Data Capture from High-Performance Computing Facilities: A Case Study

Venki Balasubramanian, Amir Aryani, Ian Thomas, Heinz Schmidt
RMIT University, Melbourne, Australia,
{Venki.Balasubramanian, Amir.Aryani, Ian.Edward.Thomas, Heinz.Schmidt}@rmit.edu.au

A Data Curation (DC) application is being developed in eResearch Office at the RMIT University for curating the datasets generated by various material physics simulation packages such as VASP, CRYSTAL, SIESTA and GULP across the High Performance Computing (HPC) platforms such as National Computing Infrastructure -- National Facilities (NCI-NF), Victorian Partnership for Advanced Computing (VPAC) and HPC-RMIT. The developed application aims to lift the research data curation practice in this specific field, support the researchers in curation, and simplify and automate processes. As part of this development, a large number of currently existing data collections will be curated.

In the course of this application development various challenges were encountered. This paper exemplifies these challenges and shows how some of these were resolved. The paper also describes some of the unresolved challenges and the issues related to these cases. In addition, some common issues related to the e-Research community are addressed. The following sections describe the key challenges addressed by the DC application:

INTEROPERABILITY

The DC application is made up of multiple systems, and one key challenge is to make these systems work together. Unlike typical enterprise applications, e-Research deals with different domain-specific software systems and has specific challenges in interoperability of data. This may be due to lack of established e-Research standards for data representation and transfer. To allow data to flow between these systems, a number of data adapter (convertors) are required, and we learned that not all these systems have well-documented file formats.

There are few established eResearch software platforms such as Fedora Commons [1] and Mytardis [2]. However, these systems attempt to solve only specific parts of the complete (end-to-end) e-Research application space. For instance, Fedora addresses the problem of digital object repositories. There are software solutions that attempt to address the larger picture, such as Kepler for e-Science workflows, but these solutions do not address the specific parts in as much detail.

SECURITY

Security is always an issue in any application that involves data transfer between other systems. The challenge on e-Research is to create reliable data repositories and process for who can access them. Hence, we distinguish between safety of data as well as backup, storage, archiving, retention, authentication and access (identifying users and managing access to data). For a software solution, which is composed of multiple systems, it is crucial that the user authentication and authorisation happen seamlessly. The Australian Access Federation and other federated services are a solution here, enabling us to authorise researches prior to access to local repositories.

SYSTEMS OF SYSTEMS

A significant part of the DC project is in collecting (harvesting) data from existing systems. In an abstract view the components of DC project are systems by themselves. To make this project happen, the first challenge was to understand the functionality of these systems, and what the expected outcomes from each system would be.

In e-Research, we may have limited control over external legacy systems in terms of reliability and availability. In this way e-Research differs from enterprise architectures where a business may routinely have contractual or SLA agreements with external entities. In e-Research, it appears that relationships are often more ad-hoc involving orchestration of available data sources and transformers. Hence, the reliability profile of the individual components may not be known or not be under control, which may lead to uncertain overall reliability.

REUSABILITY

Reusability saves much in time and development cost. The core part of the DC project is the institute metadata repository. We learned that by utilising *myTardis,* an open source project led by Monash University, we could avoid significant development effort. However, we discovered that we needed to extend its functionality. Following the open source methodology, we aimed to make these extensions reusable by the community.

Researchers based in specific discipline create software that are specific to their domain of interest and they solve their own specific problems. This seems to be limitation for software engineers in developing scalable and adaptable software for the long term. The use of open source software can give more opportunities for researchers to build their own customized repositories from open source.

e-Research is still at a level of infancy with respect to acceptability by researchers because some believe, rightly or wrongly, that curation of data for the community takes time from their already busy schedules. Therefore it is imperative to emphasis the importance of curation and collection of data for upcoming generation of scientists.

DOMAIN USERS

The DC project is designed to harvest research data and publish the metadata. We observed that the project would not be successful without the collaboration of the researchers (domain users). They should see the clear benefits in the project outcome and advantages of using it. However, software should be effortless to use as much as possible, so it will not be seen as an obstacle to the users research activities. Rather it should assist them in storage, retrieve and access to their research data. This can be matter of survival or demise for any software solution in the longer term.

Our experience is that researchers wish to fit new software into their existing workflow and not change unless the solution is widely used or is a well-known approach in their own discipline. They want to see a clear advantage to changing their workflow. Not unlike most users, researchers are conservative in that they care little about technology. They simply want to get their work done.

CONCLUSION

There is a specific barrier to adoption of eResearch solutions. There is a danger in deployment of "good enough" solutions, which are specific only to a researchers problem or environment and have little room for growth.

There is a perceived lack of overall knowledge in what eResearch can provide for a researcher. For example, they might equate data curation with backup, where data curation has many more aspects. However, there is a stealth approach here: get the researchers into some sort of system (and not just as a set of spreadsheets) and then the data thus captured can then become output and input for future endeavours and may be transformed as required. However, the data has to be there first.

REFERENCES

1. Lagoze, Carl, Sandy Payette, Edwin Shin, and Chris Wilper, *Fedora: An Architecture for Complex Objects and their Relationships*, International Journal on Digital Libraries, Volume 6, Issue 2, April 2006
2. Androulakis S, Schmidberger J, Bate MA, Degori R, Beitz A, Keong C, Cameron B, McGowan S, Porter CJ, Harrison A, Hunter J, Martin JL, Kobe B, Dobson RC, Parker MW, Whisstock JC, Gray J, Treloar A, Groenewegen D, Dickson N, Buckle AM. (2008) *Federated repositories of X-ray diffraction images.* Acta Crystallogr D Biol Crystallogr. Jul; **64**(Pt 7):810-4

ABOUT THE AUTHOR(S)

Venki Balasubramanian is one of the core developers for Data Curation project. He also works in other eResearch projects in the RMIT eResearch office. He completed his PhD in sensor networks at iNext, University of Technology at Sydney under Prof. Doan B. Hoang. He also completed his Masters and Post-Graduate Diploma in Networking from University of Sydney and University of New South Wales, Australia respectively. He worked as Research Assistant in Advance Networking Lab in University of Sydney. His research interest includes eResearch, sensor and wireless ad hoc networks, QoS, web adaptation techniques for mobile devices and health care monitoring.

Amir Aryani is a software engineer in the RMIT eResearch office, and working on myTardis extensions as part of the RMIT data capture project. He is a PhD candidate in the School of Computer Science & Information Technology at RMIT University. His research interest is software engineering with focus on software evolution and maintenance.

Ian Thomas is a software developer and system administrator at the eResearch Office of RMIT University. He was formally a post-doctoral researcher in the School of Computer Science and IT at RMIT University investigating software engineering for real-time reliable systems and agent-based management systems for e-Health (with Monash University). His current work is in data curation for three domains: high-performance computing, microscopy data for materials engineering, and screen media objects (films and television).

Professor Heinz Schmidt is the E-Research Director at RMIT University. He is a member of VDI, IEEE and ACM. He is professor of Software Engineering and an Adjunct Professor, Mälardalen University, Västerås, Sweden.