Bioscience Data Platform: TARDIS In The Cloud

Steve Androulakis\textsuperscript{1}, Ian Thomas\textsuperscript{2}, Iman Yusuf\textsuperscript{3}, Steve Bennett\textsuperscript{4}, Heinz Schmidt\textsuperscript{5}, Ashley Buckle\textsuperscript{6}

\textsuperscript{1}Monash University, Clayton, Australia, steve.androulakis@monash.edu
\textsuperscript{2}RMIT University, Melbourne, Australia, ian.edward.thomas@rmit.edu.au
\textsuperscript{3}RMIT University, Melbourne, Australia, iimanyusuf@gmail.com
\textsuperscript{4}VeRSI, Melbourne, Australia, steve.bennett@versi.edu.au
\textsuperscript{5}RMIT University, Melbourne, Australia, heinz.schmidt@rmit.edu.au
\textsuperscript{6}Monash University, Clayton, Australia, ashley.buckle@monash.edu

\textbf{INTRODUCTION}

Some of the most important challenges in biology over the next 30 years pivot on understanding the structure and function of proteins and their assemblies, due to their crucial role in the functions of organelles and cells. A structural characterization of the complexity of life on the cellular level, and accurate functional models of the biophysical and biochemical properties of large molecules such as those of proteins, DNA and RNA underpin modern medicine. This requires a multidisciplinary approach that integrates experimental techniques with computational methods. Monash University has a significant program of research focusing on characterising large multi-protein complexes; the research program at RMIT focuses on characterising the structure and function of biomedically important nanomaterials. In this project, we aim to build a platform that will be critical in attacking significant problems in biology, health and disease. We refer to the platform as the Bioscience Data Platform.

\textbf{THE BIOSCIENCE DATA PLATFORM}

The Bioscience Data Platform (BDP) is a NeCTAR research tool that aims to bring existing computational systems together in a way that allows scientists to seamlessly work with data from capture through to publication. The BDP is backed by the NeCTAR-funded Australian Characterisation Environment Virtual Lab to exploit the advantage of its uniquely powerful computational infrastructure.

The BDP project is developing a set of tools that link to high throughput computing infrastructure, potentially saving the scientist days of computational time and manual effort. Typically, the platform assists scientists that collect data at an instrument facility (such as the Australian Synchrotron) to find their data catalogued and easily accessible on the storage cloud.

The BDP aims to reach all areas of the structural biology workflow, from the inception of a research project through to scientific publication. The platform has three main features that are referred to as MyTardis ‘in the cloud’, HPC platforms and, Data Driven Publications.

\textbf{MYTARDIS ‘IN THE CLOUD’}

The Bioscience Data Platform (BDP) aims to establish itself as a nationally available, central service, by extending the MyTardis codebase to function on the NeCTAR Research Cloud and the integration of several widely used research tools.

Australian research groups can sign up to the BDP and subscribe to feeds of data from instruments at MyTardis-enabled facilities such as the Australian Synchrotron and ANSTO’s Bragg Institute. A web service interface is in development to allow new pathways
for data from instruments and facilities that previously weren’t MyTardis-enabled - to the BDP.

In addition to data subscription capability, users will be able to register their own research data with MyTardis, in order to preserve, share and take advantage of the annotation and visualisation capabilities of the BDP.

**HPC Platforms**

The BDP aims to integrate commonly used research tools with HPC resources via a web interface - eliminating the need for configuration and manual collection of output. The scheduling of high-throughput processing will take place within a set of simple web interfaces, and the outputs of processes will automatically feed back into the BDP. The byproduct of this integration is the development of a generic adapter that assists the linking of varying research tools to HPC resources into the future. The BDP plans on building support for linkages of existing HPC infrastructure (ie. RMIT clusters, VPAC, NCI, MASSIVE) to commonly used research tools such as VASP, CRYSTAL, SIESTA, GULP and DFTB. These are widely used across computational physics and chemistry.

**Data Driven Publications**

The Bioscience Data Platform plans to make data a ‘first class citizen’ in the research world. A rich data-driven publication interface is in development to be made available to all users of the BDP. Visualisations will be available alongside text, and peers can annotate and draw attention to all aspects, including text and interactive elements. These same tools aim to assist the peer review process by enabling researchers to share data with scientific journals then the general public in a secure and rich environment that tells the full story behind a discovery.

The significant investment by ANDS for Data Capture and Metadata Store solutions is being leveraged to give researchers the ability to register interactive data-driven publications hosted within the BDP to Research Data Australia, including the assignment of DOIs via ANDS’ services.

**Conclusion**

The Bioscience Data Platform provides the MyTardis data management system as a national service running on cloud storage hardware with linkages to major instrument facilities and HPC resources. The BDP will give Australian research institutes and labs the ability to register for this central web-based service and subscribe to feeds of their data and metadata from instruments such as those with existing MyTardis capability at the Australian Synchrotron and ANSTO. The BDP aims to empower the biosciences community in conveniently leveraging high performance computing resources for complex simulations and analysis by developing web interfaces for commonly used research tools and their outputs. Collaboration and communication between research groups will be enhanced via rich data publications. This presentation will reveal detailed plans for the development and implementation of this NeCTAR-funded Research Tool and will include the demonstration of a prototype.

**About the Author(s)**

Steve Androulakis is a software engineer/consultant at Monash University’s eResearch Centre and Faculty of Medicine. He and A/Prof Ashley Buckle created the TARDIS project in 2008 for raw protein crystallography data and have helped it grow into a multi-institutional collaborative venture. He is also the project manager of the NeCTAR funded Research Tool, Bioscience Data Platform as well as the designated ‘benevolent dictator’ of MyTardis’ code base.
Ian Thomas is a software developer and system administrator at the eResearch Office of RMIT University. He has worked in data curation for output of high-performance computing systems, microscopy data for materials, and screen media objects (film and television). His current work is in institutional metadata stores, decision support systems for climate change modeling, and in cloud-based platforms in support of eResearch applications.

Iman Yusuf is a software developer at the eResearch Office of RMIT University, with responsibility for developing cloud-based platforms in support of eResearch applications. She is also a PhD candidate in the School of Computer Science and Information Technology at RMIT University. Her research interests are grid and cloud computing, fault tolerance, reliability, and component-based software architecture.

Heinz Schmidt is Professor of Software Engineering at RMIT University where he is the Director of eResearch and heads the Distributed Software Engineering and Architecture lab in Computer Science. Heinz is also an adjunct professor at Mälardalen University in Sweden. Heinz received his PhD from Bremen University, Germany. He has over 30 years experience with component-based and object-oriented architecture, especially in parallel and distributed systems and languages in practice, research and education. Heinz is an eminent researcher who has published over 120 refereed articles, supervised over 25 higher-degree research students, and lectures in software engineering, distributed systems and enterprise architecture. Prior to RMIT Heinz held positions at Monash University, the CSIRO and ANU in Canberra, at the German National Research Centre for Information Technology and the International Computer Science Institute at the University California, Berkeley. Prof Schmidt has led large university-industry research collaborations, in the European ESPRIT program and the Australian Collaborative Research Center program, among others with SIEMENS, ABB, DEC and Olivetti, IBM and others.

Ashley Buckle is an Associate Professor and NHMRC Senior Research Fellow in Biochemistry and Molecular Biology at Monash University. His group studies the role of proteases in human health and disease using a combination of protein crystallography and computational techniques. His lab also created MyTardis - the world’s first repository for raw X-ray diffraction data. His group develop and host a range of bioinformatics resources, and use high performance computing resources to run molecular dynamics simulations to understand the role of conformational change and flexibility in protein function.