Using a Core Scientific Metadata Model in Large-Scale Facilities

Brian Matthews, Shoaib Sufi, Damian Flannery, Laurent Lerusse,
Tom Griffin, Michael Gleaves, and Kerstin Kleese

STFC, e-Science Centre and ISIS Facility

Abstract

In this paper, we present the Core Scientific Metadata Model (CSMD), a model for the representation of scientific study metadata developed within the Science & Technology Facilities Council (STFC) to represent the data generated from scientific facilities. The model has been developed to allow management of and access to the data resources of the facilities in a uniform way, although we believe that the model has wider application, especially in areas of "structural science" such as chemistry, materials science and earth sciences. We give some motivations behind the development of the model, and an overview of its major structural elements, centred on the notion of a scientific study formed by a collection of specific investigations. We give some details of the model, with the description of each investigation associated with a particular experiment on a sample generating data, and the associated data holdings are then mapped to the investigation with the appropriate parameters. We then go on to discuss the instantiation of the metadata model within a production quality data management infrastructure, the Information CATalogue (ICAT), which has been developed within STFC for use in large-scale photon and neutron sources. Finally, we give an overview of the relationship between CSMD, and other initiatives, and give some directions for future developments.[1]

Introduction

Metadata are a key factor in the archiving and distribution of scientific data. Through the use of good metadata models, defined at the appropriate level, scientists can publish and share data, and allow the results of experiments and studies to be browsed, searched and cited. Appropriate metadata can thus encourage the reuse of data within and across scientific disciplines.

The Core Scientific Meta-Data Model (CSMD) is a study-data orientated model which has been developed at the Science & Technology Facilities Council (STFC) over the last 8 years; for earlier work see Sufi and Matthews ([2004](), [2005]()). The model is intended to capture high level information about scientific studies and the data that they produce. The Core Scientific Metadata Model (CSMD) is being used as the core metadata model within the data management infrastructure which is being developed for the large-scale scientific facilities supported by STFC including the ISIS Neutron Source[2] and the Diamond Light Source[3] In particular, it is a key feature of the Information CATalogue (ICAT), a software suite designed to manage the cataloguing and access to facilities data (Flannery et al., [2009]()). The CSMD was thus developed to support data collected within a facility's scientific workflow. However the model is also designed to be generic across scientific disciplines and has application beyond facilities science, particularly in the "structural sciences" (such as chemistry, material science, earth science, and biochemistry) which are concerned with the molecular structure of substances, and within which systematic experimental analyses are undertaken on material samples.

In this paper, we motivate and describe the scientific metadata model which has been developed within STFC, both in its overall structure, and some of the details. We also describe how the model is being used in the ICAT, and how it relates to other initiatives and how it may develop. This paper updates the report on version 2.0 of the model (Sufi & Matthews, [2005]()).

A Metadata model for Facilities Science

CSMD 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, for example, experiments, observations, measurements and simulations. Results from these investigations usually proceed through different stages: raw data are generated, this is then analysed to produce derived data which then may be refined to an end result suitable for publication.

The model thus defines a hierarchical model of the structure of scientific research around studies and investigations, with their associated information, and also a generic model of the organisation of datasets into collections and files. Specific datasets can be associated with the appropriate experimental parameters, and details of the files holding the actual data, including their location for linking. This provides a detailed description of the study, although not all information captured in specific metadata schemas would be used to search for this data or distinguish one dataset from another.

---

[2] Science and Technology Facilities Council ISIS: http://www.isis.stfc.ac.uk/index.html
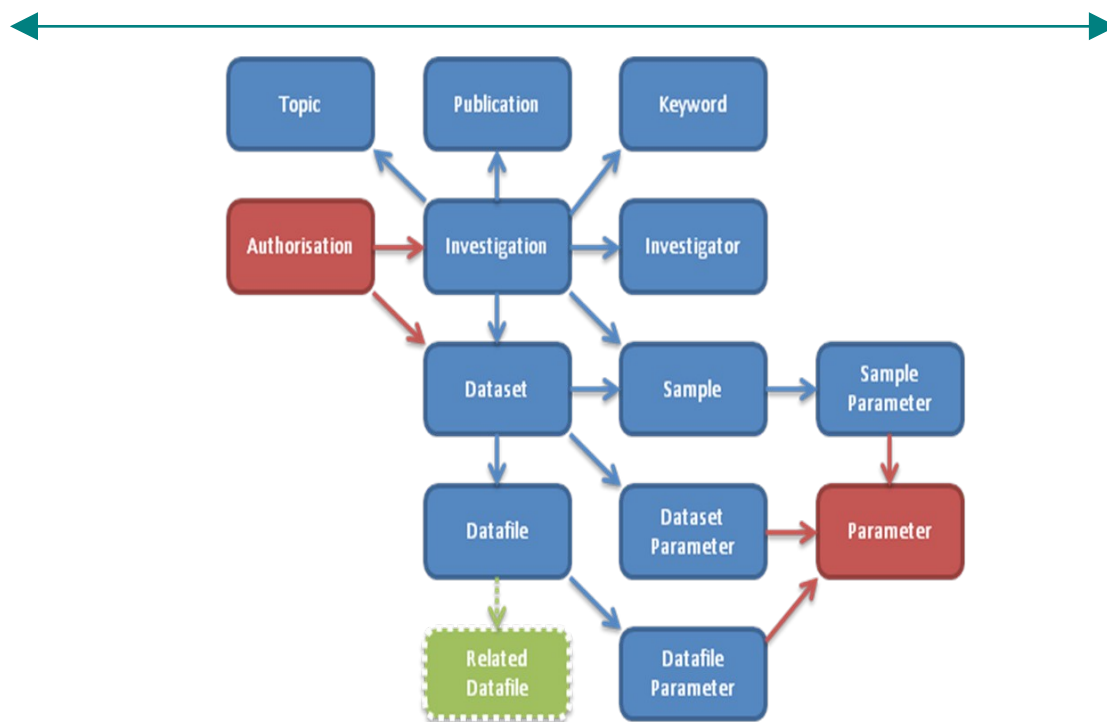[3] Diamon Light Source: http://www.diamond.ac.uk/

Figure 1. Main entities of the CSMD.

The core entities of the CSMD for a study are given in Figure 1, and are summarised as follows.

- **Investigation**. The fundamental unit of the study, including 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**. The people involved in the study, together with their institution and role in the investigation (e.g., principle investigator, research student).
- **Topic and Keyword**. Controlled and uncontrolled vocabulary to annotate and index the investigation.
- **Publication**. References 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, 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 dataset can be attached to the investigation, but subsequently, analysed datasets can also be added.
- **Datafile**. The CSMD takes a hierarchical view of data holdings, as datasets may contain other datasets 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 measureable quantities 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. Parameters can be associated at different levels, such as the sample, dataset or the datafile, and have names, units, values, and allowable data ranges.
- **Authorisation**. The CSMD can associate conditions on investigations and datasets, 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.

The Metadata Structure

The metadata within the general structure are laid in a series of classes and subclasses. We do not describe the whole model in detail for reasons of space, but rather select some areas of particular interest.

Modelling Scientific Activity

The data model describes scientific activities at different levels: the main unit is the Study, which optionally can lie in a context of a science research programme, governed by policies. Each study has an Investigator that describes who is undertaking the activity, and the Study Information that captures the details of this particular study. Studies include particular scientific investigations. The general structure of the metadata is given as a UML class diagram in Figure 2.
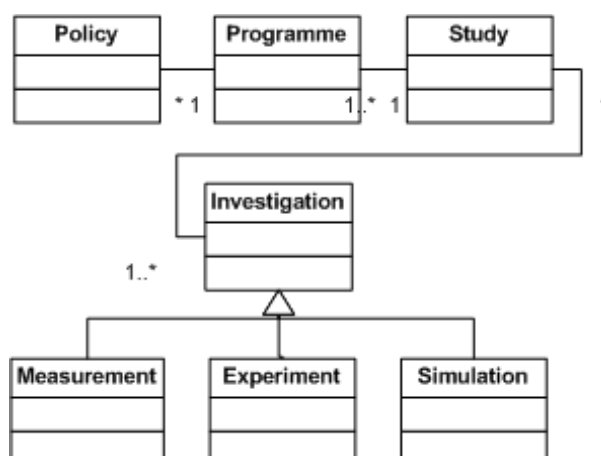

Figure 2. Model of the hierarchy of scientific data holdings.

- **Policy**. The company or government policies which initiate Programmes of work.
- **Programmes**. Related studies that have a common theme which are usually funded and resourced directly or with an intermediary organisation under the rubric of the programme.
- **Studies**. Studies investigate some aspect of science and have a Principal Investigator and/or institution, co-investigators and some specific purpose. e.g., an application for time on a facility such as ISIS.

- **Investigations** are studies or parts of studies that have links directly to data holdings, as described above. More specific types of investigations may include the following.
    - **Experiments**. Investigations into the physical behaviour of the environment usually to test a hypothesis, typically involving an instrument operating under some instrumental settings and environmental conditions, and generating datasets in files. E.g., the subjection of a material to bombardment by X-Rays of known frequency recording the resulting diffraction pattern.
    - **Measurements**. Investigations that record the state of some aspect of the environment over a sequence of points in time and space, using some passive detector, e.g., the measurement of temperature at a point on the earth surface taken hourly using a thermometer of known accuracy.
    - **Simulations**. Investigations that test a model of part of the world, and a computer simulation of the state space of that model. This will typically involve some simulation package with some initial parameters, and generate a dataset representing the result of the simulation.

Each investigation has a particular purpose and uses a particular set up of instruments or computer systems.

Classes within the model have several fields. For example, investigator has a name, address, status, institution and role within the study. For reasons of space we cannot provide a complete description of all the available classes within the metadata model. For illustration, we consider the Study class. Within a Study, there are several fields, as in Table 1.

ID	The key of the Study
NAME	Unique name given to the study.
PURPOSE	Description of purpose of study, an abstract of why these investigations are brought together.
STATUS	Ongoing or complete, as there could be additional investigations planned in the future which could be applicable to this study.
RELATED MATERIAL	Information related to the study. This could be related studies in other facilities, or on similar samples.
STUDY_CREATION_DATE	When the study was created.
STUDY_MANAGER	The user who has created the study – may not be the investigator, but rather a member of the facilities staff.

Table 1. Study Description Class Fields.

Further links in the study relate to the specific investigation. An investigation has fields for the investigators involved, together with their role and their contact details, and also references to the facility and instrument used to capture the data.

Modelling scientific data holdings

Investigations are characterised by the generation of a particular set of data on the analysis of a sample, initially raw data, but then further data representing analysed data. Other data may also be associated with the investigation, such as calibration data. Each dataset may have different parameters set. The model of data holdings used in the model needs to accommodate this complexity.

In CSMD each investigation is associated with metadata describing the data holding associated with that investigation. The metadata format given here is designed for use on general scientific data holdings, describing data logically which may be physically moved around. Thus, data holdings have three layers: the experiment, the logical data, and the physical files. The overall structure of the model for scientific data holdings is given in Figure 3, which gives more detail on Figure 1 above. Data holdings are considered as hierarchies, with datasets, which can contain sub-datasets which can be broken down into individual logical data files, generalised in the model as Atomic Data Objects (ADOs), as they may not be held in file store, but in, for example, databases. At each level of granularity, metadata can be provided giving representation information (as in Open Archival Information System (OAIS)) at the appropriate level of the data holding. At each stage of the data collection process, data are stored in a set of physical files with a physical location.
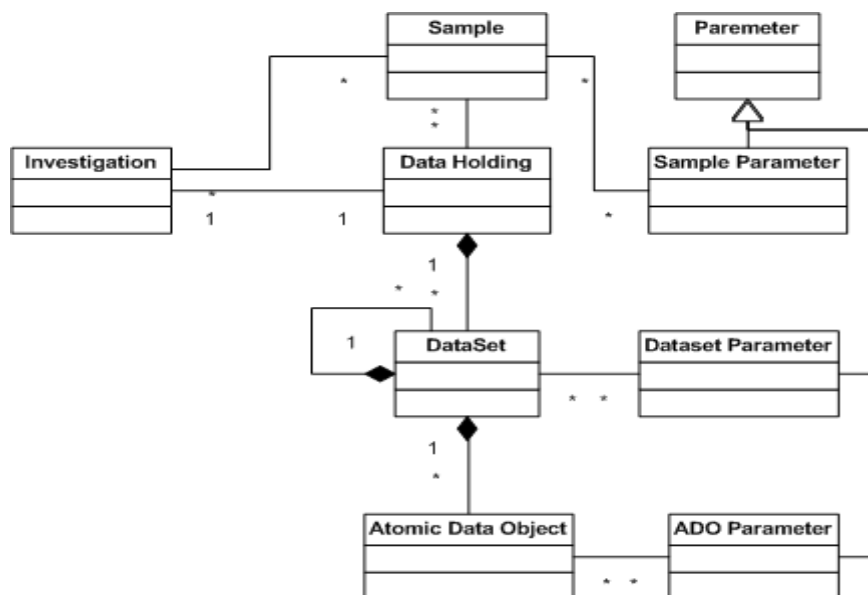

Figure 3. Model of the hierarchy of scientific data holdings.

It is possible that there may be different versions of the datasets in the holding. In a general data portal, all stages of the process should be stored and made available as reviewers of the data holdings may wish to determine the nature of the analysis performed, and other scientists may wish to use the raw data to perform different analyses. Thus type markers ('raw', 'intermediate', 'final') need to be kept with datasets and ADOs and relationships between them recorded.

The model distinguishes between the logical data holding, describing the data objects and their structural hierarchy, and the data location. The data location provides a mapping between the identifiers used in the data definition component of the metadata model, and the actual URLs of the files. This can provide facilities for describing mirror location for the whole structure, and also for individual files.

Parameters

Parameters can be associated with data holdings, datasets, or ADOs. The same metadata item is used to represent either experimental conditions or measured items stored as data points in the data collection, but are distinguished via a parameter type qualifier ('fixed' or 'measured'). Each parameter has a set of fields describing its name (e.g., temperature, pressure), its value (if fixed in as an input parameter), the units of measurement used to qualify the data points, the range of values over-which a parameter can take and the error margin expected on the value.
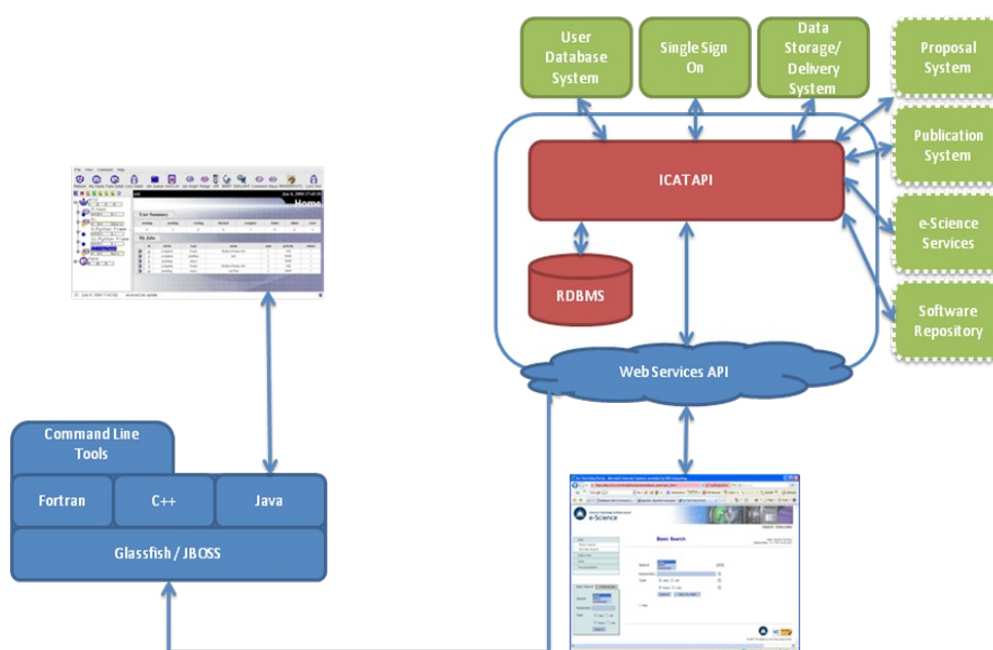

Figure 4. Architecture of ICAT.

CSMD in ICAT

An integrated approach has been taken to provide data infrastructure within STFC using the CSMD. 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 the systems supporting that stage, from proposal to publication, integrated across the lifecycle. The ICAT collects metadata across the lifecycle as automatically as is possible by interacting with associated systems almost all of which already exist as part of the operating environment. Thus core provenance data is collected from the proposal system, information about parameters from data acquisition, etc. Thus metadata is efficiently propagated through the system, maintaining accuracy and completeness, and mitigating the need for retyping.

The overall ICAT architecture is given in Figure 4. The core component, the ICAT itself, is a database storing the metadata associated with scientific resources. This provides a well defined Application Programming Interface (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 so that end user applications can interact with the ICAT. A web-based client (the "Data Portal") provides an alternative interface allowing browsing and searching of the catalogue and access to the experimental data.

Figure 5. Listing investigations in the ISIS data portal.

The back-end of the ICAT interfaces to the data storage system, for example, via to via a virtualised file-store to a mass-storage system such as STFC's Atlas Petabyte Data Store. There are also interfaces to the user database and single sign-on systems which control user identification and authentication within the facility. The ICAT is also 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, visualisation, software libraries and the publications system to cross-link with publication data can also be added.

An inherent part of design of the ICAT infrastructure is that it can be federated. 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.

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 interface. Figures 5 and 6 present two representative steps in the use of the ISIS Data Portal: Figure 5 displays a list of investigations resulting from a key word search, and Figure 6 presents accessing the underlying data holdings.
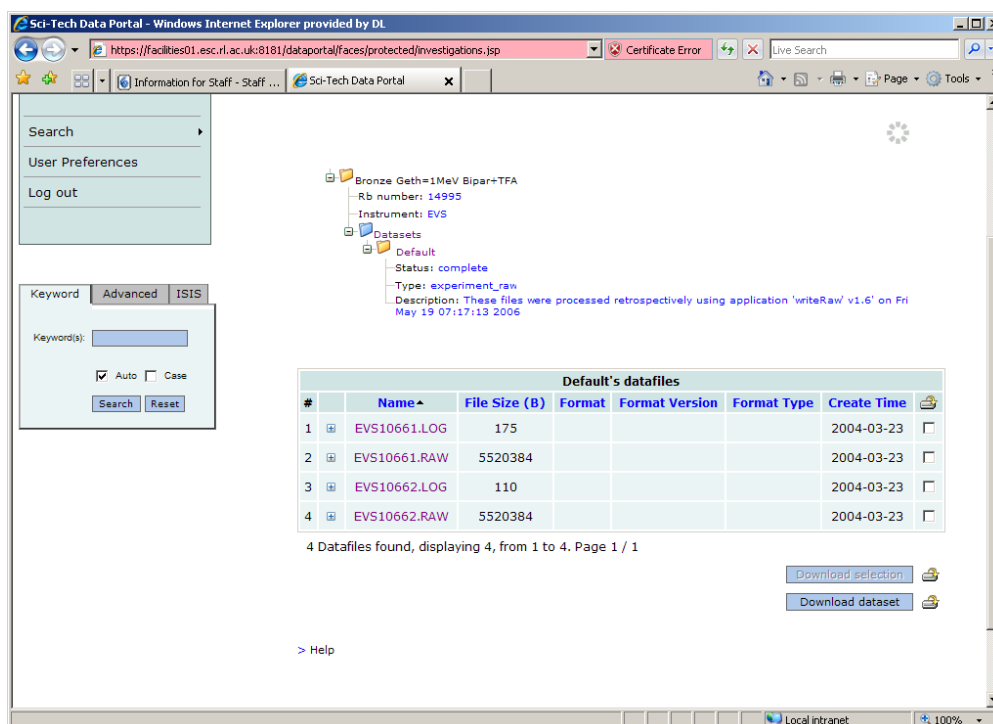

Figure 6. Accessing data via the ISIS data portal.

Conclusions

There are many metadata formats supporting specific data sources and domains. Formats which influenced the development of the CSMD included Climate and Environmental Retrieval and Archive (CERA) which was developed for earth observation data (Hoeck et al., 1995) and the Data Documentation Initiative[4] developed for the social sciences. However, there were few attempts to provide a metadata model to cover the general structure of scientific data holdings. Such a metadata for science has the requirement of being both more specialised than general metadata models, whilst being more general than metadata formats for specific domains in science, and covering a large range of metadata types. The CSMD is designed 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 CSMD was originally developed as part of the Council for the Central Laboratory of the Research Councils (CCLRC) Data Portal piloted within the e-Science programme at CCLRC (now part of STFC), using sample data from the ISIS and Synchrotron Radiation Source (SRS) facilities (Ashby et al., 2001a, 2001b; Sufi, Matthews & Kleese van Dam, 2003). The CSMD was then used as the base metadata model on a variety of UK e-Science projects, including the Natural Enviroment Research Council (NERC) e-Minerals (Blanshard, Kleese van Dam & Dove, 2003),

---

[4] The Data Documentation Initiative: http://www.icpsr.umich.edu/DDI/

Engineering and Physical Sciences Council (EPSRC) e-Materials (Blanshard et al., 2004), and the EPSRC Integrative Biology (Gavaghan, Simpson, Lloyd, Mac Randal & Boyd, 2005) projects. Further, it has been used as a template on a variety of other projects in the e-Science field; for example, the EPSRC MyGrid project adopted version 1 and enhanced the provenance information (Sharman et al., 2004) and the JISC eBank project has developed the format for crystallography data (Coles et al., 2006). The Australian Archer project has also adapted the model for another data management infrastructure for use in crystallography (Androulakis et al., 2009). The model has proven adaptable to a wide variety of situations, although not all; the NERC Datagrid project initially reviewed the CSMD, but developed its own model MOLES (Metadata Objects for Linking Environmental Sciences) more suited for data collected via environmental monitoring (Lawrence, Lowry, Miller, Snaith & Woolf, 2009). The CSMD thus can be seen as more suited to experimental science, typically an analysis of a sample in a laboratory or facility.

Over the last three years, the major current activity has been to develop a production ICAT for use in facilities, especially in ISIS and the Diamond Light Source, and with ICAT v.3.3 a robust software environment now exists. This project is using a relational schema based on version 2 of the CSMD but with modifications. Experience has shown that the model should in some respects be simpler than the full model as some metadata is hard to collect and of limited value. The ICAT infrastructure and its metadata model has attracted interest across the wider photon and neutron source community both in Europe, where it is being evaluated by the Institut Laue Langevin (ILL) neutron source, the European Synchrotron Radiation Facility (ESRF) synchrotron source, and the Paul Scherrer Institut in Switzerland amongst others, and in the wider world, such as the Australian National Synchrotron. A common metadata format for facilities' scientific data allows the possibility of providing integrated access to facilities data for its common international community.

Work on the metadata model is continuing. When the model was first conceived there were few metadata standards available. In the 8 years since the initial development of the model many metadata standards have been developed for example for bibliographic records such as the Dublin Core Metadata Initiative and its application profiles[5] for metadata registries, such as ISO11179 (2004), and for archival practice and curation such as PREMIS[6] as well as in more specialised domains. In order to integrate the metadata and the data into this wider infrastructure, we would integrate the use of the metadata with existing standards. For example, in order to have a common search mechanism over library and data portals, a base level of simple metadata is required; this can be provided by Dublin Core and, CSMD can provide Dublin Core metadata (Matthews, Wilson & Kleese van Dam, 2002). Further work would include presenting the model as Dublin Core via an application profile (Ball, 2009).

---

[5] The Dublin Core Metadata Initiative: http://www.dublincore.org

[6] PREMIS Data Dictionary for Preservation Metadata (version 2.0):
http://www.loc.gov/standards/premis/

Further we would see the CSMD as a core component with base elements for representing data. We propose to extend the metadata structure in a modular fashion to cover areas such as associating publications to datasets, providing access to simulation data as well as experimental data, and providing more detail on secondary analysis data, as part of an effort to further integrate the ICAT with analysis tools. Further, the role of ontologies to provide controlled vocabularies to improve annotation and search for particular facilities (e.g., catalogues of instruments and beamlines) or within particular disciplines has also been considered, but not fully supported. Another area under active consideration is providing a richer security model. The current system provides a simple access-control system based on membership of teams and their roles. However, facilities are moving towards setting formal data policies, and these should be mapped onto the datasets in a rule based system.

Future considerations on the use of the CSMD will consider the requirements of Digital Curation (preservation, enrichment and availability) upon the metadata record; metadata population strategies in the scientific process; and re-expression as an ontology. Experience to date has shown that the CSMD covers a wide area of scientific research work in sufficient detail in a robust yet usable fashion. We would anticipate that the model would be suitable as a common core for other more domain specific metadata models; ultimately to allow the rich discovery and exploitation of the scientific record into the future.

Acknowledgements

References

Androulakis, S., Buckle, A.M., Atkinson, I., Groenewegen, D., Nicholas, N., Treloar, A., et al (2009). ARCHER – e-research tools for research data management. *International Journal of Digital Curation*, *4*(1).

Ashby, J.V., Bicarregui, J.C., Boyd, D.R.S., Kleese van Dam, K., Lambert, S.C., Matthews, B.M., et al. (2001a). The CLRC Data Portal. *British National Conference on Databases, 2001*.

Ashby, J.V., Bicarregui, J.C., Boyd, D.R.S., Kleese van Dam, K., Lambert, S.C., Matthews, B.M., et al. (2001b). A Multidisciplinary Scientific Data Portal. *HPCN 2001: Int. Conf. on High Performance and Networking Europe, Amsterdam, 2001*.

Ball, A. (2009). *Scientific data application profile scoping study report.* University of Bath, UK: UKOLN. Retrieved from https://pims.jisc.ac.uk/outputs/view/2619

Blanshard, L., Kleese van Dam, K., & Dove, M. (2003). Environment from the
    molecular level e-science project and its use of CLRC's web services based
    data portal. *Proceeding of the 1st. International Conference on Web Services.*

Blanshard, L., Tyler, R., & Kleese van Dam, K. (2004). eMaterials: Integrating grid
    computation and data management services. *UK e-Science Programme All
    Hands Meeting (AHM2004), Nottingham.*

Coles, S.J., Frey, J.G., Hursthouse, M.B., Light, M.E., Milsted, A.J., Carr, L.A., et al.
    (2006). An e-Science environment for service crystallography -from
    submission to dissemination. *Journal of Chemical Information and Modeling,
    Special Issue on eScience.*

Flannery, D., Matthews, B., Griffin, T., Bicarregui, J., Gleaves, M., Lerusse, L., et al.
    (2009). ICAT: Integrating data infrastructure for facilities based science. In
    Proceedings 5th IEEE International Conference on e-Science, Oxford, UK.

Gavaghan, D.J., Simpson, A.C., Lloyd, S., Mac Randal, D.F., & Boyd, D.R.S. (2005).
    Towards a Grid infrastructure to support integrative approaches to biological
    research. *Philosophical Transactions of the Royal Society Series A, 363,* 1829-
    1841.

Hoeck, H., Thiemann, H., Lautenschlager, M., Jessel, I., Marx, B., & Reinke, M.
    (1995). *The CERA metadata model: Technical report no. 9.* Hamburg,
    Germany: DKRZ - German Climate Computer Centre. Retrieved from
    http://www.mad.zmaw.de/fileadmin/extern/documents/reports/ReportNo.09.pdf)

ISO 11179. (2004). *Information Technology -- Metadata Registries (MDR) -- Part 1:
    Framework*

Lawrence, B.N., Lowry, R., Miller, P., Snaith, H., & Woolf, A. (2009). Information in
    environmental data grids. *Philosophical Transactions of the Royal Society A,
    367,* 1003-1014.

Matthews, B.M., Wilson, M.D., & Kleese van Dam, K. (2002). Accessing the Outputs
    of Scientific Projects. In *Proceedings of Current Research Information Systems
    2002, Kassel, Germany.*

Sharman, N., Alpdemir, N., Ferris, J., Greenwood, M., Li, P., & Wroe, C. (2004). The
    my grid information model. *UK e-Science All Hands Meeting 2004
    Nottingham, England.*

Sufi, S., & Matthews, B. (2004). *CCLRC Scientific Metadata Model:Version 2*
    (DLTechnical Reports, DL-TR-2004-001, 2004). Retrieved from
    http://epubs.cclrc.ac.uk/work-details?w=30324

Sufi, S., & Matthews, B. (2005). The CCLRC scientific metadata model: A metadata model for the exploitation of scientific studies and associated data. In D. Talia, A. Bilas, & M. Dikaiakos (Series Eds.), *CoreGRID Series: Vol. 3. Contributions in knowledge and data management in grids*. Springer-Verlag.

Sufi, S., Matthews, B., & Kleese van Dam, K. (2003, September, 2 - 4). An interdisciplinary model for the representation of scientific studies and associated data holdings. *UK e-Science All Hands meeting, Nottingham.*