Melding Data Management with Computational Workflows

Neil Killeen[1], Jason Lohrey[2], Wilson Liu[3], Slavisa Garic[4], David Abramson[5], Simon Milton[6], Andrew Lonie[7] and Gary Egan[8]

[1]University of Melbourne, Melbourne, Australia, nkilleen@unimelb.edu.au
[2]Arcitecta Pty. Ltd., Melbourne, Australia, Jason.lohrey@arcitecta.com
[3]University of Melbourne, Melbourne, Australia, wliu5@unimelb.edu.au
[4]Monash University, Melbourne, Australia, slavisa.garic@monash.edu.garic
[5]MonashUniversity, Melbourne, Australia, david.abramson@monash.edu
[6]University of Melbourne, Melbourne, Australia, simon.milton@unimelb.edu.au
[7]University of Melbourne, Melbourne, Australia, alonie@unimelb.edu.au
[8]Monash University, Melbourne, Australia, gary.egan@monash.edu

INTRODUCTION

Contemporary scientific research increasingly involves managing and processing large amounts of distributed data within collaborations of globally distributed teams of researchers. The modern eResearch paradigm provides new models of scientific collaboration and requires new scientific research processes [1,2,3] to meet this need.

Often researchers find themselves dealing with practical limitations in data handling processes that are frustrating and significantly diminish the quality of the scientific outcomes. Examples of these limitations are: (i) no automatic management of distributed data, (ii) unmanaged multiple copies of data (and loss of knowledge of which copy is 'correct'), (iii) decoupled separation of metadata (often held in a dedicated database) and data (often held in file systems), (iv) derived results are stored and distributed in unmanaged spreadsheets with poorly defined and un-integrated quality control, processing and analysis workflows.

To meet contemporary research data challenges, inter- and intra-institution researchers need systems that combine three key components into an integrated whole: (i) distributed data (raw, processed, analysed), (ii) meta-data and (iii) workflows (both experimental and processing). Historically, workflow engines and data management have developed independently and there are simply very few generic unified systems. The current (and future) data deluge is forcing the issues of better management; researchers must automate data capture and processing through streamlined workflows and furthermore integrate the management and application of these workflows with the management of their data. Within the Australian community, this has also been partly addressed by National projects such as the ANDS data capture projects.

The creation of fully generic software tools and techniques that can be applied ubiquitously across research domains is complex. However, an approach of selective generalization, in which generic tools are specialized for specific research domains (or research patterns) is an effective methodology [e.g. 4]. In this way, the structure and key processes of many experimental research projects may be simplified and codified into standardised processes.

This project seeks to solve these problems, with the aim of producing researcher friendly fully integrated data, metadata, and computation management that can be driven from simple user interfaces.

THE PROJECT

The Distributed and Reflective Informatics System (DaRIS) was built with the Mediaflux™ digital asset management platform to enable the capture of experimental methods, metadata and data [5]. That work is being extended through an Australian Research Council (ARC) Linkage project between the University of Melbourne, Monash University and Arcitecta Pty Ltd to combine experimental methods with computational workflows [8]. The project will meld data management through DaRIS with generic computational workflows using frameworks such as Nimrod [6] and Kepler [7] workflow systems.

The existing DaRIS system consists of a data model (called PSSD) to manage experimental methods, a web-based end-user portal (Google Web Toolkit based), a set of plugin services to manipulate the data model, domain-specific meta-data definitions and experimental method definitions (describing data acquisitions and how the state of a subject may change during an experimental process). DaRIS has been primarily utilised to manage raw Neuroimaging (e.g. MR images) data and meta-data. However, its generic framework is extensible to other domains and the system is in use with other bio-medical imaging domains.

DaRIS is implemented with the Mediaflux™ platform, leveraging its substantial capabilities. Mediaflux™ is

an extensible Java-based asset management application; it supplies services, service frameworks, a data model framework, a portal library and a generic desktop environment. It is capable of fully operating in a distributed federation of systems.

To combine data management and experimental methods with computational workflows as seamlessly as possible, there are a range of technologies and enhancements that must be developed. In order to achieve the project goals, changes have been required to all of DaRIS, Mediaflux™ and Kepler:

- The development of a transform step within the PSSD Method framework to enable a pipeline of operations to be execute by one or more computers on input data currently managed by DaRIS
    - The transform step may be initiated manually by the researcher or triggered automatically by processes such as the receipt of new data into the system
    - Includes the concept of a Transform Provider; an abstraction of some service that can perform transformations such as 1) a stand-alone executable program, 2) a Mediaflux™ service, 3) a workflow engine such as Nimrod/Kepler, 4) a job scheduling system such as PBS
    - The ability to store the transform parameters
    - The ability to store the resultant data in DaRIS
- Kepler/Nimrod
    - Implement Mediaflux™ actors and a Kepler transform provider
    - Extend Kepler to allow end-user control of computational processes (suspend, resume, abort)
- Identity Management is a critical issue in all heterogenous environments. The transform step will execute on diverse resources on behalf of the user and so therefore a credential must be available to authorize use of resources. In the absence of a unified identity management across diverse resources, the approach taken is to enhance Mediaflux™ to provide the generation of identity tokens:
    - Represent the actor that generated it
    - Can only be used to execute specific services on specific resources
    - Will expire when explicitly revoked/completed or timeout.
    - Associate credentials on other systems with these tokens. E.g. store a grid certificate securely within Mediaflux™ so the certificate can be retrieved to authorise a third-party resource
- The PSSD data model has been utilised primarily for holding raw data and only experimental Methods. It will be enhanced to:
    - Manage other types of data such as derived images (e.g. a region-of-interest, a registered image) and analysed data products (e.g. a table of results)
    - Extend the Method framework to include the Transform Step
- The DaRIS portal will need to be enhanced to support the new capabilities:
    - Integration with Transform Steps
    - The ability to 'manage' the vastly more complex and larger amounts of processed data in a way that does not impede research processes
    - Query and data access capabilities to enable researchers to easily find and explore their data.

The high-level project objectives will be presented, along with example use-cases. This is a research and development project, with many unknowns. The issues encountered and requisite technology frameworks will be discussed along with those areas of uncertainty requiring future research.

REFERENCES

1. Berman F. (2008), *Making Research and Education Cyberinfrastructure Real*, EDUCASE Review, Vol.43 No.4.
2. Hey T., Tansley S. and Tolle K. (eds) 2009, *The Fourth Paradigm – Data-Intensive Scientific Discovery*, Published by Microsoft Research, ISBN: 978-0-9825442-0-4
3. Multimedia Victoria (2006*), eResearch: an eRevolution - Victorian Government eResearch Initiatives*, Report 578, Corporate Public Affairs State of Victoria
4. Lohrey, Killeen and Egan, 2009, *An integrated object model and method framework for subject-centric e-Research applications,* Frontiers of Neuroinformatics, 3:19. doi:10.3389/neuro.11.019.2009
5. http://www.arcitecta.com
6. Abramson, D., Enticott, C and Altinas, I. *"Nimrod/K: Towards Massively Parallel Dynamic Grid Workflows",* IEEE Supercomputing 2008, Austin, Texas, November 2008.
7. Abramson, D., Bethwaite, B., Enticott, C. Garic, S., Peachey, T., Michailova, A, Amirriazi, S and Chitters, R., *"Robust Workflows for Science and Engineering",* ACM MTAGS09, 2nd Workshop on Many-Task Computing on Grids and Supercomputers, co-located with ACM/IEEE SC09 (International Conference for High Performance, Networking, Storage and Analysis), Portland, Oregon - November 16th, 2009.
8. *An integrative and distributed data management and workflow framework for e-research in biomedical imaging*, ARC Linkage (R1) 2011-2013, Prof G F Egan, Dr S K Milton, Mr J Lohrey, Dr A J Lonie, Prof D A Abramson.

ABOUT THE AUTHOR(S)

Neil Killeen (Centre for Neuroscience Research and Information Technology Services - Research, the University of Melbourne) works in research computing and eResearch within the Neuroimaging and broader University communities. He leads the development of the DaRIS data management system and supports it use within the University of Melbourne and at National Imaging Facilty nodes. Neil has a PhD in Astrophysics, and worked previously for the CSIRO's Australia Telescope National Facility in Sydney, where he managed and developed radio astronomy software and techniques, as well as undertaking astrophysical research. Neil played a leading role in the Australian contribution to the International Virtual Observatory, a global astronomy informatics program.

Jason Lohrey, CTO of Arcitecta, is the conceiver and architect of the distributed data management platform Mediaflux™ and Arcitecta's high performance XML encoded object database engine (XODB). Jason has more than 20 years experience in all facets of IT systems provision including conception, contracts, requirements, R&D, manufacture, delivery and support. He has experience in commercial, government, science and academic environments for domains including real-time process control systems, film and television, digital asset management, database design and research. Jason is connected with the research community through collaborative engagements with government and tertiary institutions in Australia.

Wei (Wilson) Liu (Centre for Neuroscience Research, the University of Melbourne) is a software engineer with an extensive research computing and informatics development background within Neuroimaging; Wilson is the lead developer for the DaRIS system. He has a Master of Software Systems Engineering from the University of Melbourne and a Bachelor of Engineering from Hebei University of Technology, China.

Slavisa Garic holds an Honours Degree in Computer Science. He began work on Nimrod upon the completion of his studies in 2002. Since then he has worked on the development of Nimrod/G toolkit and has been the core developer of the Nimrod back-end services within MeSsAGE Lab. He has extensive experience in working with current Grid middleware toolkits as well as relational database systems such as PostgreSQL.

David Abramson has been involved in computer architecture and high performance computing research since 1979. Previous to joining Monash University in 1997, he has held appointments at Griffith University, CSIRO and RMIT. He is currently Director of the Monash e-Education Centre, science director of the Monash e-Research Centre and a Professor of Computer Science in the Faculty of Information Technology at Monash University, Australia. He is a fellow of the Association for Computing Machinery and the Academy of Science and Technological Engineering, and a Senior Member of the IEEE. Abramson's current interests are in high performance computer systems design and software engineering tools for programming parallel, distributed supercomputers and stained glass windows.

Simon Milton holds a senior lectureship in the Department of Computing and Information Systems at The University of Melbourne. He is interested in the ontological foundations of IT modelling, the implications of top-level ontological commitments in IT modelling, as well as the value and use of ontologies in biomedicine. Dr Milton received his PhD from the University of Tasmania's Department of Information Systems in which he reported the first comprehensive analysis of data modelling languages using a common-sense realist ontology.

Andrew Lonie is the Head of the Life Sciences Computation Centre at the Victorian Life Sciences Computation Initiative. His research interests include genomics, statistical genetics and knowledge engineering.

Gary Egan is an NHMRC Principal Research Fellow and the Professor and Director of Monash Biomedical Imaging, a research platform that encompasses the biomedical imaging research facilities at Monash University, Melbourne. He has published over 170 papers and over 250 abstracts in peer reviewed journals and received over $18 million in research funding. He undertakes high resolution structural and functional brain mapping research and clinical neuroimaging research in Multiple Sclerosis and Huntington's disease. He is head of the VLSCI/LSCC Computational Imaging Theme, the lead investigator of the Victorian Biomedical Imaging Capability, and Deputy Director of the Australian National Imaging Facility