Platforms for Research Data Management: Lessons Learned

Steve Androulakis[1], Anthony Beitz[1], Prof. Paul Bonnington[1], Calvin Chow[1], Virginia Gutierre[1], Nigel Holdgate[1], Dr. Grischa Meyer[1], Russell Sim[1], and Simon Yu[1]

[1]Monash University, Melbourne, Australia, merc@monash.edu

Summary

With the growing importance in properly capturing, managing, publishing, and reusing research data and its associated metadata and the recent injection of research infrastructure funds by the the Australian Government (NCRIS & Super Science initiatives) and the distribution and management of those funds by Australian National Data Service (ANDS); there has been a proliferation of new platforms for Research Data Management (RDM) at Australian research institutions. This trend is expected to continue over the next few years due to additional Australian research infrastructure investments.

Through Monash University's experiences in: developing research data management platforms for the DART and ARCHER projects; determining research data capture, management, and reuse solutions for various research disciplines from EIF funds; and in deploying innovative research capabilities within a university this paper provides useful guidance in the selection, development, and deployment of platforms for research data management.

Background

Research data and its associated metadata are no longer considered as simply research by-products; they are now used by researchers to validate, reproduce, and extend research. This in turn facilitates collaborative research and leads to new research outcomes and enhanced research practices. To maximise the benefits of this asset, the curation and management of research data needs to start at the point of capture and proliferate through to publication; and can be collectively referred to as a research data management platform.

Recently, the Australian Government, through the Australian National Data Service (ANDS), has generously injected a large amount of funding into research data management. This has led to a proliferation of new research data management platforms at Australian research institutions. This trend is expected to continue over the next few years due to additional Australian research infrastructure investments (such as NeCTAR, RDSI and future NCRIS-like investments).

Monash University has accumulated a wealth of experience in research data platforms due to: developing research data management platforms for the DART and ARCHER projects; determining research data capture, management, and reuse solutions for various research disciplines (including Protein Crystallography, Historical Studies, Asian Studies, Climate Research, and Optical Microscopy) from EIF funds; and in deploying innovative research capabilities within a university. Along the way, Monash learnt some lessons which provide useful guidance in the selection, development, and deployment of platforms for research data management.

Selection of Research Data Management Platforms

A RDM platform must fit in with researchers' tools and their environment. As these vary from research discipline to research discipline, it is unrealistic to believe that a single institutional solution will fit with all its researchers' needs. Research institutions will most likely host a range of RDM platforms for their researchers.

Researchers tend to have a stronger allegiance with their research discipline than their institution, and a stronger allegiance with their institution than the state. Therefore, researchers will generally adopt a research discipline solution over an institutional solution, even if the institutional solution is technically superior.

Developing a new RDM platform for a discipline takes time and is costly, so, building a new RDM platform from scratch should be considered only as a last resort. If a research discipline already has an RDM platform (or there is an emerging one) and it meets the needs of your researchers in the discipline, then adopt it. If it isn't suitable, then see if an existing solution can be re-purposed. Most likely, the re-purposing will require some degree of software development.

Development of Research Data Management Platforms

As stated above, the development of a new RDM platform should only be considered as a last resort. The development of software for researchers is quite different from typical business software. Researchers generally work in an interpretive mode, with characteristics including: an iterative process; open-ended; and thriving on ambiguity. They are very research outcome focused and due to the nature of their research, their requirements change over time and the solution may only be required for a very short period of time.

In developing RDM platforms for researchers, we've found the following principals and methods to be very successful:

- Solution must be lead by researchers and owned by their research discipline (not the institution)
- Adopt Agile software development methodologies
- Software should be easily deployable at other institutions
- Support researchers in marketing the RDM platform to their research discipline
- Encourage others to embrace and extend the solution
- It's unrealistic to expect any RDM platform to suit every discipline's needs without specific development or customisation
- By the same token, it's unrealistic on time and resources to have every discipline receive significant RDM development effort tailored to them
- Don't develop an RDM platform to be as discipline-generic as possible from the beginning: you risk being the "jack of all trades, master of none" and hence used by nobody
- Cherry pick initial disciplines as catalysts for the RDM platform to work with, find commonalities and points of divergence in research data/workflows and evolve the software architecture in response over time

Deployment of Research Data Management Platforms

Once an RDM platform has been selected or developed for a research discipline, it will need to be deployed. Most likely, this will be done at an institutional level (as institutions view themselves as the primary place for their researchers to store their research data).

As there may be many hosted RDM platforms, institutional support will need to be minimal, with most of the administrative support provided by the research discipline.

Institutional IT service providers can assist with the hosting of the RDM platform 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.

Conclusion

Research institutions will be hosting a range of RDM platforms for their researchers.

Developing a new RDM platform for a discipline should be considered only as a last resort. If a research discipline already has an RDM platform and it suits the researchers needs then adopt it.

When developing an RDM platform for a research discipline, ensure that development is lead by researchers, the solution is owned by the research discipline, Agile software development methodologies are adopted, and that the software is easily deployable at other institutions.

Institutional IT service providers can assist in the deployment of an RDM platform by providing virtual machines, databases, data storage, and assistance in deploying the software.

About the Author(s)

For each author, please include a short bio of 150-200 words. This information can extend onto a third page if needed. The information about the topic should take up no more than two pages.

Anthony Beitz, Manager of the Monash e-Research Centre, started his career in 1992 at the Telstra Research Laboratories. In 2006; he joined Monash University, where he managed the development of a prototype system which successfully demonstrated the management of research data from its acquisition to publication (DART Project); and then successfully lead the development of a Data Management Portal for research data (ARCHER Project). He then joined the Monash e-Research Centre in 2008 as its Technical Manager.

Prof. Paul Bonnington, Director of the Monash e-Research Centre, is a member of the Go8 Digital Futures group, and on the Board of Directors for the Victorian Partnership for Advanced Computing. He is a member of the steering committees for the Victorian Life Sciences Computing Initiative (VLSCI), Victorian e-Research Strategic Initiative (VeRSI) and National Computational Infrastructure's Specialist Facility for Imaging and Visualisation (MASSIVE). Paul is also a member of CSIRO's e-Research Council, and currently serving on the National Research Infrastructure Roadmap Expert Working Group for e-Research. He recently served as the Chair of the Steering Committee for the Australian National Data Service Establishment Project. Paul is a Professor in association with the School of Mathematical Sciences at Monash University.

Steve Androulakis is a software developer with the Monash e-Research Centre. He is the lead developer of the MyTARDIS project. His focus is on solving research data management problems, particularly in areas of structural biology.

Calvin Chow, MeRC Program Manager (Software Development) is a certified project manager and an Agile SCRUM master with over 14 years of software development experience. He is the project manager for the 8 Data Capture projects funded by Australian National Data Services (ANDS). Calvin graduated from Monash University in 1996 with 1st class honours in Computer Science and Engineering.

Simon Yu is a Senior Software Developer at Monash eResearch Centre since 2007. He is the lead developer for 4 of the 8 ANDS Data Capture projects: Ecosystem, Climate and Weather, Interferome, Multimedia Collections and ARROW. Prior to that, he was the lead developer for the ARCHER project and a java developer on the Persistent Identifier Linking Infrastructure (PILIN) project. Simon has a Master of Technology (Internet and Web Computing) from RMIT University.

Virginia Gutierre is Research Systems Facilitator at Monash eResearch Centre. She is the senior business analyst on the ANDS Data Capture projects at Monash University. Virginia has a honours degree in Software Engineering and Commerce from University of Melbourne.