Integration of modern data management practice with scientific workflows

Neil E B Killeen\textsuperscript{1,2}, Jason M Lohrey\textsuperscript{3}, Michael Farrell\textsuperscript{4}, Wilson Liu\textsuperscript{1}, Slavisa Garic\textsuperscript{5}, David Abramson\textsuperscript{5}, Hoang Nguyen\textsuperscript{5} and Gary Egan\textsuperscript{6}

\textsuperscript{1} Centre for Neuroscience Research, University of Melbourne, Victoria, Australia
\textsuperscript{2} Information Technology Services, University of Melbourne, Victoria, Australia
\textsuperscript{3} Arcitecta Pty. Ltd, Victoria, Australia
\textsuperscript{4} Florey Neuroscience Institutes, Melbourne, Victoria, Australia
\textsuperscript{5} Faculty of Information Technology, Monash University, Clayton, 3800, Victoria, Australia
\textsuperscript{6} Faculty of Medicine, Nursing and Health Sciences, Monash University, Clayton, 3800, Victoria, Australia
Overview

• Introduction
• Tools & Technologies used
• Integration of those tools
• New Components
• Case Study and Results
  – Application and Computation results
• Conclusions and future work
Introduction

• Practical limitations with data handling processes
  – Poor data management
    • raw and processed
  – Poorly defined quality controls
  – Poorly defined processing workflows

• Researchers should aim to
  – Automate data capture
  – Automate processing via workflows
  – Integrate workflows with data management
Introduction

• The goal of the work presented
  – Create an integrated software environment
    • Address the challenges mentioned earlier

• Technologies
  – MediaFlux - Asset management platform
  – DaRIS - PSSD data model framework
  – Kepler - Scientific Workflow Management System
  – Nimrod - Nimrod/G and Nimrod/K
Introduction

• Case Study
  – Automated analysis of fMRI brain scans
  – Look for correlation between cerebral blood flow and the pain experienced by patients
MediaFlux
(Arcitecta Pty Ltd)

• Service oriented data operating system
  – Manages provenance of all data
  – Manages relationships of all data

• Server is developed in Java
  – Client access by HTTP, NFS, Java, .Net, JavaScript

• Manages datasets in “Assets”
  – May have 0 or more content components (binary or other data)
  – Associated metadata which is egested as plain text XML

http://www.arcitecta.com/Products/Mediaflux
MediaFlux
(Arcitecta Pty Ltd)

• Capabilities presented as set of services
  – Services used for finding, storing and retrieving assets
  – Services for data analysis and transformation

• MediaFlux servers can be grouped
  – Globally servers are identified with a unique ‘uuid’
  – Globally assets are identified with a unique (‘uuid’, ‘asset id’) pair

• Strong authorization model
  – Each repository has independent access control

http://www.arcitecta.com/Products/Mediaflux
DaRIS
(Centre for Neuroscience Research, University of Melbourne)

- **Data model – PSSD**
  - Applies to studies on variety of subject types
  - Objects have citable identifiers

- **Set of MediaFlux Plugin service**
  - Data-model driven
  - DaRIS includes clients for uploading specialised data formats

- **Web base portal**
  - Google Web Toolkit implementation
  - It has knowledge of the framework
  - Has no knowledge of domain specific metadata or Methods

DaRIS
(Centre for Neuroscience Research, University of Melbourne)

Nimrod Tools
(MeSSsAGE Lab - Monash University)

- Nimrod/G
  - Parametric experiment execution
  - Distributed scheduling
    - Manages the scheduling across all available resources
    - Can scheduler to meet user defined time or budget constrains
  - Interfaces with Grid Middleware
    - Condor, SGE, Globus Toolkit
    - Cloud systems – Amazon’s EC2 or Microsoft’s Azure

http://messagelab.monash.edu.au
Nimrod Tools
(MeSSsAGE Lab - Monash University)

• Nimrod/K
  – Provides Nimrod/G functionality
    • Built on Kepler’s runtime engine
    • Uses Nimrod/G to run computations on the Grid
  – Scheduler that can be changed dynamically
  – New custom Tagged Dataflow Arcitecture Director
    • TDA supports concurrent threads of execution in the workflow itself
    • No change required to existing actors to run under TDA Director

http://messagelab.monash.edu.au
Nimrod/K – GridJob Actor
Integration
(Arcitecture – PSSD Methods)

• Presently experimental methods have
  – Subject state change – a record of state change in the experimental system
  – Data Set – the acquisition of data and related metadata
  – Branching – a decision point where one or more paths may be followed

• Transition between these steps is manual
Integration
(Arcitecture – PSSD Methods)

• Transform Step
  – Context on how to execute data transformation – eg. computational workflow

• Defines
  – Inputs/outputs
  – Optional parameters
  – Transformation itself

• Output of one Transform Step may be used as an input to another Transform Step
Integration
(Architecture – PSSD Methods)

- **Transform Service Provider**
  - MediaFlux Service
  - Must be able to control and monitor executions of Transform Steps

- **Executed on one or more resources**
  - Researcher is provided with options to choose

- **Re-execute option available**
  - With new parameters
  - Existing data could be kept or discarded – decision made by a researcher
New Components
(Implementation)

• MediaFlux Transform Framework
  – Transform Provider – controls and monitors the transformation
  – Transform Definition – definition of transform process
    • Parameter definitions
  – Transform Instance – actual instance of the transform in MediaFlux
    • Current state of the execution
    • Parameter values
  – Transform Output – stored as assets in MediaFlux

http://www.arcitecta.com/Products/Mediaflux
New Components
(Arictecture – PSSD Methods)

http://www.arcitecta.com/Products/Mediaflux
New Components
(Implementation)

• PSSD Data Model
  – Transform Step in PSSD
    • References Transform Definition (how to execute external process or computational workflow)
    • Select inputs and outputs for the Transform Step
  – Virtual Subjects
  – PSSD Services
  – New metadata for processed data

New Components
(Implementation)

• Nimrod/K
  – Kepler Transform Service Provider
  – Remote control of the Workflow Execution
    • Start, Suspend, Resume and Terminate
  – MediaFlux Actors
    • Retrieve assets from a MediaFlux repository
    • Create and store new assets to a MediaFlux repository
  – Improvements to the Nimrod/K core
    • Nimrod/K Director – Execution control
    • GridJob Actor – Directory transfers, file filtering, placeholder substitution

http://messagelab.monash.edu.au
Case Study - Background
(Imaging Clinical Pain)

- Identify brain activation associated with the low back pain in people with musculoskeletal disease
  - fMRI for acquired for two groups of participants
  - Data acquired consisted of
    - Structural brain image
    - Functional brain images acquired every 6 seconds
    - Participants pain rating acquired every minute
  - Identical analysis of data for all participants
    - Healthy participants used pain ratings of matched patient with low back pain
• Measured regional cerebral blood flow (rCBF)
  – Estimated from fMRI
• Correlated with pain ratings acquired during fMRI scan
  – Together with rCBF these help identify brain structures that have some relation to pain intensity
• Healthy to in-pain participant matching was done on basis of age and sex
Case Study - Data
(Imaging Clinical Pain)

• Data was captured manually and stored in a DaRIS repository at University of Melbourne
• PSSD model was used to organise the data
  – Asset created for each subject (participant)
  – Pain rating was included with corresponding acquired data
Case Study - Execution
(Imaging Clinical Pain)

• Created a Kepler Workflow
• Process manually started in PSSD model
• Nimrod/K runs the selected workflow with user specified parameters
  – Dataset IDs, server IDs etc
• Produces a DataSet for each analyzed subject
Case Study - Execution
(Imaging Clinical Pain)
Case Study - Execution
(Imaging Clinical Pain)
Case Study – Results
(Imaging Clinical Pain)

• Results show that low back pain participants have
  – Decreased levels of blood flow in subgenual cingulate cortex
  – Increased levels of blood flow in left inferior frontal gyrus
Case Study

Computational Results
Conclusion & Future Work

• This paper discusses the enhancement done to DaRIS repository and Nimrod/K workflow environment
  – Transform Framework
  – Remote control of Nimrod/K workflow execution
  – New actors that interact with MediaFlux
  – Improvement of current Nimrod/K components
    • Nimrod/K Director
    • GridJob actor
Conclusion & Future Work

• Identity management
• DaRIS user interface
• New data transport mechanisms in Nimrod/K
• Improve remote control interface
  – Have a service that allows users to execute workflows on dedicated remote hosts
Questions?