the RDM website, leading the ongoing development of data management planning, curating strategically important data collections, and managing the Monash University Research Repository. The Library has an ongoing professional development programme to equip subject librarians to actively participate in RDM advocacy across the University.

MeRC provides a bridge between the researchers and eSolutions, leads technical innovation in RDM infrastructure and provides RDM advice that is mainly technical in nature to researchers.

eSolutions provides platforms-as-a-service, software-as-a service, software development capabilities and technical support for technical RDM infrastructure.

4. Overview of Monash University's non-technical research data management infrastructure

Monash has spent time considering the breadth of possibilities in non-technical RDM infrastructure activities and subsequently divided activities into themes or groups. These themes have evolved over time and at the time of writing are:

- Governance, policy and strategy;
- Information and advice; and
- Knowledge and skills.

When Monash decided on its initial approach it was referred to as 'little steps', for even small steps in the right direction were recognised as significant improvements.

4.1 Governance, policy and strategy

The first iterations of the current Monash RDM Policy were long, quite descriptive and experimented with organising the information in different ways – for example, by responsibility or by stakeholder. The separation of content relating to higher-level policy principles and that relating to procedures enabled the fine-tuning of the Policy, and provided a better understanding of the potential role of the Procedures. The policy defines researchers, data management and research data and identifies stakeholders and their roles and responsibilities. Further refinement involved extensive consultation, which proved to be an effective outreach activity in its own right. To be effective it required understanding the researchers and the role they play, as well as establishing partnerships between the University's areas of responsibility for policy, risk and audit, legal matters, research management, records and archives, commercialisation, copyright/intellectual property, ethics and research training so that all could contribute to improving the policy and helping with harmonisation.

Monash policies extend to all campuses, including those in Malaysia and South Africa, and wherever possible the RDM Policy is linked to related documents/policies/legislation, of which there are many. This process exposed the interconnections between RDM and other University policies, including those relating to intellectual property, ethics and recordkeeping.

The initial process of engagement focused on looking for early adopters across faculties by choosing research groups and approaching them about their RDM needs. It was notable that all researchers spoken with wanted to improve their data management and welcomed advice and assistance. This process was not intended to focus on HDRs, but discussions with senior researchers led to a focus on this cohort because fostering best practice amongst HDRs is more likely to change the overall environment in the long run, as they are the future of research practice.

One of the first faculties to adopt the new Monash RDM Policy was the Faculty of Pharmacy & Pharmaceutical Sciences (PPS), a discipline that produces and values raw data. The Associate Dean of Research was a member of both the Monash Research Committee and the Research Governance Implementation Committee and consequently was prepared to take the lead in promoting the objectives of the new Policy within the faculty while developing practical processes to improve the management of their research data.

PPS was a very useful area in which to begin at Monash, because it is a small single-campus research-active faculty with a culture of discoverable records for drug development. Some groups within this faculty have been early adopters of electronic laboratory notebooks, which offer interesting challenges. Laboratory notebooks have long been a key source of data about research and replicating their unique attributes in the digital form for the long term has proven difficult, not least because of the cost of implementation.

The outcomes of engagement with PPS have been:

- Establishment of a faculty subcommittee, led by the Associate Dean of Research, that includes researchers, faculty information technology staff, the Research Manager and representatives from MeRC and the Library;
- Connection of the faculty's local central data storage to the University's large institutional research data store, enabling nightly archival transfers;
- Mandated network storage for Honours students with agreed file directory structures and naming protocols for funded projects (although persuading the students to adhere to this has been an ongoing challenge);
- New hardcopy laboratory notebook protocols (with electronic ones to be addressed in future); and
- Faculty leadership in training and cultural change in the area of RDM.

This project has also been a useful demonstration of the way that diverse sections of the University can work together to achieve aims in RDM. The Associate Dean of Research provides an introduction to Honours and HDR students in association with Library staff on campus, who follow up after induction. In the project the Library has taken on the role of encouraging and marking progress through the 'little steps' approach, while MeRC and eSolutions contribute to and facilitate technical solutions.

This Faculty recognised the benefits of improved research practices and the promotion of cultural change. Professor Chris Porter, the Associate Dean of Research, when asked for an impression of the experience and progress in late 2011, responded by saying:

"Forcing ourselves to get in early with addressing the challenges of research data management has been very valuable. 'Wins' – such as central storage that is easily accessible by all staff – are proving to be much more powerful than perhaps we first thought. These systems are starting to focus attention on behaviours, and in many respects are changing the way data is viewed, shared, accessed etc. So a physical solution is actually changing the way we work – which in the long run will be a more useful benefit than safe storage."

This illustrates the importance of both technical solutions and the role of people in order to achieve a long-term impact. While of course there is more to do, the shared responsibility approach is working and will continue to bring improvements.

4.2 Information and advice

The RDM website (a component of the Library's website) is the main focus of the information and advice theme. It raises researchers' awareness of RDM and provides guidelines in dealing with common RDM issues. The website contains current information about:

- Planning, ownership, copyright, ethics and consent, storage and backup;
- Case studies;
- Training seminars and material; and
- Relevant resources (such as Monash's RDM Strategy and policy documents).

RDM information and advice are also available through a range of introductory printed material and a strong focus on interpersonal communication.

Figure 8.1 – Screenshot of Monash University's research data management website

4.3 Knowledge and skills

Much work is taking place in the area of skills development. In particular, new HDR students are made aware of data management through a range of events including induction sessions, presentations, consultations with groups and individuals, attendance at special events such as the annual Monash e-Research e-XPO and extended sessions run by learning skills advisors and subject librarians.

Librarians, frequently in collaboration with MeRC and faculty staff, have contributed to a range of local programmes, campus research days and HDR activities. The Library has also developed programmes that include RDM, such as the Focus series of workshops that run at least twice yearly for HDRs in Medicine, Nursing and Health Sciences. Working with the Monash Institute of Graduate Research, the Library has contributed RDM content to inductions and supervisor training and has been running 2 hour data management planning workshops for new HDR students. More formal procedures may emerge in future, for example tying confirmation of candidature to some form of data plan or asking HDR students to deposit data associated with theses. The Monash Institute of Graduate Research has been developing a more formal training and coursework programme for HDR students and early adopting faculties began some of these programmes in 2012. The Library has been working with the Institute and with faculties to provide RDM and other discipline-tailored skills development opportunities.

While policies and procedures are about best practice, it is possible and desirable to frame RDM in different ways when talking to researchers. After much engagement with its researchers, Monash has evolved communication strategies that focus on benefits and ask researchers to imagine a future in which better RDM has made a real difference to everyone's day-to-day working life. In the belief that in the long run carrots are more effective than sticks, Monash tries to take a 'what's in it for you' approach to RDM training.

4.4 Sustainability of non-technical research data management infrastructure

In order to sustain non-technical RDM infrastructure activities, Monash is seeking to integrate them into the wider outreach and teaching roles of the University so that RDM is seen as an overall part of good research practice rather than an exceptional or special activity. To achieve this, the role of Data Management Coordinator is expected to develop into that of a 'broker', one that provides input and advice into a broad range of University engagement with researchers.

5. Monash University's technical research data management infrastructure

Research institutions are increasingly aware of the need to ensure that their approach to RDM contributes to their institutional research strategies and goals and to this end they require fit for purpose RDM infrastructure to enable leading-edge research, increase research outcomes and impact, validate research, reduce legal risk, and attract the best researchers and additional research income. All this must be within a clearly promulgated institutional context to ensure that funds invested in infrastructure are not wasted.

Researchers need RDM infrastructure to achieve new research outcomes and higher research impact, to safeguard their data and deal with the data deluge, to facilitate collaboration, enable data reuse and validate their research. They also need to conform to funding body requirements, institutional policies, legal obligations, codes of conduct and cultural norms. Researchers themselves are generally focused on research outcomes, have goals that may change over time, and may be more loyal to their research community than their institution. As a result, if research institutions deploy RDM infrastructure that does not serve researchers' needs, researchers are likely to deploy their own, engage with external RDM infrastructure or move to a research institution that they perceive will provide for their needs.

Researcher and institutional needs can, therefore, be markedly different even though both recognise that research data and its associated metadata are critical to achieving their goals. Monash has considered these issues in the selection, development and deployment of RDM infrastructure and in the ongoing work to make RDM sustainable, and the resulting experience is now being used to develop the next generation of RDM infrastructure.

5.1 Enhancing research by integrating research data management infrastructure into the research workflow

Research workflow can be modelled in various ways, some more sophisticated than others. In order to demonstrate where research data management infrastructure fits into the research

process, the workflow can be modelled as a simple linear sequence of activities, as shown in Figure 8.2.

Stage	Description
Conceive	Conceiving a hypothesis to test.
Design	Designing a number of experiments to test the hypothesis.
Experiment/ Gather	For the sciences, conducting the experiments and collecting the results. For the arts, gathering research material.
Analyse	Analysing the results of the experiments and determining their findings.
Collaborate	Sharing findings with colleagues in order to gain different perspectives and to promote discussion.
Publish	Describing findings and making them publicly available.
Expose	Maximising research impact by disseminating the findings to key target communities.

Figure 8.2 – Associating research data management infrastructure with the research process

This model demonstrates that the relationship between the research process, RDM principles and practical data management starts at the concept stage and continues beyond publication to the ongoing dissemination of findings. RDM and its associated platforms span the total lifecycle of research data by supporting the gathering and production of digital research data, analysis and

sharing of results and publication of outcomes and by ensuring ongoing impact of the research through discovery and reuse of the data by others. Publicly available research data needs to be well-curated, its metadata harvestable in major formats, the data source registered with key repositories (such as institutional, national, and community portals), and references to the research data placed in key electronic journals. Underlying the research data platforms is research data storage. Recognition of the extent of the confluence between the research process and RDM has led Monash to implement RDM infrastructure which is either tailored to specific disciplines or else has enough flexibility to be adapted or re-purposed to ensure that it supports the whole research process. Adoption of these platforms has led to better research practices, new research outcomes and facilitated collaborative research.

An example of a RDM platform that not only supports research at the project level but also provides broader benefits is the MyTARDIS platform for Protein Crystallography. This platform, developed by Monash, is facilitating:

- *Better research practices* – Instances of MyTARDIS now exist at the Australian Synchrotron and at many universities around Australia. As raw data is captured by the Synchrotron's Protein Crystallography beamlines, metadata is automatically extracted. If the owner of the data belongs to a university with a local instance of MyTARDIS, the metadata and raw data are automatically shipped and catalogued in the researcher's local instance of MyTARDIS.
- *New research outcomes* – Monash's instance of MyTARDIS has been collecting raw datasets for a number of years. Researchers analyse the raw datasets in order to solve the crystal's atomic structure but not all structures can be solved. However, new analysis techniques are now solving atomic structures that were previously unsolvable. These new techniques are now being applied to earlier unsolved raw datasets stored in MyTARDIS. This recently led to the resolution of the PlyC atomic structure. PlyC is a potential replacement for antibiotics and its discovery is a significant research outcome.
- *Collaborative research* – Data collections in MyTARDIS can be kept private, shared with other researchers, or made public. Data collections in MyTARDIS can also be registered in Research Data Australia – or with TARDIS, The Australian Repositories for Diffraction ImageS (6).

5.2 High-level overview of research data management infrastructure

To provide a different perspective on the Monash RDM infrastructure, it can be split into six major categories, as shown in Figure 8.3.

Category	Description
Federated Collections	Components in this category facilitate the federation, discovery, and reuse of data.
RDM Platforms	RDM platforms facilitate the capture, curation, organisation, analysis, sharing, discovery, and reuse of data.
Advisory	Raises awareness in RDM and includes institutional RDM policies, strategies, practices, and tools and encourages RDM planning.
Data Source	The source of the digital research data, be it either manually generated, from an instrument, from a compute node or reused data.
Storage	Provides the fundamental storage for the research data. This can generally take two forms: Market, which is frequently accessed or updated; and Vault, which is archival, with only periodic access.
Support	This category contains various components which enrich the

	research data collection's metadata, provide persistent identifiers, enable authentication and authorisation, and facilitate citation.
--	--

Figure 8.3 – Overview of research data management infrastructure

This entire infrastructure need not be provided by an individual research institution; some elements may be provided by a state, national, international or commercial service provider. Monash has sought to identify a variety of options where appropriate.

5.3 Strategies in selecting, developing and deploying research data management infrastructure

5.3.1 Selection strategy

Researchers predominantly engage with a RDM platform. For this platform to be effective and have high utility it must fit with researchers' tools, workflows, instrumentation, methodologies, environment and, most importantly, culture. As most of these vary from discipline to discipline, it is unrealistic to believe that a single RDM platform will consistently meet researchers' needs. Indeed, research institutions should expect that a range of RDM platforms will be required in order to accommodate their researchers.

Researchers frequently have strong allegiances with their research community, sometimes ahead of their institution. Consequently, researchers may be inclined to choose to adopt a research community RDM platform over an established institutional solution, even if the institutional solution and service is technically superior; the need to collaborate with and contribute to their professional community takes precedence. Monash's view is that if a research community already has a well-developed RDM platform and strategy, and that if this platform meets the needs of the institution's researchers in the discipline, then the institution should facilitate its adoption by those researchers. In some cases, this might mean re-purposing the RDM platform to meet local needs, requiring some degree of software development.

Monash has many research disciplines, so it is unrealistic for the University to support a RDM platform for every discipline. Therefore, Monash has sought to strategically select the discipline-specific RDM platforms that it supports. Methods of doing this included filtering on research disciplines that are ready, willing and able to engage and focusing on disciplines that are strategically important to the institution's research impact and funding. Cloud based services, which alleviate the need for local expertise and support and may also reduce the costs and certain risks in providing these platforms to researchers, have also been considered.

5.3.2 Development strategy

Monash's philosophy is that given the potentially high development costs, numerous risks and challenges of software development projects, development of new RDM platforms should be a last resort. However, given the uniqueness of some research requirements and environments, the development of new tailored RDM platforms may at times be required.

In the Monash experience, the process of developing software for researchers is often quite different from that required for mainstream business software. Development partnerships between software developers and researchers generally work in an interpretive mode, with characteristics including an iterative process that is open-ended and thrives on ambiguity. Researchers are ultimately focused on the research outcome, and their requirements may change over time. In some cases, therefore, the software solution may only be required in the short term.

When developing RDM platforms for researchers at Monash, the following key principles and methods have been successful:

- Researchers must lead the development of solutions;
- These solutions must be owned by their research community rather than the institution;
- Existing RDM platforms should be adapted to save development time and reduce the risks of green field development;
- Agile software development methodologies should be used;
- Software should be easily deployable at other institutions; and
- Support should be provided for researchers in marketing the RDM platform to their research discipline.

5.3.3 Deployment strategy

5.3.3.1 Federated collections infrastructure

In the context of a research institution, federated data collections may have an internal or broader scope. Monash's view has been that it is preferable for federated collections infrastructure to be hosted and supported outside the institution. This raised the question of whether the university needed a central institutional research data repository.

An institutional research data repository has two major uses: to showcase a research institution's data and to track data holdings so that researchers within the institution can discover and reuse accessible collections. Monash has not implemented a comprehensive institutional data repository. Instead, it disseminates metadata to federated collections infrastructure from a range of data tools (such as research data management platforms and the Monash University Research Repository) using various interchange protocols and formats. Research Data Australia federates Australian research data collections and enables each research institution to showcase its own collections. This has been Monash's preferred method for comprehensive discovery.

5.3.3.2 Research data management platforms

The number of RDM platforms is growing. They can be hosted by research institutions, regional government based service providers, national government based service providers, international organisations or commercial organisations so it is important that research institutions are flexible in their approach to providing and supporting RDM infrastructure. Since a single research institution cannot afford to host and support a RDM platform for every discipline, it can be assumed that many RDM platforms used by its researchers will be hosted externally. The exception to this is in the case of sensitive data, such as that produced from clinical trials, which for legal and/or ethical reasons may need to be hosted on a local RDM platform.

Monash assists with hosting RDM platforms by providing virtual machines, a range of popular databases, plenty of freely available research data storage (with good access time) and assistance in deploying the software.

5.3.3.3 Data storage

The RDSI project noted above is now establishing national research data storage, which will impact on Monash's strategies for its own data storage infrastructure. Monash is acting as a local storage node as part of RDSI, and is considering how to maximise storage opportunities between the node and its own infrastructure. Movement of the data between these two options in order to maximise discovery and reuse is of particular interest. Monash views itself as the primary place for its researchers to store their research data because it has a longevity that exceeds that of regional or project-funded organisations.

5.4 Integration of research data management infrastructure with internal and external systems

5.4.1 Integration between institutional administrative systems and research data management platforms

When registering research data collections associated with a publication and/or to encourage reuse, it is desirable to include metadata about the related researchers, associated grants (including granting bodies) and associated publications. RDM platforms at Monash will increasingly include mechanisms to enable the association between a data collection and the related research administration metadata, the association of a research data collection with a persistent identifier, and the public release of a data collection via publication, syndication and/or harvesting. To achieve this, institutions need to consider techniques for making their research administration information systems accessible by RDM platforms. At Monash, this is provided through a web service that interrogates the central research administration system.

5.4.2 Integration between external systems and research data management platforms

To promote collaboration and the reuse of research data, and to simplify and expedite workflow, RDM platforms need to integrate with a variety of external systems, including such aspects as:

- Authentication and authorisation;
- Citation;
- Dissemination;
- The minting and maintenance of persistent identifiers; and
- Appropriate provision of analysis infrastructure (such as high performance computing).

Since one of the main aspects of RDM is data sharing and as most researchers will have collaborators external to their institution, RDM platforms that support single-sign-on through federated access technologies, such as Shibboleth, will be more convenient to use and are more likely to be adopted. Institutions will need to link their authentication systems with relevant identity providers and Monash has adopted the solutions provided by the Australian Access Federation (7).

RDM platforms are likely to disseminate metadata describing their research data collections to a range of repositories, including national/institutional catalogues of research data collections, community portals and electronic journals. Metadata may also need to be disseminated in a range of formats and using a variety of techniques. For instance, metadata contributed to Research Data Australia uses RIF-CS(8) to describe the data collection and its associated researchers, grants, and services, and is disseminated using OAI-PMH.

When a research data collection is made available as part of a paper/journal/blog, a persistent identifier is useful in enabling the data collection to be relocated. When selecting an identifier service, institutions should consider the longevity of the identifier provider service, cost in generating an identifier and the extent to which the type of identifier is used within the research community. Monash has chosen to use the Digital Object Identifier (DOI) service provided by ANDS and DataCite (9) to encourage persistent identification of the datasets and facilitate citation and impact tracking.

5.5 Monash's technical research data management infrastructure

5.5.1 Research data storage

Many researchers still store their research data on personal storage devices, such as CDs, USB sticks, portable hard drives and laptops. These kinds of media are easily lost, difficult to collaborate around and are easily corrupted especially if they are stored under sub-optimal conditions, while this approach also results in fragmented research data collections.

In the past many researchers using Monash's enterprise networked storage found their default quota too small. Increasing the quota was relatively expensive, so researchers began purchasing personal storage devices and/or being forced to make hard decisions about what to keep and what to delete. Monash responded by commissioning a petabyte research data store, known as LaRDS (Large Research Data Store). To reduce the barriers to uptake and to promote the fundamentals of good research data management practice, storage allocation is generous and provided free-of-charge. Access can also be customised. To enhance performance and reduce intra-university network congestion, some research teams have been provided with a local high speed NAS (network-attached storage), functioning as a staging post, which is synchronised to LaRDS on a regular basis.

As LaRDS retains multiple copies of each data file in order to ensure data integrity and reliability, its total storage capacity is much greater than the effective capacity. In April 2013, LaRDS stored 1.489PB of unique research data and had an effective capacity of 2.14PB. Figure 8.4 shows the growth in LaRDS in 2011-2013.

Figure 8.4 – Growth in research data storage at Monash University

Researchers in the long tail of data management often require only a safe place for storing their data, from which they can easily share it with colleagues. Further, researchers in general do not want to spend time thinking about what they keep and what they throw away. LaRDS provides these researchers with access to very large and reliable research digital data storage.

5.5.2 Research data management planning

Monash provides advice on RDM planning in the form of guidelines posted on its RDM website and through a data planning checklist document. The checklist guides researchers through a comprehensive planning process that addresses all the aspects of data management covered by the *Australian Code for the Responsible Conduct of Research* as well as the Monash University Research Data Management Policy (10). As the contents of this checklist and supporting processes are still being refined and there is currently no requirement in Australia to lodge a data management plan, Monash does not provide software tools to automate data planning.

5.5.3 Monash University research repository

The Monash University Research Repository contains content representing Monash's research activity. It was initially put into service as an open access publications repository but has since been extended to expose research data holdings. The repository provides a place to securely store and centrally manage selected research data, collections and related publications so that they are globally accessible online. It contains accepted versions of published works such as books, book chapters, journal articles and conference papers; non-published manuscripts and grey literature such as theses, technical reports, working papers and conference posters; and research data

holdings including data sets, image collections, audio and video files. Its primary role has been to host research material not available elsewhere and it contains over 90,000 records. Its role in RDM is expected to grow.

5.5.4 Discipline-specific research data management platforms

At Monash, a diverse range of discipline-specific RDM platforms are used by researchers. This section provides a sample of them.

CDD (Collaborative Drug Discovery) Vault

CDD Vault (11) offers a web-based solution for managing drug discovery data, primarily around small molecules and associated bio-assay data. Licensing is fee based.

DaRIS

DaRIS is a framework for managing data and metadata. It is primarily used to supply a secure repository for biomedical imaging data and metadata, although there is no actual technical restriction on data types. Licensing is part open source and part fee-based.

Healthy Food Basket

The Healthy Food Basket (12) was developed by Nutrition and Dietetics researchers at Monash University to measure and monitor the cost and affordability of a healthy basket of food for typical families. Its supporting IT platform facilitates the collection of data from supermarkets using a mobile app, reducing errors in collection and transcription, simplifying the management and reuse of the collected data, and expediting analysis and the production of reports.

Interferome

Interferome (13) assimilates a large number of data sets, including detailed annotation and quantitative data, from the microarray analysis pipeline and makes this available to researchers by providing enhanced search capabilities that allow them to query more than 2000 data points. This platform also has the ability to publish metadata about research data collections to Research Data Australia. The service promotes citations and data reuse and enables new discoveries from old data. It also facilitates comprehensive analyses like tissue expression and regulatory analysis. Licensing is open source.

MyTARDIS

MyTARDIS (14) began as an automated solution for managing and sharing raw protein crystallography data. Since then, efforts from many independent projects have enhanced and evolved the central MyTARDIS product. New features such as data staging mounts, automated metadata extractors, parameter set creation and high performance computing task scheduling have been added to meet researcher needs. A diagram showing an implementation of MyTARDIS sourcing research data from beamlines at the Australian Synchrotron is shown in Figure 8.5. Licensing is open source.

Figure 8.5 – Implementation of MyTARDIS

OMERO

Data-intensive research depends on tools that manage multi-dimensional, heterogeneous datasets. Open Microscopy Environment Remote Objects (OMERO) (15) is an open-source software platform that enables access to and use of a wide range of biological data. OMERO uses a server-based middleware application to provide a unified interface for images, matrices and tables. Licensing is open source.

OzFlux Repository

Ecosystem research in Australia investigates the role of ecosystems in the cycling of water and carbon between biospheric and atmospheric stores and the response of these ecosystems to changes in these cycles. Effective research is hampered by the lack of coordination in data collection, archiving and quality control from independently-managed measurement stations across remote Australia. The OzFlux Repository (16) standardises and automates the collection of data, archives and controls the quality of measurements from a network of measurement stations, integrates complementary data streams from different sources into a single data and metadata repository and facilitates the linking of data through Research Data Australia to encourage reuse. The licensing is open source.

5.5.5 Versatile research data management platforms

Most researchers do not have access to a discipline-specific RDM platform, so research institutions need to provide researchers with access to one or more versatile RDM platforms. As the research terminology and processes can differ greatly between research disciplines, it is expected that one RDM platform will not be enough.

At Monash, MyTARDIS has been used effectively by many scientific disciplines. Besides its officially supported disciplines of protein crystallography, electron microscopy and proteomics, it has also been used for medical imaging, quantum physics and material science. Monash also provides researchers with access to Arcitecta Desktop, a web operating system for metadata and data that enables users from individuals to distributed groups to ingest, discover and share any type of data. This tool is a commercial product for which Monash has a site licence.

5.6 Sustainability of technical research data management infrastructure

Monash's approach to promoting the sustainability of its RDM infrastructure involves:

- Thinking about sustainability from the start, when deploying or developing new RDM infrastructure;
- Making a strategic assessment before deciding to support any new discipline-specific infrastructure at Monash;
- Considering cloud based solutions, in order to reduce operational costs and outsource specialist support;
- Adopting or adapting a community solution before developing a new one from scratch;
- Promoting a sense of ownership by the relevant researchers and their community in discipline-specific infrastructure. This encourages adoption and provides better funding opportunities, because a community can access a larger amount of funding than an individual institution;
- Providing a nursery to nurture and develop new RDM services; and
- Devolving responsibility for RDM to the Monash community.

5.7. What's next? – the Virtual Lab

Generally, RDM infrastructure is independent of other research components. A new Australian research infrastructure project, the Characterisation Virtualisation Laboratory (CVL), integrates high performance computing, direct data capture from key instrumentation, RDM, data storage and visualisation infrastructure with analysis and imaging tools to support next-generation instruments. The CVL (see Figure 8.6) is a cloud-based environment for the analysis and visualisation of multi-modal and multi-scale imaging data. It is being developed by four universities (Monash, the University of Sydney, the University of Queensland and the Australian National University) together with four national imaging facilities (Australian Microscopy and Microanalysis Research Facility, National Imaging Facility, Australian Nuclear Science and Technology Organisation and the Australian Synchrotron). This remote desktop environment allows researchers access to a wide range of existing tools and services provided through a centrally managed environment.

As RDM infrastructure becomes more tightly integrated with other research infrastructure and as RDM practices become a common part of researchers' everyday practice, it is foreseeable that RDM will become indistinguishable from other aspects of the research process. The CVL takes the next step along this path.

Figure 8.6 – The Characterisation Virtual Laboratory desktop environment being applied to structural biology

6. Conclusion

From 2006 to 2013 Monash University has taken a strategic approach to RDM that aligns with the goals of the institution and its research agenda, while also providing infrastructure based on discipline and researcher needs. To build RDM capabilities, key areas of the University with specialist skills have worked together with senior management in a collaborative environment. This has created a pool of expertise in RDM with a shared focus on common outcomes, which will drive Monash forward as researcher needs and university goals evolve.

At an institutional level Monash has created and endorsed RDM policies and strategies; established organisational structures to plan, implement and promote RDM; developed an RDM website, guidelines and other resources; developed programmes to build capability; established technical infrastructure and deployed a number of RDM platforms; created a large and widely accessible institutional digital data store; and established the Monash University Research Repository. Jointly this provides Monash with flexible research infrastructure for the future.

From the researcher's perspective Monash offers a range of discipline-specific RDM platforms, provides advice and has started conversations on the process of curating and exposing research data collections. In strategically significant areas Monash is providing discipline-specific RDM

platforms and for more general areas it is providing one or more versatile RDM solutions to meet these needs. Monash will continue to develop infrastructure to meet the ever expanding and diverse needs of its researchers.

Monash embraced RDM to improve research practice, attract researchers and research income, encourage reuse of data to increase research impact, validate research outcomes and reduce legal risk. For this to be sustainable Monash strives to integrate RDM within all research processes, practices and training. The goal is for RDM to be standard research practice, including being part of research funding. The complex infrastructure now in place represents the first steps towards maximising the value of the institution's research.

7. References

- 1) <http://www.nhmrc.gov.au/guidelines/publications/r39>
- 2) <http://rdsi.uq.edu.au>
- 3) <http://www.nectar.org.au>
- 4) <http://www.and.s.org.au/about/approach.html#ardc>
- 5) <http://researchdata.and.s.org.au>
- 6) <http://tardis.edu.au>
- 7) <http://www.aaf.edu.au>
- 8) <http://and.s.org.au/guides/cpguide/cpgrifcs.html>
- 9) <http://www.datacite.org>
- 10) <http://www.policy.monash.edu/policy-bank/academic/research/research-data-management-policy.html>
- 11) <http://www.collaborativedrug.com>
- 12) <http://hfb.its.monash.edu.au>
- 13) <http://interferome.its.monash.edu.au/interferome/home.jspx>
- 14) <http://mytardis.its.monash.edu.au>
- 15) <http://www.openmicroscopy.org>
- 16) <http://ozflux.its.monash.edu.au>