The CSIRO eResearch Workflow Service

Tim Ho, Joel Ludbey, Alfred Uhlherr
CSIRO Information Management & Technology, Clayton, Australia
Tim.Ho@csiro.au, Joel.Ludbey@csiro.au, Alfred.Uhlherr@csiro.au

Introduction

Over the past five years CSIRO has developed a comprehensive suite of eResearch services to support science workflows through investment in infrastructure, capability development, and software and systems technologies. The focus is on developing world-class eResearch tools and advanced facilities that provide a modern collaborative environment for facilitating multidisciplinary and interdisciplinary collaboration between CSIRO staff and their research and industry partners.

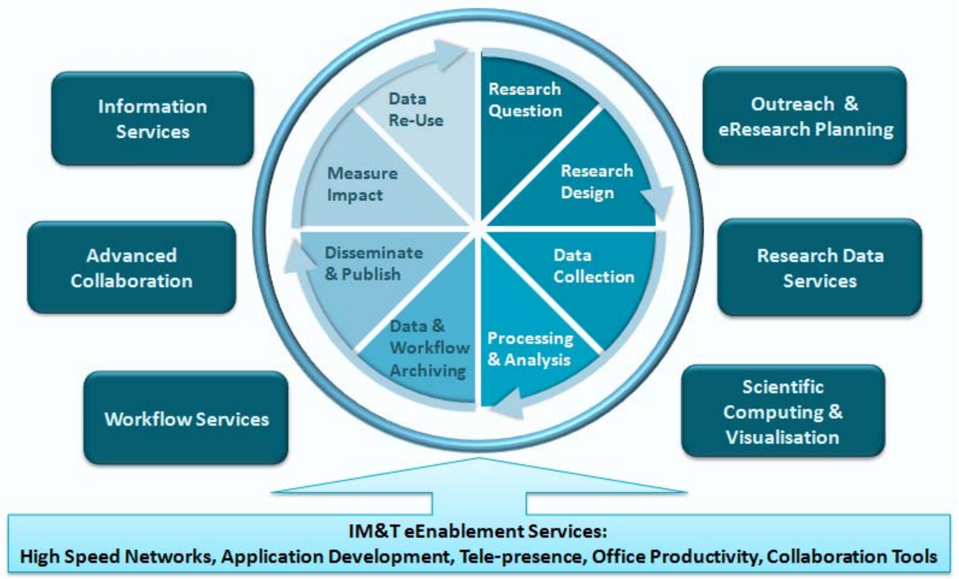
Information Services
Advanced Collaboration
Workflow Services
Data Re-Use
Research Question
Research Design
Data Collection
Processing & Analysis
Data & Workflow Archiving
Disseminate & Publish
Measure Impact
Outreach & eResearch Planning
Research Data Services
Scientific Computing & Visualisation
IM&T eEnablement Services:
High Speed Networks, Application Development, Tele-presence, Office Productivity, Collaboration Tools

Figure 1 CSIRO eResearch Services

Scientific Workflows

Scientific research is becoming increasingly compute- and data-intensive. In many domains, such as bioinformatics and environmental sciences, it is no longer possible for scientists to carry out their day-to-day research activities without the significant use of computational techniques. In addition, advanced modelling techniques, such as analysis tools that utilise MPI for parallel computations, enable (and encourage) researchers to construct complex distributed solutions that scale to many CPUs and/or nodes for solving very large research problems.

Such work often requires a series of structured activities and computations, which are typically invoked in a routine manner. For example, data sets created by scientific instruments are processed and verified by pre-processors (e.g. format translation), ensuring that the data can be used in other simulation programs for further analysis. In many cases, the computations can be very complex and structured with intricate dependencies.

Scientific workflows can be used to describe such tasks. They typically record details such as what software components are required, how to invoke those software components (e.g. compute resources), what intermediate data products are created, where the data should be stored, how the data may be used by other software tools (e.g. visualisation) and whether human intervention is required. Ideally, such workflows should be portable, reusable and reproducible.

Significantly, modern desktop computers can no longer provide sufficient computing capacity for complex simulations and modelling workflows. The advent of advanced application development techniques such as MPI, OpenMP and GPU programming has enabled researchers to build science processing pipelines that take advantage of the vast amount of compute and storage capacity available from High Performance Computing (HPC) systems.

CSIRO eResearch Services

CSIRO Information Management & Technology (IM&T) provides staff and affiliates with access to a range of eResearch services [1] that support scientific workflows and facilitate the transition from desktop computing to high-end computing infrastructure. These include:

- *Compute services* – enabling scientific modelling and analysis to be performed efficiently and effectively.
- *Data services* – supporting the end-to-end lifecycle of research data.
- *Visualisation services* – providing resources and expertise in visualising science data.
- *Software* – a comprehensive science software portfolio providing a large range of open source and commercial software products.

An emerging trend for these eResearch services is that researchers can be overwhelmed by the many facilities that are available to them, and have difficulties in identifying which resources are best fit to their research, or how to optimise their usage of these resources. Moreover, there is a growing need to capture data and workflow provenance to enable reproducible science, so that research results can be validated and knowledge can be reused and shared in the scientific community.

To meet these needs CSIRO now offers a specialised *eResearch Workflow Service* to staff and affiliates. Client engagement is a key component of this service, in which team members engage with researchers to understand their needs and deliver against them. Importantly, the team facilitates client access to other eResearch services through assisting users to develop fit-for-purpose and sustainable workflow solutions that draw from CSIRO's continuum of computing services.

As part of the Workflow Service, staff also promote the importance of capturing workflow provenance and provide advice and assistance on building manageable workflows using workflow management tools such as CSIRO's Workspace [2] and Galaxy [3]. We run workflow collaboration projects using a biannual call for proposals [4], in which we assign a workflow specialist to each accepted project, who will work with the project stakeholders to complete an agreed set of deliverables within the project timeline (typically 6 months). The goal of these projects is to help researcher create reusable, portable, reproducible and automated workflows.

Examples of workflow projects delivered to date include the following:

- Researchers from the Biosecurity Microscopy Facility at the Australian Animal Health Laboratories (AAHL) [5] now use an image processing workflow in Workspace, which automates the processing of electron tomographs of viruses that involves data intensive and highly repetitive operations.
- Creation of data processing and visualisation Workspace plugins for the Zebedee scanner [6], allowing researchers to create workflows that visualise and quantify forest regions in regards to biomass stocks, vegetation densities and distributions.
- A Galaxy pilot program offers an online data analysis platform for bioinformatics research. There are now more than 200 Galaxy users and more than 400 Galaxy tools made available through this platform. We are currently establishing a production Galaxy service that uses CSIRO's private cloud and HPC infrastructure.

In this presentation we describe how the eResearch Workflow Service addresses research requirements through established eResearch services, helps researchers develop reproducible workflows, and enables science outcomes through cross-team collaboration with researchers from CSIRO and partners.

References

1. CSIRO IM&T Scientific Computing. Available from https://wiki.csiro.au/display/ASC/, accessed 27 May 2015.
2. CSIRO Workspace. Available from: https://research.csiro.au/workspace/, accessed 22 May 2015.
3. The Galaxy Project. Available from: https://galaxyproject.org/, accessed 22 May 2015.
4. eResearch Collaboration Projects. Available from https://wiki.csiro.au/x/CYEiGw, accessed 27 May 2015.
5. Australian Animal Health Laboratory. Available from: http://www.csiro.au/en/Research/Facilities/AAHL, accessed 25 May 2015.
6. Zebedee: handheld laser scanner. Available from: http://www.csiro.au/en/Research/DPF/Areas/Autonomous-systems/Mapping/Zebedee, accessed 25 May 2015.

About The Authors

Dr Tim Ho is the User Services Manager in CSIRO IM&T Scientific Computing, where he is responsible for leading and developing eResearch capabilities and support services in high performance computing, cloud computing, workflow management and software portfolio management and deployment. He is also the manager of the Galaxy project, which will deliver an online data analysis platform for the bioinformatics community in CSIRO. He was previously an HPC application specialist responsible for profiling, optimising and parallelising science applications. Prior to joining IM&T he worked for a number of software development projects, including a Boeing corrosion modelling framework, a scalable data abstraction layer for distributed applications and the Guard parallel relative debugger, which later became Cray's comparative debugger. He has also introduced an active data model, which addresses the management of the lifecycle of derived data. Tim holds a Ph.D. in the areas of high performance computing and data management from Monash University.

Joel Ludbey is a scientific workflow specialist working within the Scientific Computing User Services team in CSIRO IM&T. He currently operates as the main lead in several projects under CSIRO's eResearch Collaboration Program, providing his expertise in helping scientists manage and develop scientific workflows that enable them to make intelligent use of the scientific computing resources available to them both internal and external to CSIRO. He has previously worked with the CSIRO IM&T Data Services team, providing valuable support to scientists in the management, transformation and preservation of scientific datasets as well as the management and deployment of several data transfer services. Prior to this he worked as a system administrator for the Australian Research Collaboration Services (ARCS), helping to provide and support several services aimed at provisioning a distributed compute and data management grid across multiple high performance computing sites in Australia. Joel holds an honours degree in computer science and software development from Deakin University.

Dr Alf Uhlherr is the Senior Manager for Scientific Computing Services in CSIRO Information Management & Technology. He completed his B.Sc. (Hons) and Ph.D. at Monash University in 1987 and 1991 respectively, majoring in physical chemistry. Following a postdoctoral fellowship at University of Cambridge, he has worked for over twenty years at CSIRO in Melbourne. His research has focused on computational methods for analysis of soft matter and development of new materials. In his present role Alf is responsible for delivering a wide range of eResearch services to internal and affiliate users, including high performance computing, data intensive computing, advanced visualisation, software applications and access to partner facilities such as NCI, MASSIVE and QCIF. His current governance responsibilities include serving on the National Computational Merit Allocation Committee (NCMAC), facility director for the Pawsey Supercomputing Centre, and CSIRO representative on the Australian eResearch Organisations (AeRO) committee.