Data Analysis Workbench

an on-line and off-line toolkit for synchrotron science

EMBL
BUILT ON eclipse
ESRF

WEPKS019
Sandor Brockhauser, EMBL
Matthew Gerring, ESRF
Andrew Götz, ESRF
Olof Svensson, ESRF

"Connecting analysis, visualization and collection in an elegant framework"

Workflows and Pipelines

What is a Workflow Engine?

Workflows (also called pipelines) are chains of linked actions or 'actors'. The linkages define an algorithm which incorporates these actors. Messages are passed between actors and represent the point of execution of the algorithm similar to script programming, however multiple messages may exist in a pipeline at one time. The actors are made up of 'Sources', which add jobs to the pipeline queue, 'Transformers' which complete tasks within the workflow and 'Sinks' which receive and stop propagation of the current message.

The user defines a set of sequential and/or parallel actions using the actors. This model is interpreted by a workflow engine. The workflow engine uses multi-threading on the CPU allowing an efficient model to be developed by a user. The workflow engine will load the CPU(s) of the host machine(s) efficiently because the engine uses CPU threads when the user creates a model with branches and/or with multiple jobs in the pipeline (using Sources). This approach avoids the need for the creator of the workflow to program threads directly. Some workflow engines have user interfaces allowing a graphical design to be made of the workflow algorithm. For DAWB the engine 'Ptolemy II' was chosen,

How did you use one?

We created a user interface for configuring the model of the engine and allowed the engine to run on a web service. This service allows the workflow to run from the data analysis workbench or from other programs, such as a Tango server. The current applications are in macromolecular crystallography, spectroscopy and image packing. For instance there two new models running a legacy Fortran code in real time as a detector changes for a spectroscopy experiment. More information in the case studies below.

DAWB Architecture for Design of Experiments
Currently being tested for mxv2, id14-4 and plans for testing on other beamlines.

Client, n Workbench Layer
(where n is small)

User Interface (based on rich client platform)			
Work Flow Engine (based on Ptolemy II)			
Hardware Monitoring	Hardware Work Flow	Image Monitoring	Data Analysis Work Flow

Using JMX

Tango, m,
(usually 1)

Hardware Layer, 1
x-devices
where x is large

Algorithm tested for high data rates and large folder contents

"Able to drive analysis and complete logic to design experiments which can be run automatically or with user interaction"

Data Acquisition
- Spec Motors and Macros
- Taco Devices
- Hardware Repository
- Detectors
- Other tango servers

Data Analysis
- Edna plugins
- Workbench actors
- External codes (where not edna) C / Fortran / etc.
- Python numpy
- Other work flow actors, KNIME, Kepler, Taverna etc

Images to disk

Current design for linking data analysis and acquisition in a user or beamline scientist configurable way.

How hard is it to add your own actors?

Much of the time is is not necessary to write your own code, running commands and EDNA are already supported but what happens when you do?

1. Python Script Actor
One of the most popular actors is a python script which receives data from the upstream workflow and allows the custom code to be run. The workbench also provides an IDE for editing python based on pydev.

Supports sub-routines and gives access to workflow variables
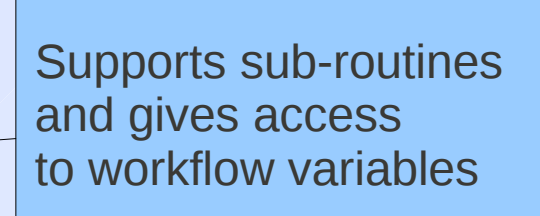
Context sensitive color-scheme, debugging and team tools

2. Java Actor
You can create your own Java actors and deploy them in the workbench without releasing their source code. A special clause has been added to the DAWB license agreement to allow this to happen. There is a guide on the web site for creating and adding your own Java actor.

Easy to implement powerful, thread safe designs
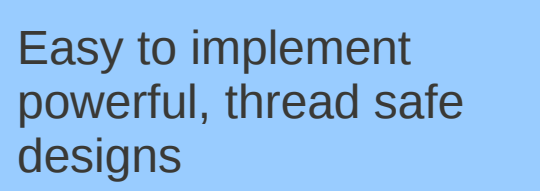
Custom message type supports various data types and integrates with Python
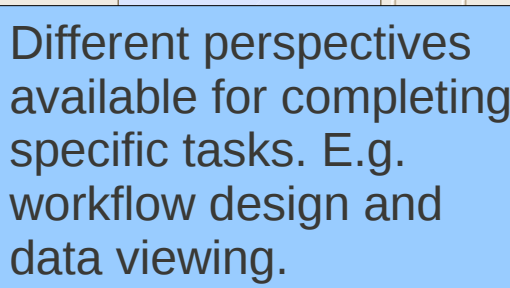
Viewing, Slicing and Plotting of Data

Support for a wide range of file formats

A wide range of plot options. Including log, zoom, pan, annotations and unlimited undo/redo

Results can be dragged to the workbench as links or by copying

New projects may be created and edited with different roles such as workflow editing or data viewing.

User interface tasks are monitored and cancelable and run in separate threads.

Different perspectives available for completing specific tasks. E.g. workflow design and data viewing.

The whole data set can be opened as a gallery of images similar to a photo album.

Individual images when selected are shown in a larger part.

Ability to zoom on a region and then inspect this region further using numerical tools
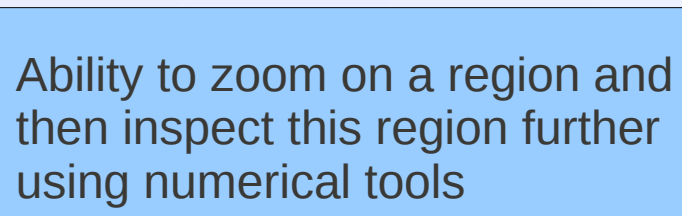
Slicing images out of hdf5 stacks

Choosing the plot axes.

Slices may be done as summed ranges and the summed slice displayed back in the main window

Ability to change the intensity range to slice by intensity as well as image index.

Ability to view the full tree of data including meta data

Case Studies

### 1. Enhanced characterization workflow for Macromolecular Crystallography

A workflow model was developed in the workbench for aligning the kappa hardware on mx beamlines at the ESRF. A more detailed paper is being published on this work. The screenshot below shows the steps from an early version of this workflow. The actors are able to move hardware and complete a data collection algorithm, in this case defined in a spec macro. There are also python actors for transforming data and EDNA python actors for completing sub-workflows, running more complex software and using a data model developed inside EDNA.

2. Running JESF with zero dead time on the TEXAS experiment

A workflow was developed to connect to a detector writing to shared memory at the ESRF. The connected pipeline gives users the option to run any analysis and visualize the results in real time while connected to the live detector. The initial testing was completed running a Fortran code developed to take a spectra and complete provide information like noise, EXAFS and the Fourier transform. The user can open multiple windows and monitor the plots in real time while the workflow is running.

3. Packing many images into a hdf5 file

A workflow was developed to pack image directories containing many (~10k) images of size 2048x2048. This allows large (~100Gb) directory contents to be combined into a single file. Slicing and analysis tools can then be used. It is possible to transport and store this file more efficiently than with individual images. Although the workflow looks simple, the multi-threaded approach of Ptolemy II ensures that images are read and written asynchronously making the most of hdf5 writing and the memory of the local machine.
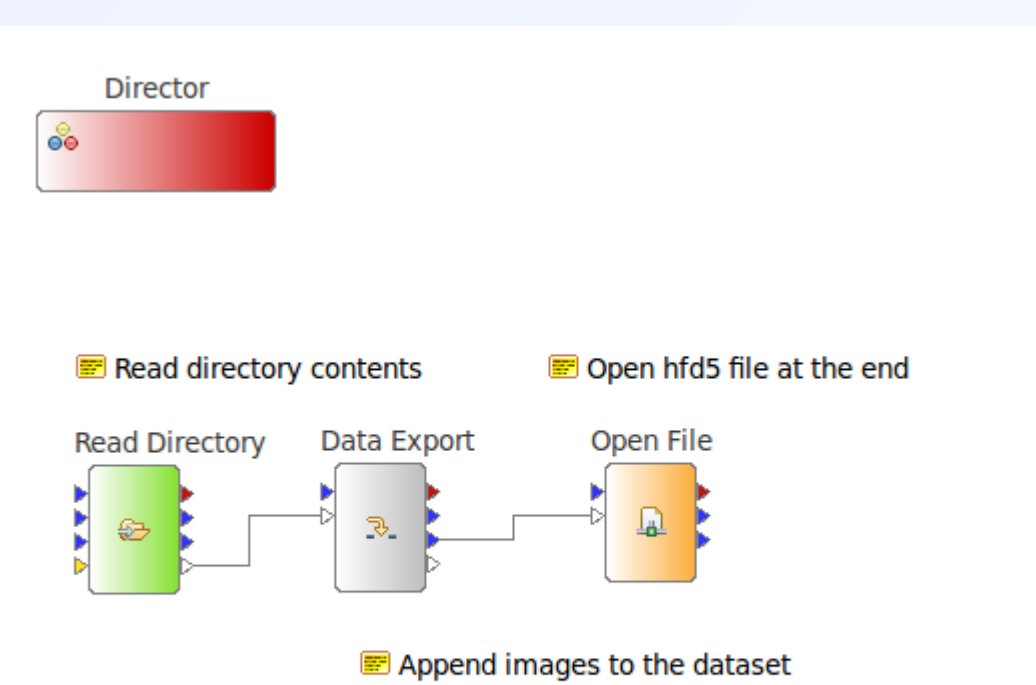
Download Data Analysis Workbench

You can download and try the workbench for yourself at www.dawb.org. The software is available on linux, windows and mac and is open source.

www.dawb.org

icalepcs 2011