ICAT: Integrating data infrastructure for facilities based science

Damian Flannery, Brian Matthews, Tom Griffin, Juan Bicarregui
STFC Rutherford Appleton Laboratory,
Didcot, OX11 0QX, UK
brian.matthews@stfc.ac.uk

Michael Gleaves, Laurent Lerusse, Roger Downing
STFC Daresbury Laboratory,
Warrington, WA4 4AD, UK

Alun Ashton
Diamond Light Source Ltd,
Didcot, OX11 0DE, UK

Shoaib Sufi, Glen Drinkwater, Kerstin Kleese
Devigo Limited,
Daresbury, WA4 4FS, UK

Abstract— Scientific facilities, in particular large-scale photon and neutron sources, have demanding requirements to manage the increasing quantities of experimental data they generate in a systematic and secure way. In this paper, we describe the ICAT infrastructure for cataloguing facility-generated experimental data which has been in development within STFC and DLS for several years. We consider the factors which have influenced its design and describe its architecture and metadata model, a key tool in the management of data. We go on to give an outline of its current implementation and use, with plans for its future development.

Keywords - data management; data curation; facilities; metadata; synchrotron source; neutron source.

I. Introduction

Neutron and photon sources are a class of major scientific facilities serving an expanding user community of 25,000 to 30,000 scientists across Europe, and a much wider community across the world, within disciplines such as crystallography, materials science, proteomics, biology and even archaeology. Two new photon (light) sources became operational very recently in Europe (SOLEIL, the Diamond Light Source (DLS)) and several others are under construction or will undergo major upgrades (ALBA, PETRA-III, EU-XFEL, ESRF). The ISIS pulsed neutron source has just inaugurated a second target station and the ILL will also undergo a major upgrade of its experimental facilities. The modern electronic detectors and high-throughput automated experiments in these facilities mean that experiments are of increasing complexity, increasingly done by international research groups and many of them will be performed in more than one laboratory. The resulting raw data as well as the processed data needs to be accessible over the Internet and remain on-line until the results are published and in many cases much longer to allow re-processing and to allow for the preservation of knowledge. Together these facilities will produce as much and maybe soon more data than projected for the Large Hadron Collider at CERN. This upcoming “data deluge” as recognized in for example [1], makes efficient and sustainable data management essential.

The current situation at many of the facilities, and in particular at the photon sources leaves data management almost entirely to the individual users who have traditionally carried away data on portable media. These media are notoriously unsuitable to guarantee the longevity and availability of precious and costly experimental data. Not only is this becoming unfeasible considering the dramatic increase in size of some of the data sets, it is also counterproductive for the scientific workflow and in the end constitutes a dramatic loss for the whole scientific community.

Consequently, such facilities have a strong requirement for a systematic approach to the management of data across the lifecycle. Further, the unified operating environment with an established staff and administrative processes provide an ideal environment for developing such a system. However, unlike in projects which are encouraging data sharing between research groups such as NERC DataGrid [2] and Archer [3], there has been no common software environment offered by the facilities community.

The ICAT system for cataloguing facility-generated experimental data has been in development within STFC over several years for in use at both the ISIS Facility and the Diamond Light Source. It forms part of an infrastructure supporting data management across the scientific lifecycle.

This paper will focus on the key software elements that allow ICAT to perform effectively on large-scale catalogues of experimental data. In the rest of this paper we motivate the development of the ICAT infrastructure within STFC facilities, and in particular the ISIS facility and describe the architecture of the current ICAT system, with a particular emphasis on the metadata model used to represent information on scientific investigations and their data holdings. We then present some details of the implementation.

Cataloguing of 'raw' facility data is only one aspect of the ICAT system. This paper will also highlight the possibilities afforded by the cataloguing of 'reduced' ‘derived’, and 'published' data, and opportunities for federation of distinct facility catalogues based around ICAT.

II. Supporting Facilities Science

Large-scale facilities are advanced scientific environment which have demanding computing requirements. Modern instruments can generate data in extremely large volumes, and as many instruments as possible are placed around target areas or beam-lines in order to maximize the output from the expensive neutron or synchrotron x-ray resource. Consequently, the data volumes are large and increasing, especially from synchrotron sources, and the data throughput is very high. Consequently the data management creates a need for data transfer and storage. The diverse communities involved in building instruments and software and also the different academic communities and disciplines, has lead to a proliferation in data formats and software interfaces. Similarly there is a diversity the terminologies used for concepts. While there are number of concepts in common (e.g. temperature and pressure and their units) there is a tendency to use different names.

The data collected has a large number of parameters, both measured from the operating environment (e.g. temperature, pressure) and from the sample (typically angles from a scattering pattern) and this requires a multi-variate analysis, typically over several steps. To handle the data volume and to use bespoke software, distributed computation such as Grid systems are required to access high-performance computation.

Facility users are typically from university research groups, but also from a number of commercial organizations such as pharmaceutical companies, and in both cases the data can be sensitive. Consequently, there is a need to manage different data access requirements, sharing data with a research team in different institutions, and restricting access to non-authorised individuals.

Finally, as expensive investments (DLS cost some £400M programme to commission), governments wish to maximise the science output from facilities. Thus there is a need to maximise the use of data for the original data collectors, by capturing, organising and presenting it to them in a manner so that it can be analysed with the most up-to-date techniques, and not be a subject of unnecessarily repetition of the experiment through lost or poor quality data. Further, there is an increased recognition that output can be maximised by managing data for the long-term so that is can be reused by future scientists rather than re-doing the experiment.

Thus when considering how to provide infrastructure to support facilities-based science, it is helpful to consider the whole of the research lifecycle involved, from submitting applications for use of the facility, through sample preparation and instrument configuration and calibration, through data acquisition and storage, secondary data filtering, analysis and visualisation to reporting within the research community, informally and through formal publication. By taking an integrated approach, taking into account the provenance of the data (Creation, Ownership, History), the infrastructure can maximise the potential for science arising from the data.

III. The ICAT Infrastructure

As a consequence of the above situation analysis of facilities science, an integrated approach has been taken to provide data infrastructure within STFC.

The core component is an information catalogue – the ICAT - which collates metadata about the experiment from different stages of the experimental lifecycle by integrating with a number of different systems supporting that stage, from proposal to publication. Thus systems which could be integrated across the lifecycle would include:

- **Proposals.** Once awarded beam-time at a facility, an entry is created in ICAT that describes the proposed experiment. Thus we can collect descriptive information on who is doing what experiment for what purpose.
- **Experiment.** Data collected from the experiment will be indexed by ICAT (with additional experimental conditions) and made available to the experimental team
- **Analysed Data.** The end user will have the capability to upload any desired analysed data and associate it with their experiments.
- **Publication.** Using ICAT the scientist is also e able to associate publications with the experiment and reference data from publications.

The ICAT collects metadata across the experimental lifecycle as automatically as is possible by interacting with a number of other associated systems almost all of which already exist as part of the operating environment. The major interactions with associated systems are illustrated in Fig. 1. Thus core provenance data is collected from the proposal system, information about parameters from data acquisition, etc. Thus the system is efficiently propagating metadata through the system, maintaining accuracy and completeness, and negating the need for retyping. There are a number of features which ICAT needs to accommodate to support its user community.

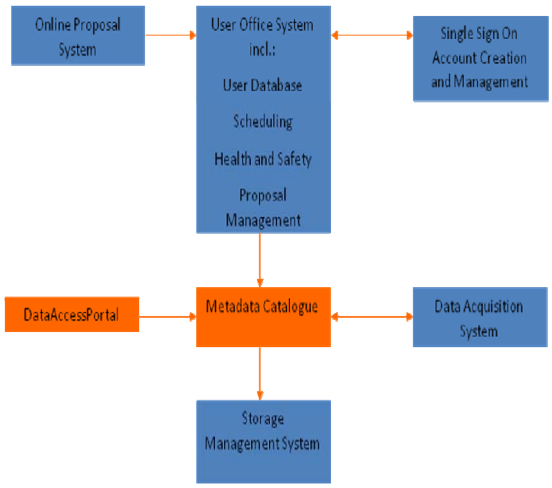

Figure 1. The ICAT and its associated systems.

- **Secure distributed access to data.** The user community for facilities is distributed across institutions who wish to access their data when they return to their home institutions. Therefore, ICAT provides a web-based front-end accessible from the scientist's desktop, which allows secure access to their data, while enabling the setting of appropriate access conditions so that it can be shared across experimental teams or collaborators.

- **Flexible data searching**. Data can be annotated with appropriate tags and keywords, most of which come from the proposal system. These annotations are used to search for data in a meaningful way, via a taxonomy of scientific terms, via resources (instruments, beam-lines) provided at facilities, or through sample and experimental parameters, such as temperature, pressure etc
- **A scalable architecture**. The ICAT infrastructure needs to be able to be scaled to allow rapid access to large volumes of data both in absolute size and also in the number of discrete data files which are indexed.
- **An extensible and flexible architecture.** The system should be adaptable to local requirements at different facilities and to be tailored to the needs of particular user communities. ICAT has adopted a modular open source approach which allows code to be shared and adapted.
- **Integration with analysis tools**. Data searching and access should be integrated with data analysis and visualisation tools to further process the data. ICAT allows data access to end user programs via its API.
- **Access to high performance resources**. In order to analyse the large quantities of data, the ICAT should provide seamless access to high-performance resources, such as computing clusters and large-scale data storage.
- **Linking to other scientific outputs**. Scientific data is part of a larger science lifecycle which includes other outputs, particularly formal publications, but also more informal forms of scientific communication, such as laboratory notebooks, blogs and newsgroups. ICAT should provide a means to link data with publications.
- **Data Policy.** Facilities are developing data policies to formalise access to data for the user community. While publically funded data is in principle publically available, in practice it is made available to investigators initially for their research. Thus ICAT enforces for ISIS a 3 year embargo on data (an additional year can be requested), and commercially funded data is never made public. The Instrument Scientists can access all data from the beam-line they are responsible for. Calibration data is public, but secondary analysis data that involves additional IPR is private for perpetuity unless explicitly shared by user

The overall ICAT architecture is given in Fig. 2. The core component, the ICAT itself, is a database storing the metadata associated with scientific resources. This provides a well defined API that provides a uniform interface to experimental data and a mechanism to link all aspects of research from proposal through to publication. This is published as a web-service interface and allows bindings in a number of languages including C++, Fortran, and Java so that end user applications can interact with the ICAT. A web-based front end (the "Data Portal") provides an alternative interface allowing browsing and searching of the catalogue and access to the experimental data.

The back-end of the ICAT interfaces to the data storage system, for example to a virtualised file store on a mass-storage system, such as the Storage Resource Broker[1]; this can be tailored to other storage systems. There are also interfaces to the user database and single sign-on systems which control user identification and authentication within the facility. The ICAT can also be linked to other systems which supply it with data, especially the proposal system, initiating investigations. Further interfaces to e-Science services such as high-performance computing or visualisation, to the publications system cross-linking with publication data and software libraries can also be added.

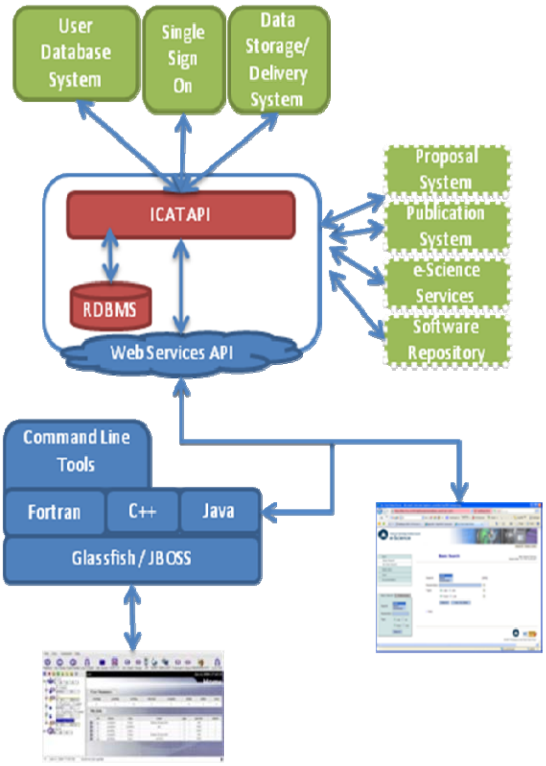

Figure 2. Architecture of ICAT.

An inherent part of design of the ICAT infrastructure is that they can be federated together. Users may typically use more than one facility and wish to access their data on each via one interface. Given an ICAT installation at each, then a common data portal can search and browse all the databases at one time via the common ICAT API.

IV. The ICAT Data Model

A key aspect of the ICAT infrastructure is the data model used within the metadata catalogue. This database schema conforms to the Core Scientific Metadata Model (CSMD) which has been designed to capture information about experiments and the data they produce within facilities science, and has been the result of an analysis of science practice over a number of years and a number of projects to allow the user to search for interesting data.

1 http://www.sdsc.edu/srb/index.php/Main_Page

The currently implemented version is based on the CCLRC Scientific Metadata Model v2 [4] with many extensions. For more details on the model see [5].

The model is organised around a notion of Studies, a study being a body of scientific work on a particular subject of investigation. During a study, a scientist would perform a number of investigations e.g. experiments, observations, measurements and simulations. Results from these investigations usually run through different stages: raw data, analysed or derived data and end results. Data should be grouped accordingly, and associated with the appropriate experimental parameters. Not all information captured in specific metadata schemas would be used to search for this data or distinguish one data set from another, give possibility to select special parameter. The CSMD is designed to be a common general format/standard for Scientific Studies and their associated data holdings.

Thus this model:

- Forms a specification for the types of metadata which should be captured during Scientific Studies
- Allows citation, collaboration, exploitation and integration of information on scientific studies.
- Allow easy integration of distributed heterogeneous metadata systems into a homogeneous (albeit virtual) platform

The CSMD has been developed to be a core system which is extensible and can be specialised to particular scientific domains, so it does not make assumptions about the specific terminology of the domain.

The core scientific metadata is a study-data set orientated model holding information about Investigations and their associated datasets and datafiles. The core entities of the CSMD are given in Fig. 3, and are summarised as follows.

- **Investigation**: forms the fundamental unit of the model, with a title, abstract, dates, and unique identifiers referencing the particular study. Also associated with the investigation are the facility and instrument used to collect data.
- **Investigator**: described the people involved in the study, together with their institution and role in the study (e.g. principle investigator, research student).
- **Topic and Keyword**: provide controlled and uncontrolled vocabulary to annotate and index the investigation.
- **Publication**: provides links to publications associated with (motivating or derived from) the investigation.
- **Sample**: information on the material sample under investigation within the study. The model has fields for a sample's name, chemical formula and any associated special information about it, such as specific safety information on a toxic material.
- **Dataset**: one or more datasets can be associated with an investigation, representing different runs or analyses on the sample. Initially a raw data set can be attached to the investigation, but subsequently, analysed datasets can also be associated.

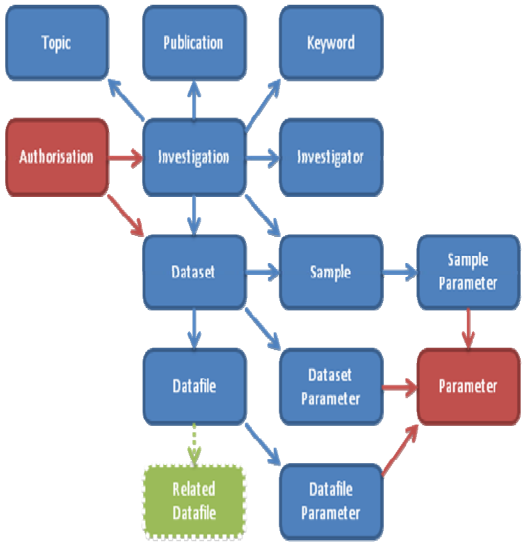

Figure 3. The main entities of the CSMD

- **Datafile**: the CSMD takes a hierarchical view of data holdings, as data sets may contain other dataset as well as units of storage, typically datafiles. Each datafile has more detailed information, including its name, version, location, data format, creation and modification time, and fixity information such as a Checksum.
- **Parameter**: parameters describe physical entities associated with the investigation, such as temperature, pressure, or scattering angle, describing either the parameters of the sample, the environment the data was collected in, or the parameters being measured. Thus parameters are associated with the sample, dataset or the datafile, and have names, units, values, and allowable data ranges.
- **Authorisation**: the CSMD can associate conditions on investigations and data sets, so that user specified access conditions can be specified. Thus the authorisation entity can record which user in which role can access data on specific investigations.

V. IMPLEMENTATION

The components of the ICAT software are given in Fig. 4. It has been design in a modular architecture using a Service Oriented approach with clear functional boundaries for each component, Core functionalities have been grouped together, customisable presentation layers are separated from the function layer to achieve easy maintenance, easy customisation, insulation from changes to underlying modules. The core component, the ICAT itself, is a database storing the metadata associated with scientific resources. An EJB layer provides a well defined API that specifies a uniform interface to experimental data and a mechanism to link all aspects of research from proposal through to publication.

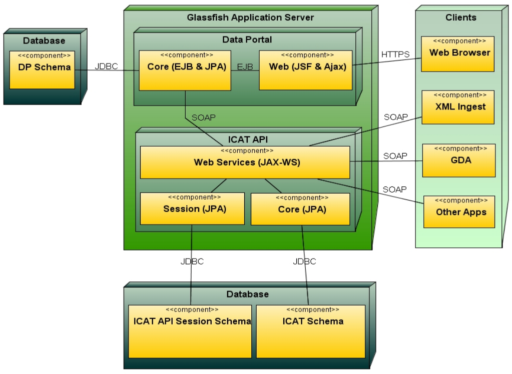

Figure 4. The component diagram of the ICAT suite.

The system is deployed using the Glassfish application server, with core functionality implemented as Java Objects using JPA. For deployment EJB3 Session Beans bind the core API, user database and data delivery aspects together. Users are required to authenticate in order to obtain session token, which is then used in all subsequent API calls to for authorisation. STFC currently use an Oracle system for query optimisation, although the system is designed to be application server and database independent.

All interactions with the ICAT catalogue are through the ICAT API which provides a functional library to interact with the CSMD model. Fig. 5 gives a description of a typical function in the library, the `addSample` function, which associates a sample with an investigation. The Web Service API is provided as a WSDL description. The interface is defined in Java but the web service can be accessed from any language capable of communicating with a SOAP web service. .NET (C#) is also used in part of the ISIS production system.. The metadata also can be represented in XML, with an associated XML schema. An XML ingest programme is also provided with the ICAT suite to allow metadata sets to be more easily uploaded into the data catalogue.

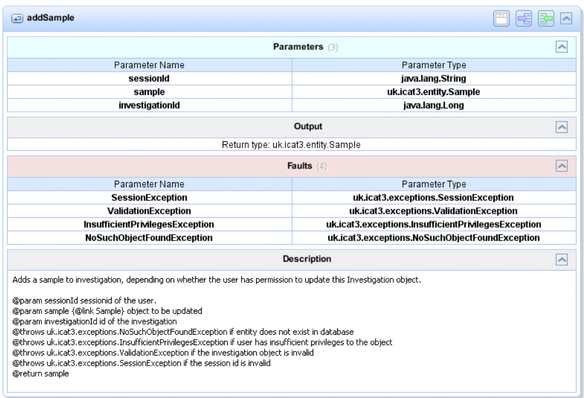

Figure 5. The addSample function

VI. A Data Portal

A web based user client has been developed to the ICAT infrastructure – the Data Portal. This provides the end user with a tool to access their data holdings. Versions have been provided for ISIS and Diamond which offer slightly differing functionality and a look and feel tailored to the local style guides and terminologies. For example, Diamond uses "beamline" where ISIS might use "instrument", and this is reflected in the instrument. Figs 6, 7 and 8 give typical pages from the ISIS data portal.

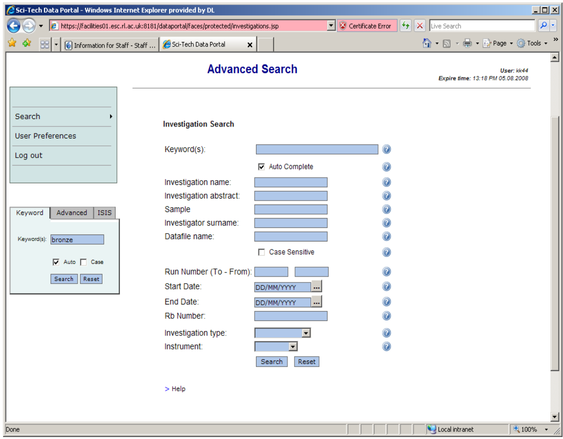

Figure 6. Searching for data sets via the data portal

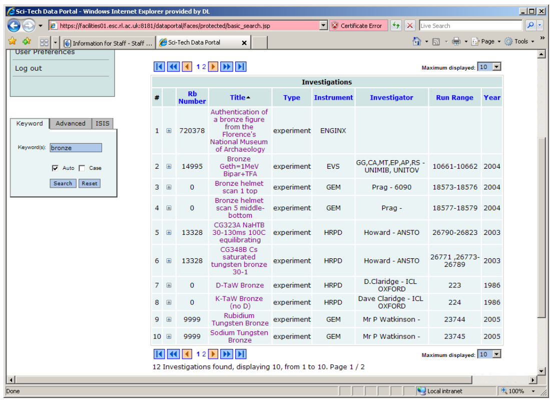

Figure 7. Listing investigations in the data portal

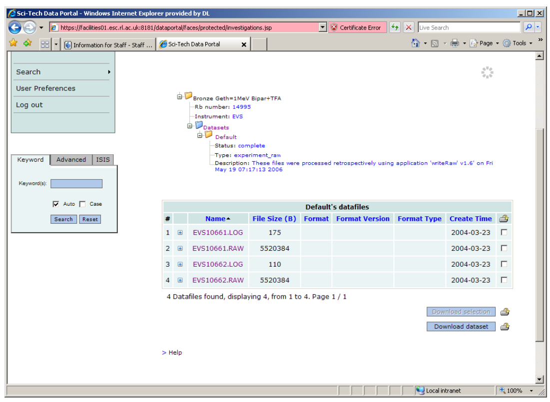

Figure 8. Accessing data sets via the data portal

Figure 6 presents a search interface where users can search for data sets which match particular parameters. Figure 7 gives a detailed view of a list of results of a search, and Figure 8 presents data access to the underlying data store via the data portal.

VII. CONCLUSIONS

The ICAT infrastructure has been under development for some years; the first experimental prototype was developed as early as 2001 [6], and the underlying metadata model and architecture, has been used in a number of projects over the last 8 years, including eMaterials [7]. The CSMD metadata model has been used in a number of different projects, including *myGrid* [8], *eBank* [9] and *Archer* [3] and has become known as a reference model for scientific metadata [5]. However, scientific facility users are a demanding community who require a high level of reliability and usability, and expensive large-scale facilities demand high-quality software to maintain their services.

The current version, ICAT V3.3, represents a new implementation of the ICAT suite based on the experienced of supporting large scale data management robustly and efficiently. The introduction of the ICAT API to directly manipulate the model, rather than using XML ingest together with the use of enterprise Java Beans and the web service interface has lead to a quality product which is reliable and efficient to meet the needs of a production service rather than a research experiment. For example, using Oracle and EJB has meant that tables of 25 million records can be queried in seconds. These key software elements allow ICAT to perform effectively on large-scale catalogues of experimental data e.g. the current ISIS catalogue of neutron data has approximately 4Tb of raw data, collected since 1984 in the form of some 3 million files and rising, can be searched in a matter of seconds from a web-based portal. Potentially, the three MX beam lines at DLS produce of the order 1TB per day. The ICAT experience thus represents a strong exemplar of how research software can be brought to service.

A. Future developments

There are further features which are not supported by the current ICAT infrastructure which would also be desirable. The current data portal user client is considered to be rather inflexible in its search and data access methods, and does not allow access to all desired features, such as setting access rights, and a new client is being piloted. This will allow greater flexibility in the management of a user's data assets and allow cross-facility searches. Further plans would be to extend this further so that analysis and visualisation tools can be directly access via the ICAT client, so that it supports the functionality supported by *Virtual Research Environments,* such as *myExperiment* [10], and also providing access to Grid and high-performance computing resources. Facilities scientists typically have strong attachment to particular tools and techniques to support analysis, such as MatLab and IDL, so the interface supplied must be flexible and modular so that such tools can be plugged into the existing infrastructure.

The current ICAT suite is independent of the particular facility and discipline used, so it needs to be tailored to each facilities terminology and indexing is currently using free-text terms. The use of Ontologies would make the management of terminologies for facilities easier and more explicit. Early experiments in building an ISIS Ontology have been undertaken [11]; this work can be built upon to support better and more accurate annotation and querying and allow easier exchange of information with user communities. However in a production environment, care needs to be taken to maintain a current and relevant controlled vocabulary.

ICAT supports an authentication and authorisation system supporting a limited role-based access control system. As already mentioned data policies are being developed to support facilities. ICAT could be extended to support richer access control to enforce policies through a rule based system. STFC has begun to investigate these features in the European Consequence project[2].

Currently in use at both the ISIS Facility and the Diamond Light Source at the STFC Rutherford Appleton Laboratory, the ICAT V3.3 release is now fully tested, documented and streamlined for adoption at other facilities. Although there is a substantial community of Neutron and X-Ray Synchrotron facilities around the world, support for data management had been developed on a more institutional based ad-hoc basis, and a common approach has not been developed. The ICAT software suite offers the possibility of such an approach, and as a consequence has been released as an open source development on Google Code[3], and is being investigated to support data management in a number of international facilities, including the Australian National Synchrotron[4], and the Institut Laue-Langevin (ILL)[5], an international reactor based neutron source in Grenoble France. The ICAT is also used as a component in at the Oakridge National Laboratory, USA, as part of its Neutron Scattering Portal[6] [12]. The emerging PaNData[7] community is beginning to explore the potential of coordinating effort for data management support across facilities internationally.

The adoption of a common tools and models to support the data management within large-scale facilities represents an opportunity for the traditionally self-supporting photon and neutron communities to establish common data and metadata standards and a common data infrastructure. Data can then be shared across facilities under common data policies, allowing users to more easily merge results which have different characteristics, allowing a more complete analysis of the target sample to be undertaken. The data management infrastructure can further be linked to computational resources and shared analysis and visualisation software. Thus ICAT could form a component in a step change in scientific practise in the field.

ACKNOWLEDGMENT

The ICAT and Data Portal has been a combined effort over a number of years. We would like to thank our colleagues within STFC and Diamond Light Source for their support and contributions to this effort.

2 http://www.consequence-project.eu/
3 http://code.google.com/p/icatproject/
4 http://www.synchrotron.org.au/
5 http://www.ill.eu/
6 http://neutrons.ornl.gov/portal/
7 http://pandata.neutron-eu.net/main/

REFERENCES

[1] A. J. G. Hey, and A. E. Trefethen. "The Data Deluge: An e-Science Perspective". In: Grid Computing - Making the Global Infrastructure a Reality, pp. 809-824, Wiley and Sons. ISBN 0470853190, 2003.

[2] B.N Lawrence, R Lowry, P Miller, H Snaith and A Woolf. "Information in environmental data grids," Phil. Trans. Roy. Soc. A vol. 367, pp. 1003-1014, 2009.

[3] S. Androulakis, A. M Buckle, I. Atkinson, D. Groenewegen, N. Nicholas, A. Treloar, and A. Beitz. "ARCHER – e-Research Tools for Research Data Management", International Journal of Digital Curation, Vol 4, No 1. 2009.

[4] S Sufi, B Matthews. "CCLRC Scientific Metadata Model: Version 2". DL Technical Reports, DL-TR-2004-001, http://epubs.stfc.ac.uk/work-details?w=30324, 2004.

[5] B. Matthews, S. Sufi, D. Flannery, L. Lerusse, T. Griffin, M. Gleaves, and K. Kleese. "Using a Core Scientific Metadata Model in Large-Scale Facilities", Proc. 5th International Digital Curation Conference, Dec. 2009.

[6] J V Ashby, J C Bicarregui, D R S Boyd, K. Kleese van Dam, S C Lambert, B M Matthews, and K D O'Neil. "A multidisciplinary scientific data portal", Proc. 9th Int. Conf. on High Performance Computing and Networking Europe (HPCN Europe 2001), Amsterdam, LNCS. Vol. 2110, eds. Bob Hertzberger, Alfons Hoekstra, Roy Williams (Springer-Verlag), p13-22, 2001.

[7] L. Blanshard, K. Kleese van Dam, and M. Dove "Environment from the Molecular Level e-Science project and its use of CLRC's Web Services based Data Portal", Proc. of the 1st. Int. Conf. on Web Services, 2003.

[8] N. Sharman, N. Alpdemir, J. Ferris, M. Greenwood, P. Li and C. Wroe. "The myGrid Information Model", Proc. of the UK e-Science All Hands Meeting, Nottingham, England, 2004.

[9] S. J. Coles, J. G. Frey, M. B. Hursthouse, M. E. Light, A. J. Milsted, L. A. Carr, D. de Roure, C. J. Gutteridge, H. R. Mills, K. E. Meacham, M. Surridge, E. Lyon, R. Heery, M. Duke, and M. Day. "An e-Science environment for service crystallography - from submission to dissemination", Journal of Chemical Information and Modeling, Special Issue on eScience. 2006.

[10] D. De Roure, C.Goble, and R. Stevens. "The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows", Future Generation Computer Systems 25, pp. 561-567. 2009.

[11] L B Casely-Hayford, and S Sufi. "The ISIS Facilities Ontology and OntoMaintainer", Proc. of the UK e-Science All Hands Meeting, Nottingham, England, 2006.

[12] M. L. Green and S. D. Miller. "Multitier Portal Architecture for Thin- and Thick-client Neutron Scattering Experiment Support", Proc of Int. Workshop on Grid Computing Environments 2007, Reno, NV, USA. http://casci.rit.edu/proceedings/gce2007, 2007.