ISIS Data, Metadata, Proposals and the CCLRC Data Portal

Damian Flannery
ISIS Facility, CCLRC Rutherford Appleton Laboratory, Chilton, Oxon, UK

Abstract

The storage, retrieval and management of data is fast becoming a major concern for all large scale facilities. In addition to the 20 years worth of archived raw data, the ISIS facility produces ~700GB of Neutron/Muon data each year, and with the introduction of Target Station II, this rate of data collection is set to rise still further. Looking to the future, the full value of these data resources will only be realised if they are easily searchable, accessible and reusable. To this end, a number of significant software developments are taking place at ISIS. These include a new electronic proposal system (ISIS On-line Proposal System) which automates the submission process and provides access to a rich source of metadata which can be fed directly into the experimental setup; the migration towards the use of a Storage Resource Broker (SRB) for ISIS data storage that provides heterogeneous access to resources based on their logical names rather than their physical names or locations; and more importantly, the development of a metadata catalogue (ICAT) comprising information describing each experiment plus links to any associated raw and log files. Such a catalogue provides many opportunities; from the capture of an instrument configuration at a particular point in time, to the tracking of an experiment from conceptualization (proposal) to its completion (publication). Furthermore, the ICAT will be exposed to the CCLRC Data Portal enabling all ISIS data to be searched and retrieved along with data from many other facilities contributing to the CCLRC Data Portal project (see www.e-science.cclrc.ac.uk).

Introduction

CCLRC is placing a major emphasis upon e-Science, and at ISIS there are a number of plans in place to improve the way that data collected at the facility (which is costly to obtain) is accessed, analysed and visualised using 'e-Science' techniques so that it can be fully exploited by the scientific community.

The capture of metadata and the provision of seamless access to the data produced at the facility is a fundamental component to achieving this goal. This paper describes the current developments that are progressing in this area at ISIS.

ISIS Online Proposal System

From a data management perspective, one of the most important, consistent and reliable sources of metadata come from facility beam-time proposals.

Examples of information that can be obtained from these proposals include:

- experimental team
- title and abstract of experiment
- instrument details – including name, days requested, preferred contact etc.
- sample details – including sample material, chemical formula, weight, volume etc.
- sample environment details – including equipment needed, temperature, pressure and magnetic field ranges of experiment
- sample safety details – any hazards associated with experiment

At ISIS, we have developed the ISIS Online Proposal System [1] to automate the capture of this metadata. This web-based system (developed using Oracle 9i, Java Server Pages, Struts and CSS) enables users of the ISIS facility to electronically create, submit and collaborate on applications for beam-time.

This evolution from the previous word form / paper-based system greatly reduces work for the departmental administration staff and can have a major impact on metadata annotation of experimental data produced from the facility.

The Information obtained from proposals will be used to pre-populate fields in the Instrument Control Program (ICP) making it easier for instrument scientists and users of the facility to set up their experiment. Furthermore, this information can also be used to define access permissions for data resulting from an experiment.

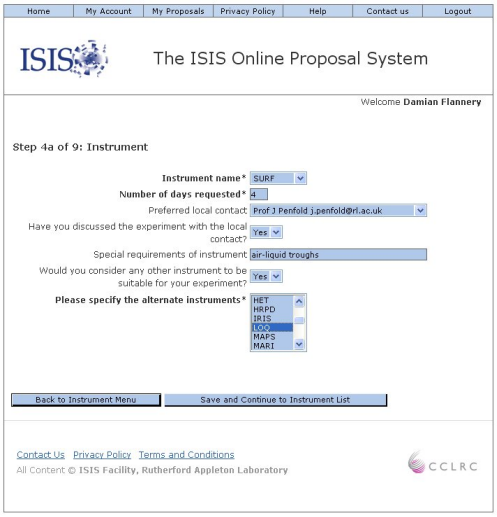

Figure .1 – Screenshot from the ISIS Online Proposal System

ISIS Metadata Catalogue (ICAT)
In order to maximise the value of data produced from the facility, it must be fully searchable. To address this issue, ISIS are developing a metadata catalogue which is based on the CCLRC metadata model [2]. The ICAT is essentially a database that contains information describing experiments.

Besides the obvious advantage of improving access to data, the catalogue has many more potential benefits. For example, the catalogue introduces the concept of provenance e.g. information on data creation, ownership, history, what processing has taken place on data (including software and versions), what analyses it has been used in, what result sets have been produced from it, and the level of confidence in the quality of information can all be modelled in a metadata catalogue.

Moreover the catalogue will act as a 'project repository' as described in [3], linking all aspects of research including proposals, raw data, treated data, analysed data, results and finally publications.

Additional benefits include the ability to relate data not just to an instrument, but to an instrument with a particular calibration and the ability to perform statistical analysis. This analysis can range from facility statistics e.g. the number of neutrons counted on a particular instrument during a cycle, to system statistics as a basis for future computing related capacity planning (particularly relevant for the Target Station 2 project at ISIS).

The ICAT system is based on a multi-tier architecture and is divided into the following layers:

- A data source layer that is used by the middleware layer to persist state permanently. The ISIS catalogue data source layer will consist of an Oracle 9i RDMS.
- A middleware layer that contains the core business logic exposed as services. The middleware layer in the ICAT system consists of a number of components based on the Enterprise Java Beans (EJB) technology and will be exposed as standardised xml web-services.
- A presentation layer that contains components dealing with user interfaces and user interaction. This layer will consist of a web-based application (Java Server Pages (JSP) and Struts) that will enable users to search/view/create/edit/delete and download data and metadata.

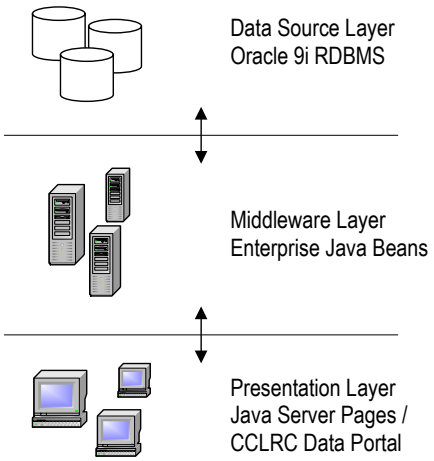

Figure .2 – Architectural diagram of the ISIS Catalogue

The principle advantage of this architecture is that by viewing the application as a collection of services, these services can be leveraged and made available to other applications. For example, Open Genie [4], a data analysis tool developed and used at ISIS, can be modified so that data can be easily located and manipulated from within this framework. Additionally, it will be possible to develop different flavours of clients, for example a GUI client application to view/edit/download data and metadata supplementing the web based version.

Storage Resource Broker (SRB)
As discussed previously, the ICAT contains pointers to actual data files. This causes an administration challenge in that data files are located by an absolute file path. If this path changes, e.g. a set of files are copied to another medium for long-term storage, then this change must be reflected back in the ICAT to preserve integrity.

Crucially, however, ISIS is migrating towards the use of the San Diego Supercomputing Centre's Storage Resource Broker (SRB). The SRB is a client-server middleware that provides uniform access to distributed resources [5]. In essence, the SRB provides a way to access data sets and resources based on their logical attributes rather than their physical names or locations.

SRB provides access to data via a uniform API and set of commands. Drivers to various storage devices can be configured e.g. Disk Farms, Tape Robots etc. This means that files can then be manipulated via these SRB commands with the underlying storage type remaining transparent to the user. The ICAT can thus link to files via an SRB URL, delegating the responsibility of maintaining the physical locations and access mechanisms of data to the SRB.

SRB is rich in features and introduces possibilities such as:

- Replication – The ability to replicate data. This is potentially advantageous when considering the download of large neutron data sets from diverse locations around the world.
- Federation – The ability to link to other SRB servers, potentially at other facilities.
- Authentication – Standard secure password authentication and support for certificate based authentication.
- Ticketing – The ability to enable users to access selected data sets on a temporary basis.

CCLRC Data Portal

The CCLRC operates many scientific facilities in which large quantities of data are stored across the organisation in the form of many files and databases. The CCLRC Data Portal [6] project is an e-Science initiative that aims to provide a gateway to this scientific data, and enable access to the many GRID based services that are being developed by the CCLRC e-Science centre [7].

The over-all concept of the Data Portal relies on each facility producing its own metadata catalogue (e.g. ICAT). Each facility then must provide a standard web-service based interface to the catalogue / database. This means that when someone performs a search using the Data Portal, each facility can be searched for results (transparently to the user).

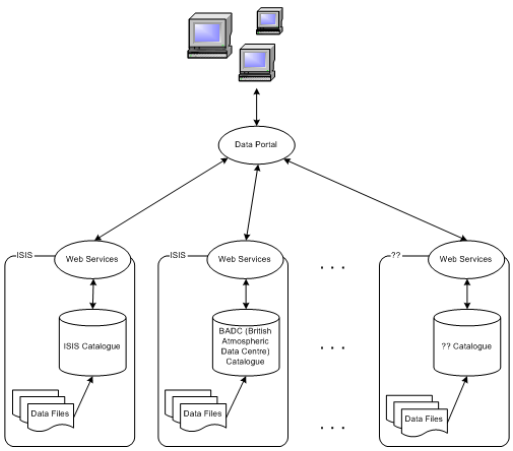

Figure.3 – Data Portal communication with facility metadata catalogues.

This project offers many altruistic benefits to the research community, for example, the repetition of experiments can be avoided by discovering data that already exists, collaborations can be built by identifying that someone is working in a similar area, and there is even the possibility reanalysing existing data when better analysis tools become available. Though, perhaps the biggest attraction is the integration of the HPC portal and the Visualisation portal [8] services. Once users have located their data in the Data Portal they can invoke these services on their data.

Significantly, the Data, HPC and Visualisation portal's all use SRB to access data, further reinforcing the case of using SRB as a storage mechanism.

Conclusion

This purpose of this work is to unite the many disparate sources of information currently distributed within ISIS and to provide access to many of the GRID based services currently being developed by the CCLRC e-Science centre. There are many challenges in such a pursuit, including issues of security, data ownership etc. However, these will be overcome; delivering a significant step forward in the way data management and science is performed at ISIS.

References

[1] The ISIS Online Proposal System http://proposal.isis.rl.ac.uk/proposal
[2] Matthews BM, Sufi SA. The CCLRC Scientific Metadata Model - Version 1 http://www-dienst.rl.ac.uk/library/2002/tr/dltr-2002001.pdf
[3] Watson PW. Databases and the GRID http://www.nesc.ac.uk/technical_papers/PaulWatsonDatabasesAndTheGrid.pdf
[4] Open GENIE web-site http://www.isis.rl.ac.uk/OpenGENIE/
[5] Rajasekar A. et al. Storage Resource Broker – Managing Distributed Data in a Grid http://www.npaci.edu/DICE/Pubs/CSI-paper-sent.doc
[6] Data Portal web-site http://www.e-science.cclrc.ac.uk/projects/dataportal/
[7] CCLRC e-Science centre web-site www.e-science.cclrc.ac.uk/
[8] Grid services portal web-site www.e-science.cclrc.ac.uk/projects/gridservicesportal