CAMERA: Focussing on instrument based research

David Abramson[1]*, Jake Carroll*[1]*, Michael Mallon*[1]*, Aswin Narayanan*[1]*,*
Edan Scriven[1]*, Zane van Iperen*[1]

[1]**University of Queensland, Brisbane, AUSTRALIA,**
{firstname.lastname}@uq.edu.au

INTRODUCTION

Digital instruments such as microscopes, scanners and sequencers now underpin much scientific research, and are a critical element in many modern laboratories. They provide the ability to image and analyse objects across a huge range of modalities, for example optical, electron, magnetic resonance and x-ray imaging, to name a few, including devices such as gene sequencers that use a combination of imaging techniques and chemistry. Data from such instruments is rarely useful as raw data, and often needs to be processed and analysed using powerful high performance computers. Data also needs to be stored and archived for later use by the originating research team and their collaborators. While it is possible to move data manually between instruments, storage systems and computers, it is preferable to make access as transparent as possible to simplify the job of the researchers. Furthermore, while digital instruments have been common for some years, the exponential growth in data volume and velocity is challenging for computational infrastructure, and this demands innovative and powerful solutions be developed.

This paper discussed a new infrastructure called CAMERA, which facilitates the **CA**pture, **M**anag**E**ment, sto**R**age and **A**nalysis of data. While CAMERA defines a new framework, it leverages existing open source and commercial software and infrastructure extensively. At the University of Queensland specifically, CAMERA builds on and extends: the UQ Research Data Management platform (RDM) [4] which simplifies the task of requesting storage, the Metropolitan Data Caching Infrastructure (MeDiCI) [5] which simplifies the process of accessing data, and a variety of image repository stacks. CAMERA supports a complete life cycle for instrument gathered data, seamlessly rendering it on a range of instruments, cloud computers, desktops and high performance computers. Importantly, CAMERA encourages best-of-breed of data repositories and meta-data management systems without unnecessary data replication. As data grows exponentially, the latter minimises storage requirements without losing functionality.

A TYPICAL INSTRUMENT WORKFLOW

Figure 1 shows a typical instrument workflow in a modern laboratory setting. Data is captured on an instrument, and often needs to be pre-processed. The results are then uploaded, using a range of direct and indirect methods, to a repository, which extracts meta-data, and catalogues and organizes the data around experiments. Repositories, such as the Open Microscopy Portal (OMERO) [1] shown in Figure 1, are often Web enabled, providing powerful graphical user interfaces, but are not necessarily efficient communication pathways. While the data is indexed in the repositories, it is stored in archival storage systems, and is subsequently analysed using a range of computing platforms from high-powered workstations, cloud computers, and parallel high performance clusters..

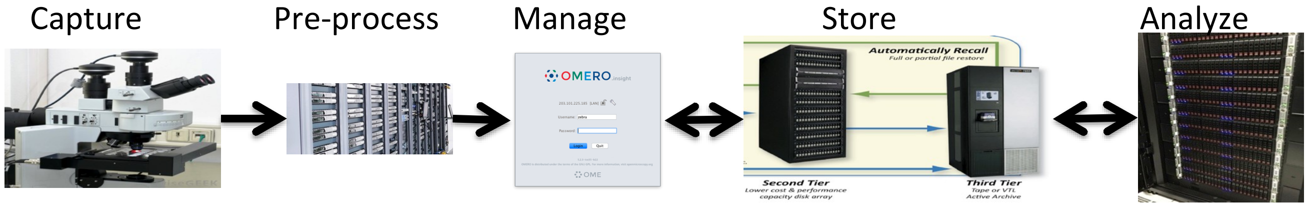
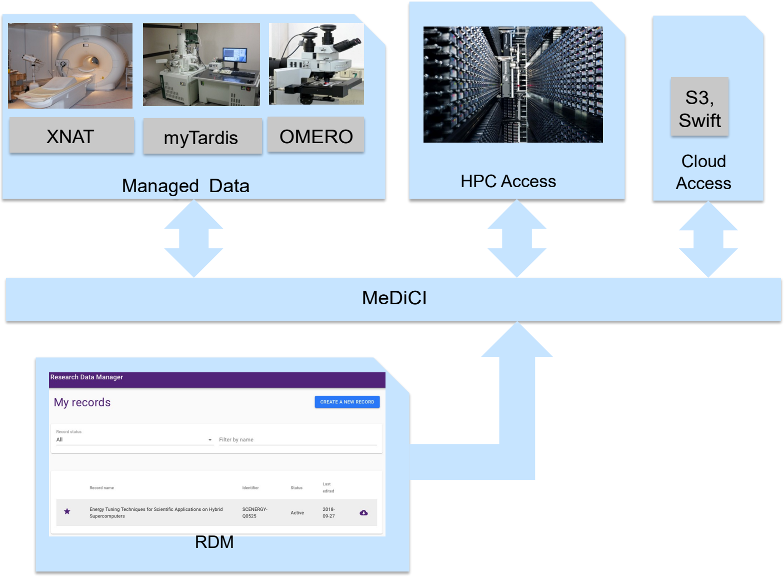

Capture
Pre-process
Manage
Store
Analyze
OMERO
Automatically Recall
Full or partial file restore
Second Tier
Lower cost & performance capacity disk array
Third Tier
Tape or VTL Active Archive

Figure 1: A typical instrument workflow

IMPLEMENTATION

As mentioned, it is not trivial to implement an efficient solution for instrument pipelines while keeping the system simple for researchers. CAMERA achieves this by leveraging powerful underlying technologies, such as high-throughput networks and storage systems while hiding this complexity from the users. Importantly, CAMERA users largely see the workflow as presented in Figure 1, without needing an understanding of the underlying mechanisms. CAMERA is not a single system, but an aggregate of the best-of-breed technologies. It defines a framework in which components that implement only part of the process can be integrated and inter-operated. This makes it possible to interconnect them end-to-end to support the workflow, making the system appear as a single coherent platform. For example, most repository systems assume that they completely manage experimental meta-data and data, and are free to store it in any way that optimizes the operations they support. However, such a world view can make it difficult to render data on a high-performance computer which expects a conventional POSIX file system, and usually, it is necessary to copy files in and out of the repository using often inefficient Web based protocols. CAMERA avoids this by providing multiple views of the underlying data structures – one that suits the repository, and another that suits a high-performance cluster.

Figure 2 shows how CAMERA achieves seamless inter-operation discussed above. It supports a number of managed data repositories; currently for optical microscopy we have implemented OMERO, and for multi-modal scanners such as MRI, PET-CT, etc we use MyTardis [3] and XNAT [2]. These repositories are the primary view of the data from the instrument and the operator, and facilitate meta-data management, search, sharing, and simple visualisation and processing tasks. Importantly, while the repository stack is executed on a QRIScloud node [7], data is actually stored on the MeDiCI fabric. MeDiCI stores a single copy of the data, avoiding unnecessary replication, but provides multiple access protocols and views, including NFS, SMB, parallel file system (NSD), OwnCloud [6] and OpenStack Swift [8]. MeDiCI makes it simple to expose a collection to a repository but later mount that same data on a HPC cluster, and while data movement might be necessary, it is organised transparently.



	Record name	Identifier	Status	Last edited
★	Energy Tuning Techniques for Scientific Applications on Hybrid Supercomputers	SCENERGY-Q0525	Active	2018-09-27

Figure 2: CAMERA architecture

Importantly, CAMERA leverages the UQ Research Data Manager (RDM), through which researchers request a data collection prior to their laboratory work, and link this collection to the repository. This means that the project level meta-data is captured once in RDM, and the University can manage the provenance of the data in the same way as all other research data. MeDiCI then facilitates the transparent access of the data throughout the analysis pipeline, and also provides mechanisms for sharing data with collaborators.

REFERENCES

1. https://www.openmicroscopy.org
2. https://www.xnat.org
3. http://www.mytardis.org
4. https://research.uq.edu.au/rmbt/uqrdm
5. MeDiCI, Abramson, D., Carroll, J., Jin, C. and Mallon, M., "A Metropolitan Area Infrastructure for Data Intensive Science", 13th IEEE eScience Conference, Auckland, New Zealand 24th – 27th October, 2017.
6. https://owncloud.org/
7. https://www.qriscloud.org.au
8. https://wiki.openstack.org/wiki/Swift