4th International Digital Curation Conference
December 2008

ARCHER – e-Research Tools for Research Data Management

Steve Androulakis[5], Ian Atkinson[2,4], Ashley M Buckle[5], David Groenewegen[1,2,3], Nick Nicholas[2,6], Andrew Treloar[1,2], Anthony Beitz[2,7]

[1]ANDS, [2]ARCHER, [3]ARROW, [4]James Cook University e-Research Centre,

[5]Department of Biochemistry and Molecular Biology, Monash University, [6]Link Affiliates, [7]Monash e-Research Centre

October 2008

Abstract

With new scientific instruments growing exponentially in their capability to generate research data, new infrastructure needs to be developed and deployed to allow researchers to effectively and securely manage their research data from collection, publication, and eventual dissemination to research communities. In particular, researchers need to be able to easily acquire data from instruments, store and manage potentially large quantities of data, easily process the data, share research resources and work spaces with colleagues both inside and outside of their institution, search and discover across their accessible collections, and easily publish datasets and related research artefacts. The ARCHER Project has developed production-ready generic e-Research infrastructure including: a Research Repository; Scientific Dataset Managers (both a web and desktop application); Distributed Integrated Multi-Sensor and Instrument Middleware; and a Collaborative Workspace Environment. Institutions can selectively deploy these components to greatly assist their researchers in managing their research data.

The 4th International Digital Curation Conference taking place in Edinburgh, Scotland over 1-3 December 2008 will address the theme *Radical Sharing: Transforming Science?*.

Introduction

The need for the data management infrastructure is being felt ever more acutely in the e-research community due to:

- The quantity of scientific data increasing exponentially, challenging researchers to keep track of it all;
- This large quantity of electronic data creating new challenges for collaboration;
- A much greater expectation for online publishing of data and verifiability of experiments;
- Concerns about security and privacy in many disciplines of e-research; and
- A significant push to streamline the workflows of e-research by providing centralised, persistent, and reliable storage.

The ARCHER Project (Australian ResearCH Enabling enviRonment) was funded by the Australian Government's Department of Education, Science and Technology in 2006 as an attempt to begin to address these concerns. By providing a cogent, data-centred view of the e-research enterprise. ARCHER allows researchers the flexibility of iterative and heuristic workflows and ease of collaborative management, and ensures that data remains well curated and publication-ready, with appropriate metadata, provenance, and authorisation.

ARCHER has produced a suite of tools developed jointly by Monash University, James Cook University, and the University of Queensland; drawing on, integrating and extending existing open source toolkits. It provides infrastructure to assist researchers in collecting, managing, storing, collaborating on, and publishing scientific data. The Project builds on the earlier DART project [1,2,3,4,5], taking selected proof-of-concept tools and moving them to production. ARCHER will complete its tool suite in September 2008, and is making its products and source code openly available.

In this paper, we take a closer look at the features and benefits of the ARCHER suite of data management tools, identify ARCHER's relationship with the Australian e-Research environment, and demonstrate how its components can be loosely coupled with other e-Research data management components to provide a comprehensive data management solution from data collection, through to publication, and to its eventual dissemination within a research community.

The ARCHER Toolkit

Overview

The ARCHER initiative has developed a suite of open-source production-ready generic e-Research infrastructure components to provide better management of research data, including:

- Distributed Integrated Multi-Sensor and Instrument Middleware

(DIMSIM) - concurrent data capture and analysis, and telemetry

- ARCHER Research Repository, which introduces enhancements to SRB (Storage Resource Broker)
- XDMS – web-based research data manager and curator
- HERMES – desktop client research data manager and file transfer agent
- Collaborative Workspace Development Tool (based on PLONE), for creating e-Research Portals.

Figure 1. demonstrates how these components may be integrated.

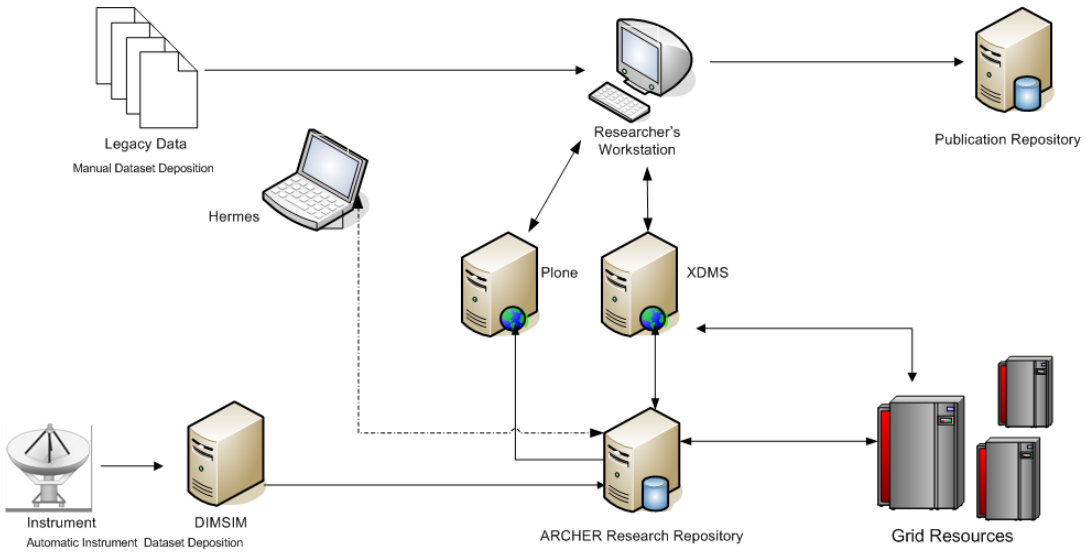

Figure 1 Potential architecture for ARCHER's deployment

DIMSIM (Distributed Integrated Multi-Sensor & Instrument Middleware)

New scientific instruments are collecting research data at phenomenal rates, and conventional practices, such as storing the collected research data on CDs or portable hard drives, will not suffice to ensure long term storage and management. Other potential challenges include: dealing with complex and distributed instruments; determining the status of a remote experiment; transferring data from a remote instrument to the desired data store; and starting an analysis while the experiment is still running.

DIMSIM solves all of these problems, and allows multiple sensors to be easily integrated. It is built on CIMA (Common Instrument Middleware Architecture) [6], which allows instruments to be more easily accessible over a network. This in turn enables: direct deposition of collected research data into a network data store, without human intervention; concurrent analysis; and remote telemetry. By having DIMSIM

deposit research data directly into a large, reliable, and secure institutional research repository, many storage concerns are alleviated. If the institutional Research Repository also supports rich metadata, then curation can begin at collection time, improving curation quality and potentially reducing its costs. Figure 2 shows a snapshot of live telemetry being captured by DIMSIM, during an X-ray crystallography experiment.

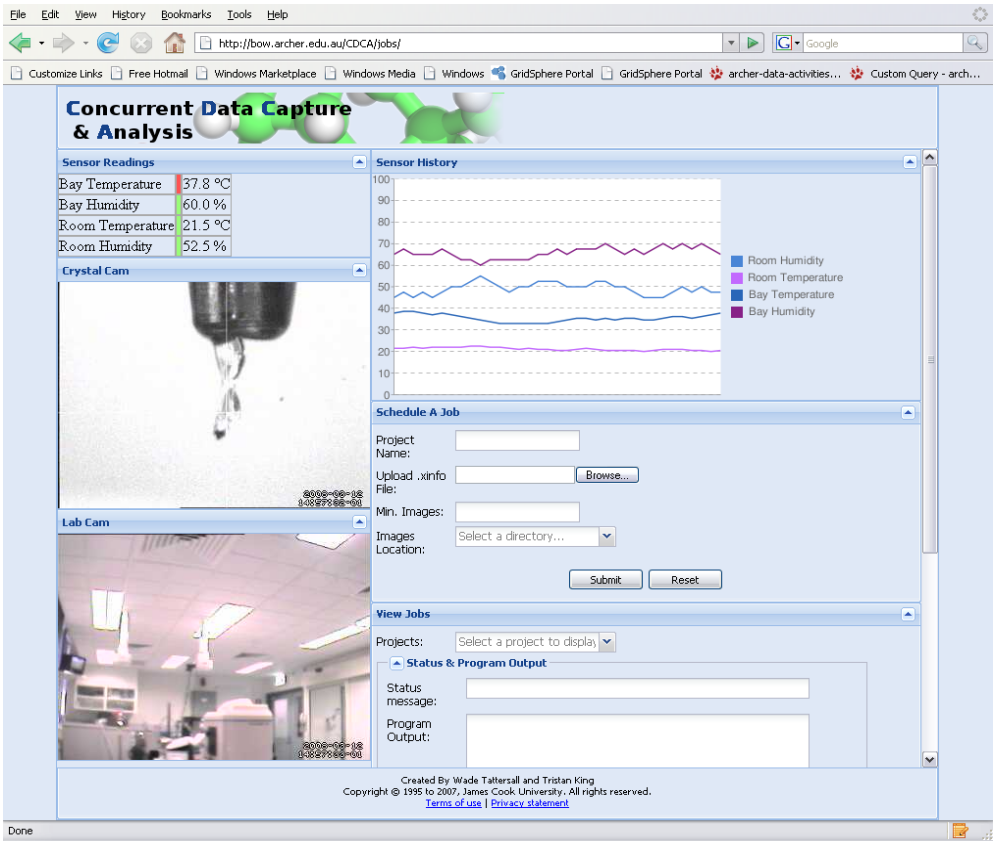

Figure 2 DIMSIM Screenshot

ARCHER Research Repository

A Research Repository (for experimental data) needs to be: secure; reliable; able to cope with large and numerous datasets; able to provide support for rich discipline-specific metadata; support good data management practices; and be easily accessible.

SRB was chosen as the foundation for ARCHER's Research Repository because of its demonstrated ability to deal with large and numerous datasets. One key limitation however was its metadata repository (MCAT), which only supports key/value pairs. ARCHER considered this inadequate to the requirements for proper data curation and so augmented with this with an additional metadata store called iCat. This is a relational database, which, although not the most flexible solution, was chosen as the new store for the research data's metadata mainly because of its ability to scale.

A schema was required which provided at the top levels a highly structured and scientific discipline agnostic approach, while allowing for additional discipline-

specific metadata. CCLRC's (Council of the Central Laboratory of the Research Councils – now the STFC (Science and Technology Facilities Council)) Scientific Metadata Model was identified as the most suitable solution, providing a rigid structure of Project(Study) ⇨ Experiment(Investigation) ⇨ DataSet ⇨ DataFile in the top levels (see Figure 3.), and offering discipline-specific schemas associated with Samples, DataSets, and DataFiles.

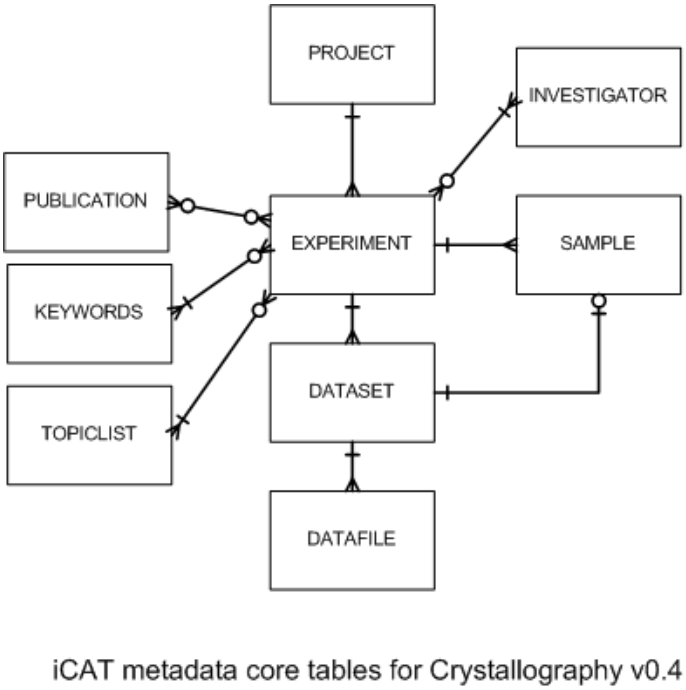

Figure 3 CCLRC Scientific Metadata Model

XDMS (Crystallography Data Management System)

XDMS is the web based data management component of the ARCHER suite of e-research infrastructure tools, and sits on top of ARCHER's Research Repository. It promotes good data management practises and provides researchers with data access, data deposit, data export, curation facilities, search and discovery services, and the ability to associate persistent identifiers with datasets.

XDMS provides two levels of metadata support: a generic core metadata profile, applicable across disciplines, using the CCLRC Scientific Metadata Model; and a domain-specific metadata profile, which is user-configurable, and editable by the ARCHER Metadata Editor. Metadata associated with the various levels within the CCLRC metadata hierarchy, including discipline-specific metadata, can be searched and browsed, enabling researchers to easily locate objects and collections.

XDMS provides support for the deposition of research data, and can automatically extract a datafile's metadata from its header and associate it with the deposited datafile. Due to connection timeout issues inherent in all web browsers, ingestion of large quantities of research data via HTTP is not practical, and deposition of multiple datafiles is better handled by ARCHER's desktop client data management component

HERMES.

XDMS can export research data in both native file format and packaged into a METS format. It can also deposit the METS package directly into a Fedora based Publication Repository. Publication repositories are where data is made available to a general audience rather than the collaboration group [7], with a guarantee of long-term persistence. These are typically provided by institutionally supported repositories, and use technologies such as Fedora and DSpace rather than SRB; so packaging is necessary for transferring the data across. As with deposition, exporting large quantities of research data is better handled by HERMES.

XDMS allows persistent identifiers to be generated for research datasets using CNRI's (Corporation for National Research Initiatives) Handle technology, enabling researchers to easily share references to their Datasets with selected colleagues.

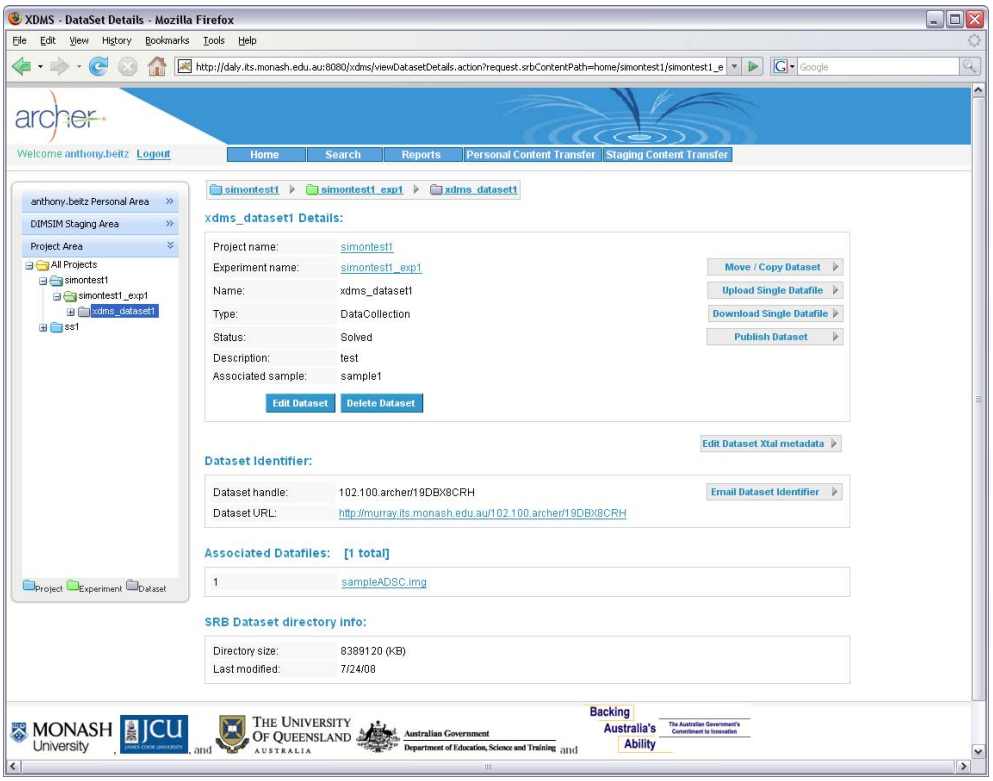

Figure 4 XDMS Screenshot

Hermes

HERMES is ARCHER's desktop client data management tool, and can sit on top of ARCHER's Research Repository. It functions as a desktop file browser, which allows browsing of local drives, Samba, SRB, GridFTP, FTP and Secure FTP file systems. HERMES allows upload and download of large files singly and in batches. This coupled with its support for a wide range of storage solutions make it ideal as a file transfer agent.

ARCHER Collaborative Workspace Development Tool

Generally, each research discipline within an institution has its own unique set of needs for e-Research technologies. These needs may also vary within the same discipline in different institutions, making it practically impossible to produce a generic e-Research portal which satisfies all researchers. Therefore some level of customisation is usually necessary.

ARCHER's approach was to develop generic e-Research components which could be selectively coupled together, which would share the same authentication system, and to adopt a portal development tool that was easily adaptable and which enabled customisation of the information architecture, including research data stored in ARCHER's Research Repository.

PLONE is a popular CMS (content management system) and was extensively customised by ARCHER to support e-research collaboration anchored to research data and projects. Part of this customisation is a PLONE plug-in, which enables PLONE to access the ARCHER Research Repository. This in turn enables links, comments, blog, and discussions to be made on stored research data.

ARCHER's Place in the Australian e-Research Environment

Australia's e-Research environment is influenced by PFC (Platforms For Collaboration), which is part of NCRIS (National Collaborative Research Infrastructure Strategy). The PFC contains a number of services, two of which can be associated with ARCHER's tools. In particular, ARCS (Australian Research Collaboration Service) and ANDS (Australian National Data Service).

This section provides a brief overview of these services and describes ARCHER's synergies with each of them.

ARCS

ARCS's objective is to provide long-term eResearch support services for the Australian research community with a particular focus on interoperability and collaboration infrastructure, tools, services and support. It offers services like:

- Video collaboration;
- Web based collaboration;
- Research Data Fabric; and
- Remote Instrumentation and Sensor Network Activities.

ANDS

The overall objective of this service is to improve researchers' practises in managing their research data, predominantly by:

- Transforming collections of Australian research data into a cohesive

network of research repositories;

- Assisting Australian research data managers to become experts in creating, managing and sharing research data under well formed and maintained data management policies;
- Increasing the amount of research data that is routinely deposited into stable, accessible and sustainable data management and preservation environments;
- Enabling researchers to find and access any relevant data in the Australian 'data commons' ; and
- Facilitating the sharing of Australian data to support international and nationally distributed multidisciplinary research teams.

ARCHER's Synergies with ARCS

One of the ARCS's collaborative tool offerings is ARCHER's enhanced version of PLONE. Its customised plug-ins make it well suited to the developing e-Research environment, and allows PLONE collaborative tools to directly access and link to research data stored in an ARCHER Research Repository, enabling researchers to easily collaborate around their research data.

Through ARCHER's work in research repositories, it has contributed to the development of the ARCS's Data Fabric. The ARCS Data Fabric is intended to make it easy for researchers to store and share their data outside their usual institutional confines. This is encouraging new collaborations to form, and providing new research opportunities.

ARCS has also adopted HERMES as the front line tool in providing access to its Data Fabric from a client desktop. Its interface's support for multiple file systems makes it very easy for researchers to move their data from one digital repository to another.

ARCHER's Synergies with ANDS

ARCHER's relationship with ANDS is that it provides software components to enable Australian researchers to better manage their research data, therefore hopefully increasing the amount of data being stored into secure, reliable, and sustainable repositories. ANDS hopes that this will help to facilitate the sharing of Australian research data, both locally and internationally.

Research Data Management – Gluing the Pieces Together

This section describes how ARCHER's components may be coupled with additional data management components, from the ARROW and TARDIS projects to provide researchers with a comprehensive data management solution.

ARROW

Australian Research Repositories Online to the World (ARROW) is a consortium consisting of Monash University (lead institution), together with the University of New South Wales, Swinburne University of Technology, and the National Library of Australia. Its objective is to identify and test software or solutions to support best practice institutional digital repositories that would contain e-prints, electronic theses, e-research and electronic journals. From this project has come the ARROW Repository (with a commercial offering known as VITAL, developed by VTLS Inc.). The ARROW repository is an institutional publication repository built on the Fedora open source repository platform [8].

TARDIS

TARDIS (The Australian Repositories for Diffraction Images) [9] is a multi-institutional collaborative venture consisting of Monash University (lead institution), together with the University of Queensland, Institute for Molecular Bioscience, the University of Melbourne, St Vincent's Health Melbourne, Bio21 Institute, ARROW, ARCHER, the Australian National University, the University of Sydney, Australian Partnership for Sustainable Repositories, eCrystals Federation Project , and the University of Southampton. It aims to facilitate the archiving and sharing of raw X-ray diffraction images, collectively known as a dataset. It has developed a number of client desktop tools which assist in the preparation and deposition of a collection of raw crystallographic datasets into an institutional publication repository. It also provides a community portal "TARDIS" which harvests metadata of published crystallographic datasets from registered institutional publication repositories, indexes the collected metadata, and then provides a federated search across institutional repositories.

Modelling the Curation and Migration of Research Data from Collection to Dissemination

As data is collected, shared, published, and disseminated; it is migrated through a range of conceptual domains [7,10]. Each of these domains can be defined by a set of attributes which describes the data objects and the repositories that store them (e.g. accessibility of the data and richness of the metadata). The boundary between these domains can be referred to as curation boundaries.

There are four domains:

- Private Research Domain, where data is initially collected and is generally only shared amongst a tight-knit research team;
- Shared Research Domain, where the team may open up the access to the research data to a select group of researchers (e.g. reviewers assessing pre-published research data);
- Public Domain, where the data is relatively open to the public; and
- Community Domain, where selected data is made available to a community for dissemination and further collaboration.

This model is explained further in figure 5. below.

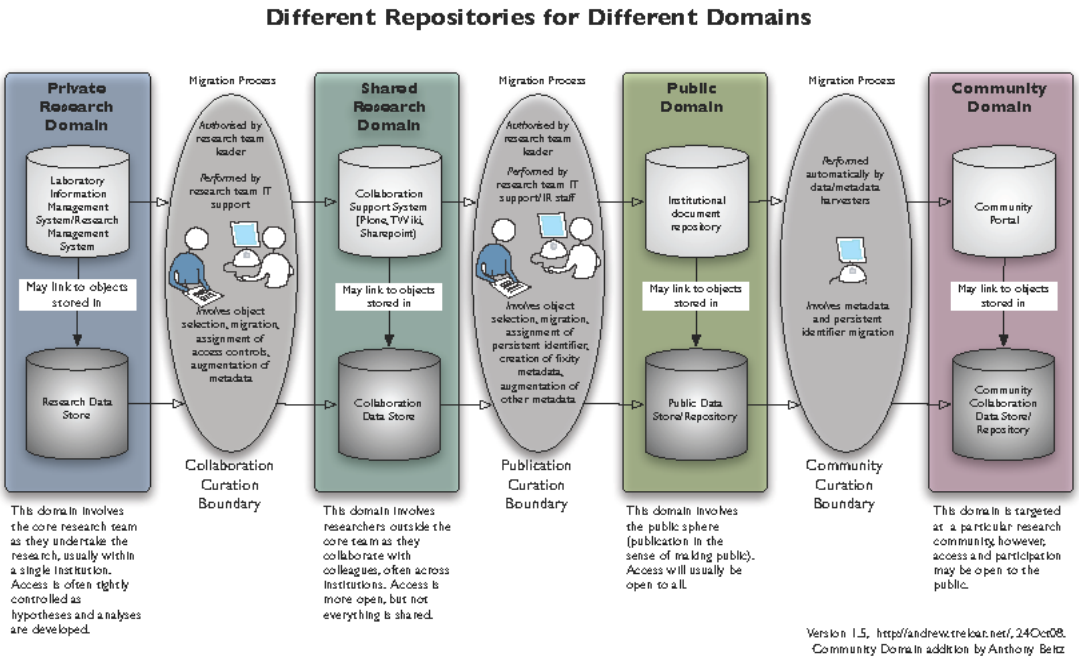

Figure 5. A Model for the Curation and Migration of Research Data

Combining ARCHER, ARROW, and TARDIS

Based on the model presented in Figure 5., ARCHER's components can be combined with an ARROW's repository and TARDIS in the following way:

1. ARCHER's components can be combined to perform management of the research data. Filling the roles outlined for the Private and Shared Research Domain.

2. Data is migrated into an ARROW repository utilising componentary common between ARCHER's XDMS and TARDIS's client desktop deposition tools.

3. The ARROW repository performs the role defined for the Public Domain.

4. The ARROW repository's metadata is made accessible, and its and other registered repositories' metadata is harvested by TARDIS.

5. TARDIS's Portal then disseminates community-relevant data to its community, filling the role for the Community Domain.

Together, ARCHER, ARROW, and TARDIS provide a comprehensive research data management solution, which can greatly assist researchers in managing their research data

Conclusions

As a result of the work performed in ARCHER, researchers in general are now much closer to having: a place to collect, store and manage experimental data; software tools focused on management of data and information; being able to easily customise a collaborative and adaptable portal web site relevant to their research field; having standardised and secure methods of storing, accessing, and analysing research results; and being easier to collaborate and share research datasets and information.

ARCHER has addressed many key issues in e-Research data management. It is enabling researchers to keep better track of their scientific data by organising it into intuitive and generic structures described by the CCLRC Scientific Metadata Model, and by making the research repository easily searchable. It is alleviating issues around the collaboration of large datasets by: enabling the storage of large research datasets, adopting a data-centric view; and by providing an initial set of collaborative tools that can be directly associated with the research data. It is also protecting the privacy and security of research data by limiting the access to a project team's research data.

Finally, the combination of ARCHER, ARROW, and TARDIS together effectively demonstrate how crystallographic research data can be well-managed and well-curated from the research data's collection, from an instrument, to its eventual publication and dissemination to the crystallographic research community, via the TARDIS portal.

Acknowledgements

ARCHER was funded by the Australian Commonwealth Department of Education, Science and Training (DEST), through the Systemic Infrastructure Initiative (SII), a part of Backing Australia's Ability – An Innovation Action Plan for the Future. We would also like to acknowledge the co-operation and support of Monash University's Biochemistry department.

References

[1] Faux N, Beitz A, Bate A, Amin AA, Atkinson I, Enticott C, Mahmood K, Swift M, Treloar A, Abramson D, Whisstock JC, Buckle AM, "eResearch Solutions for High Throughput Structural Biology", *3rd IEEE International Conference on e-Science and Grid Computing*, Bangalore, India, 2007. DOI 10.1109/e-Science.2007.43DOI 10.1109/e-Science.2007.43

[2] Ian M. Atkinson, Anthony Beitz, Ashley Buckle, Andrew Treloar, "An X-Ray Crystallography Case Study: Using the DART Toolkit to enable more effective eResearch", *APAC Conference Exhibition Australian Partnership for Advanced Computing*, Perth, 2007.

[3] Andrew Treloar, "The Data Acquisition, Accessibility, Annotation and e-Research Technologies (DART) Project: Supporting the complete e-Research Lifecycle", *Proceedings of UK e-Science Programme All Hands Meeting 2007 (AHM2007)*, Nottingham, September 2007.

[4] Andrew Treloar, "DART: Building the new collaborative e-research infrastructure", *Proceedings of Educause Australasia 2007*, Melbourne, April 2007.

[5] Andrew Treloar, "The Dataset Acquisition, Accessibility, and Annotation e-Research Technologies (DART) Project: building the new collaborative e-research infrastructure", *Proceedings of AusWeb06, the Twelfth Australian World Wide Web Conference*, Southern Cross University Press, Southern Cross University, July 2006.

[6] Ian M. Atkinson, et al, "Developing CIMA-Based Cyberinfrastructure for Remote Access to Scientific Instruments and Collaborative e-Research", *Proceedings of 5th Australasian Symposium on Grid Computing and e-Research*, Bendigo, Jan 2007.

[7] Andrew Treloar & Catherine Harboe-Ree, "Data management and the curation continuum: how the Monash experience is informing repository relationships", *Proceedings of VALA 2008*, Melbourne, February 2008.

[8] Lagoze, C., Payette, S., Shin, E. and Wilper, C., "Fedora: An Architecture for Complex Objects and their Relationships", International Journal of Digital Libraries: Special Issue on Complex Objects, Volume 6, Issue 2, April 2006.

[9] S. Androulakis, J. Schmidberger, M. A. Bate, R. DeGori, A. Beitz, C. Keong, B. Cameron, S. McGowan, C. J. Porter, A. Harrison, J. Hunter, J. L. Martin, B. Kobe, R. C. J. Dobson, M. W. Parker, J. C. Whisstock, J. Gray, A. Treloar, D. Groenewegen, N. Dickson and A. M. Buckle, "Federated repositories of X-ray diffraction images", *Acta Cryst.* (2008). D64, 810-814 , July 2008, [doi:10.1107/S0907444908015540]

[10] Andrew Treloar, David Groenewegen, Catherine Harboe-Ree, "The Data Curation Continuum - Managing Data Objects in Institutional Repositories", D-Lib Magazine Vol. 13 Num. 9/10, Sept/Oct 2007.