



FINAL REPORT DMQ5 WORK PACKAGE

Improve the intelligence of the storage framework.

Version 1.0, MAY 2007

Lead Investigator: Tristan King

Prepared by:

Tristan King tristan.king@jcu.edu.au

Franz Eilert frank.eilert@jcu.edu.au

School of Maths, Physics and Information Technology
Faculty of Science, Engineering and Information Technology
James Cook University

Executive Summary

This report presents the findings of research undertaken as a part of the DART (Dataset Acquisition, Accessibility and Annotation e-Research Technologies) project, specifically for Work Package DMQ5, 'Improve the intelligence of the storage framework.' This work package is extremely closely coupled to the DMQ2 Work Package, 'Ensuring effective and reliable connection of selected instruments and sensors to storage repositories (SRB) via CIMA middleware, and efficient use via changed work practices.'

The milestones for this work package have been achieved whilst also creating a significant component, the Data Manager, which is vital for integration with the Distributed Integrated Multi-Sensor and Instrument Middleware (DIMSIM), planned for the ARCHER project. The two key milestones were achieved as follows:

- *Triggers built into Storage Resource Broker (SRB) source stream.* - The initial focus for solving this problem was by using the workflow system Kepler. This solution triggered events when changes were made in SRB using a Kepler actor. This method was flawed as files often appeared in the SRB collection before the process writing that file to the collection had completed. Since SRB uses a standard database, it was decided that the standard database triggers could be used to perform the trigger functions. Some Proof of Concept triggers were implemented in Postgresql.
- *Production Implementation* - Production triggers have been implemented in Postgresql as part of the implementation of the X-ray Crystallography demonstrators, JCU And Indiana Instrument Service (JAINIS). Triggers will be implemented once live data becomes available from the ReefGrid project, DMQ1 and DMQ2.

A third milestone was achieved that integrated DMQ2 DART components with SRB utilising Kepler and was a result of the investigations performed in the initial milestones.

- *Kepler Workflow Production* - The DMQ2 work package required a set of triggers and workflow to ensure data was move into SRB. It was determined that a Kepler workflow would be an ideal method to create a Data Manager. A significant amount of effort was spent designing workflows for data management and building new actors to suit the requirements of the task. A significant investment was made in scoping the instruments and assessing work cases.

The project achieved its milestones and exceeded requirements by implementing production quality workflow solutions in JAINIS. It is recommended that development of Kepler workflow continue for integration with both Instrument Middleware, DIMSIM, and for Grid based processing applications be continued by the Archer project to bring value to the scientific community. Scientific workflow provides great value to any collaborative environments by automating previously labour intensive and repetitive tasks while dealing with institutional repositories and applications.

Table of Contents

1	INTRODUCTION	4
1.1	ORGANISATION OF THIS REPORT	4
2	PROJECT MILESTONES	5
2.1	AGILE SOFTWARE DEVELOPMENT	5
3	PROJECT OUTCOMES	6
3.1	BUILDING TRIGGERS INTO SRB SOURCE STREAM - SRB DIRECTORY POLLING	6
3.2	BUILDING TRIGGERS INTO SRB SOURCE STREAM - DATABASE TRIGGERS	6
3.3	KEPLER WORKFLOW PRODUCTION	6
3.4	JAINIS	7
3.5	ARCHITECTURE	8
3.6	PORTAL DEVELOPMENT	8
3.7	ERROR FIXING	8
4	RECOMMENDATIONS	9
5	ARCHIVAL STORAGE OF PROJECT DELIVERABLES	10
6	PUBLICATIONS	11
7	TERMS OF REFERENCE	11
7.1	GLOSSARY	11
8	REPORT SIGNOFF	12

1 Introduction

The original DMQ5 work package project goal was to 'Improve the intelligence of the storage framework.' The initial key elements were to build triggers into the database and then create production ready triggers. Later the work package's focus moved more onto Kepler orientated tasks including the JAINIS system.

The goal translated into two milestones:

- *Triggers built into SRB source stream.* - The initial focus for solving this problem was by using the workflow system Kepler, and
- *Production Implementation* - Production triggers being implemented in production environments such as the X-Ray Crystallography demonstrator, JAINIS, and the sensor network located on the Great Barrier Reef, ReefGrid.

The initial two milestones were achieved, although production implementation of the triggers required applications which had not been fully developed until later in the project. With the skills developed trying to utilise the Kepler workflow package to deliver triggers, a third milestone was incidentally achieved.

- *Kepler Workflow Production* - The DMQ2 work package required a set of triggers and workflow to ensure data was move into SRB. It was determined that a Kepler workflow would be an ideal method to create a Data Manager.

Kepler became a key element for the project effort at JCU as instruments and sensors and their accompanying actuators are essential to the conduct of scientific research. In many cases they provide observations in electronic format and can be connected to computer networks with varying degrees of remote interactivity. These devices vary in their architectures and type of data they capture and may generate data at various rates. The data manager and workflows developed in Kepler is a key element in automating tasks between the instruments and final results for researchers.

1.1 Organisation of this report

The purpose of this report is to detail the work undertaken in relation to the development of the intelligence in the storage framework and the subsequent utilisation of the knowledge gaining into workflow to provide integration in a number of areas of the DART project.

The report covers these distinct areas:

1. Investigation of Triggers and Production Implementation
2. Kepler development
3. JAINIS

2 Project Milestones

The following core milestones were defined for the DMQ5 work package in the original project specification:

- Triggers built into SRB source stream, and
- Production implementation.

Realistically a third milestone was achieved that integrated DMQ2 DART components with SRB utilising Kepler and was a result of the investigations performed in the initial milestones.

- Kepler Workflow Production.

The genesis of the third milestone was due to the agile methodology utilised at JCU. This is described below and the results of the project milestones are discussed in section 3, Project Outcomes.

2.1 Agile Software Development

The initial milestones did not really provide a comprehensive set of requirements and the milestones were realistically achieved early in the time line without solving the fundamental problems related to Data Management. Data management was not explicitly part of the work package but is essential to a number of aspects of the DART project.

It was decided that Kepler workflow would be the appropriate product to provide a data management service, Data Manager. After attempts to use it as a database trigger, the use of Kepler to solve data management and later x-ray crystallography data processing workflows became a prime focus of this work package. DMQ5 discoveries were incorporated into the development of the JAINIS product, based on the CIMA framework. The use of a workflow product fast tracked the storage services to SRB and combined with the Instrument Middleware in the JAINIS product, drew the attention of the Institute of Molecular Biology, University of Sydney, as well as Monash University.

3 Project Outcomes

The project goal was to ‘Improve the intelligence of the storage framework.’ This was further extended when the value of the Kepler workflow engine for various aspects of the DART project was discovered. The project outcomes as related to each milestone including the extra milestone that resulted from the completion of the first milestones is detailed below.

3.1 Building Triggers into SRB Source Stream - SRB Directory Polling

The initial focus for improving intelligence of the storage framework problem was by using the workflow system Kepler. This solution triggered events when changes were made in SRB was performed using a Kepler actor. The actor polled a specific SRB collection, and triggered a workflow to be executed when a new file was added to that collection, using that file as input to the workflow.

This method was flawed as files often appeared in the SRB collection before the process writing that file to the collection had completed, and occasionally the workflow would be executed with an incomplete file. The polling method was also very resource heavy.

3.2 Building Triggers into SRB Source Stream - Database Triggers

Since SRB uses a standard database (the MCAT, generally run under Oracle or Postgresql) to store its file allocation table, it was decided that the standard database triggers could be used to perform the trigger functions.

Some Proof of Concept triggers were implemented in Postgresql which were able to successfully notify other systems of new and modified files in SRB. Specific fields in the tables were able to be used to detect when a file was complete, proving definite advantages over the polling method.

One problem that arose was how to get the events out of the database system. Since database triggers are generally built to perform other changes to the database itself rather than other systems, this proved a challenge. The Proof of Concept model used the PL/Python scripting language to create XML-RPC messages which could notify a server of the changes that occurred.

This seemed to work well, but more research needs to occur to discover appropriate ways to distribute these messages to multiple clients quickly and security.

3.3 Kepler Workflow Production

The DMQ2 work package required a set of triggers and workflow to ensure data was move into SRB. It was determined that a Kepler workflow would be an ideal method to create a Data Manager. A significant amount of effort was spent designing workflows for data management and building new actors to suit the requirements of the task.

Kepler is a scientific workflow system out of the San Diego Supercomputing Centre. It provides a visual interface for creating scientific workflows using components called Actors which perform the specific tasks in a workflow. Kepler is written in Java and actors are created by writing classes that extend a base actor class. Below is a list of the Kepler actors that were developed during DMQ5.

1. **MATLAB actor**

This actor called MATLAB from within the workflow, passing to it a script to execute as well as initial variable values passed to the actor from other actors in the workflow. The outputs of the actor were defined as specific result variables at the end of the MATLAB execution.

This actor was put up for a code review by the Kepler board and accepted into their base cvs.

2. **Blogging actor**

This actor made it possible to post a blog entry from the workflow. It used the Blogger API to perform the blogging functions, and a simple custom keyword system to include variables passed to the actor to be included in the blog entry.

3. **SRB actors**

Kepler already came with actors to perform SRB operations, however when trying to use these actors many flaws arose. Mainly the fact that every actor that required access to an SRB server needed to be connected to a single actor which provided the connection object for them. This caused workflows to become cluttered with pointless connections between actors that distracted the actual flow of the workflow.

This was fixed by creating a standalone actor that handled the Server connections, and creating an abstract actor which provided a method to access the SRB server. Any actors requiring access to an SRB server then simply had to extend the abstract actor and call the method when it needed the connection object.

SRB actors were created using this method as required, thus the library is in no way complete. Also it may be required that the SRB server actor is modified to use Kepler's newly developed internal authentication system to authenticate to the SRB server.

3.4 **JAINIS**

For the DMQ2 CIMA project it was decided that a Kepler workflow would be a good way to create a Data Manager. A lot of effort was spent designing workflows for data management and building new actors to suit the requirements of the task.

A new entry point actor was required to receive the SOAP parcels that the CIMA architecture's separate components use to communicate. This was originally built using the Apache Axis libraries, but changed to Xfire to reduce the memory requirements of the workflow.

The DART project has created a JAINIS demonstrator with the following functionality:

- Collect experimental data from Rigaku Defractometers and store the data in SRB, and
- Workflows for running molecular analysis software, CCP4 without user interaction.

3.5 Architecture

After an initial period of development and attempted installation at the Institute for Molecular Biology (IMB) on a Rigaku Diffractometer, it became clear that the existing version of CIMA provided by Indiana University was unsatisfactory for deployment, as it contained various bugs, software incompatibilities (with SRB services developed as part of the work package) and hardware incompatibilities (X-Ray Diffractometer types).

The workflow started out as a proof of concept built with very basic requirements and using mostly existing components. This workflow was large and hard to modify, and as requirements evolved it was decided that more complex functionality was needed than what was currently available in Kepler. New actors were built to handle these requirements such as tracking multiple experiments with the same workflow and hosting experiment statuses for the JAINIS portal to access and the workflow became quite small with actors which preformed large parts of the flow, including SRB operations, purely in code. However, this was very un-workflow-like and these large actors needed to be expanded out into separate actors to make the workflow more workflow-like. The final version of the workflow expanded out all the large actors, adding reusable components (mainly the SRB actors) back into the workflow.

The JAINIS system has been deployed in real environments at IMB at UQ and the Monash medical centre. Both provided good testing environments and helps to find and fix problems with the initial deployment systems. See the DMQ2 report for more detail.

3.6 Portal Development

The original portal developed at Indiana University was written in GS portlets specifically for the Gridsphere portlet container. It was considered that the portal did not adhere to the standard portlet concept, so the portal was re-developed to adhere to portlet concepts by creating JSR168 portlets which are not container specific. The standard base portal has 3 individual portlets defined (multiple instances of these portlets can be used to compose a portal).

Sensors Portlet: This portlet is responsible for displaying the double data readings from the labjack. The portlet can currently displays temperature, humidity, and the weight on the scales (for liquid nitrogen levels).

Converted Image Portlet: Responsible for displaying converted diffraction images (for both Rigaku and Bruker), created by Kepler.

Live Video: This portlet projects live video feeds from the cameras in the laboratories.

3.7 Error Fixing

A very major consumption of time for this project was taken up by testing and debugging errors which occurred at deployment time, and could not be reproduced in simulation. Many other major hurdles were faced which hampered development, or produced inconsistent results during testing stages. These are listed below:

4 Recommendations

The project achieved its milestones and exceeded these except in case where there were dependencies on other work packages or organisations. Over the course of development many lessons were learned and because of the short time frame for development. The DMQ5 work package:

- Developed Proof of Concept SRB Triggers.
- Produced excellence in developing for and using the Kepler workflow system, and
- Developed workflow systems for use outside the initial use case for database triggers.

The project achieved its milestones and exceeded requirements by implementing production quality workflow solutions in JAINIS. It is recommended that development of Kepler workflow continue for integration with both Instrument Middleware, DIMSIM, and for Grid based processing applications be continued by the Archer project to bring value to the scientific community. Scientific workflow provides great value to any collaborative environments by automating previously labour intensive and repetitive tasks while dealing with institutional repositories and applications.

5 Archival Storage of Project Deliverables

The software produced by this work package is configuration of an existing set of application software. In the case of database triggers, these are triggers implemented within a specific type of database software, Postgresql or Oracle. PL/Python scripting language was used to create XML-RPC messages for Postgresql and a copy of the Python scripts can be zipped up and shipped to Monash University for archival storage. The Kepler actors are essentially Java classes that operate within the Kepler workflow environment. The Kepler actors will be zipped up and shipped to Monash University to be stored with the database trigger scripts and other software artefacts from the DART Project.

6 Publications

I.M. Atkinson, et al. Common Instrument Middleware Architecture: Extensions for the Australian e-Research Environment, 2nd IEEE International Conference on e-Science and Grid Computing

D.F. McMullen, et al. Toward Standards for Integration of Instruments into Grid Computing Environments, 2nd IEEE International Conference on e-Science and Grid Computing

I.M. Atkinson, et al. Developing CIMA based Remote Access for Collaborative e-Research, 4th Australasian Symposium on Grid Computing and e-Research

7 Terms of Reference

7.1 Glossary

Acronym	Definition
CIMA	Common Instrument Middleware Architecture
JAINIS	JCU And Indiana Instrument Service
SRB	Storage Resource Broker

8 Report Signoff

It is agreed between

Franz Eilert and Tristan King

and

Associate Professor Ian Atkinson

and

Dr Andrew Treloar

That the **Final Report Document** for the DART DMQ5 – ‘Improve the intelligence of the storage framework’, gives a full account of the work undertaken for the DART Project.

Franz Eilert

0401 710 272

Tristan King

07 4781 5645

- has been read and reviewed by all parties,
- shows that the DART DMQ5 – ‘Improve the intelligence of the storage framework’ has been completed satisfactorily,
- clearly outlines the deliverables stated in the DMQ5 requirements documentation have been met.

Dated this 29th day of May 2007

Signed
Chief Investigator
Associate Professor
Ian Atkinson

Signed
For and on behalf of DART
Project Director
Andrew Treloar