Accessing the Outputs of Scientific Projects

Brian M Matthews, Michael D Wilson, Kerstin Kleese-van Dam
CLRC, UK

Summary

We describe a science data portal for generic access to scientific data. This data portal is uses a *generic science metadata format* to catalogue and access science data from a range of disciplines. We describe the metadata format that is used and further discuss how this can be used in combination with library metadata formats, such as the Dublin Core, to access all the outputs of scientific projects, both data and publications.

1 Introduction

The scientific research projects have two major outputs: traditional publications, in journals and other forms of literature; and the data sets generated during the course of observations and experiments. These are then subject to analysis and visualisation to generate the results reported in the literature. Traditionally, science has concentrated on the former output as the major means of disseminating the results of research, whilst access to the latter has been restricted to small groups of individuals closely associated with the original researcher. However, modern distributed information systems offer the opportunity to provide access to both outputs to a wider audience. This allows other researchers to verify the results of the analysis, and also to reuse the data-sets to carry out secondary analysis, possibly in combination with results from elsewhere, to produce new insights without the cost of repeating the original experiment.

These data resources are stored in many file systems and databases physically distributed throughout organisations with, at present, no common way of accessing or searching them to find what data is available. It is often necessary to open and read the actual data files to find out what information they contain. There is little consistency in the information which is recorded for each data-set held and sometimes this information may not even be available on-line, being recorded only in experimenters' logbooks. This situation creates the potential for serious under-utilisation of these data resources or to the wasteful re-generation of data. It also hinders the development of cross-discipline research, as this requires good facilities for locating and combining relevant data across traditional disciplinary boundaries.

To address these problems, the concept of a *data portal* has been developed (Ashby et al. 2001a, 2001b; Houstis & Lalis 2001; NESSTAR; Ryssevik & Musgrave 1999). This offers a single method of browsing and searching the contents of scientific data resources, across a variety of scientific domains. Such a system has potentially a wide spectrum of users, from scientists working in related fields wanting to find information on a topic, through experimenters interesting in accessing and analysing their own data, to the data curators based at the facilities themselves who want to use the portal as a data management tool. In order to construct such tools, including mechanisms for cataloguing, browsing and accessing data resources, a generic metadata model for scientific data is needed. Such a metadata for science has the requirement of being both more specific than general metadata models such as the Dublin Core (Dublin Core), whilst being more general than specific metadata formats for specific domains in science, such as earth observation (Hoeck et. al. 1995). There are many metadata formats usually supporting specific data sources; a mechanism needs to be defined to access such metadata in an interoperable way from the generic metadata that preserves the meaning, and allows deeper searches into the domain specific metadata. This approach also differs from generic representations of *science data* such as XSIL (XSIL) that has elements to represent arrays and tables, but little capability to represent provenance data and other information *describing* the science data.

A common metadata format for scientific data also allows the possibility of providing a single point of access to both the major outputs of science: data and publications. By using the common or interoperable features of the generic scientific metadata model, we allow the possibility of combined searches across both domains, or alternatively, using the metadata from one domain (say scientific publications) to search and access appropriate information from the other (say retrieve relevant data sets to test the claims of the publication).

We describe a briefly describe a Science Data Portal developed in CLRC. As a major component of this project a metadata model was defined. In the main body of this paper, we describe in some detail the structure of this metadata both in its overall structure, and some of the details. Further we then discuss how this metadata model can be related to metadata formats for cataloguing

2 A Science Data Portal

A pilot system has been developed to test these ideas that enables researchers to access and search metadata about data resources held at the ISIS and SRS facilities within CLRC, and further extended to cover the British Atmospheric Data Centre (BADC). The system being developed has 3 main components: a web-based user interface; a metadata catalogue; and generic data resource interfaces. These are integrated using standard Web protocols. It is anticipated that the system will exploit the emerging Grid Service infrastructure to offer a distributed interface to scientific data resources both inside and outside CLRC.

The data resources accessible through the data portal system may be located on any one of a number of data servers. Interfaces between these existing data resources and the metadata catalogue are being implemented as *wrappers* on web services that will present the relevant metadata about each resource to the catalogue so it appears to the user to be part of the central catalogue. These wrappers are implemented as XML encoding of the specific metadata relating to that resource using the metadata model schema; wrappers are an established technique for providing such interfaces (Baru et. al 1999).

3 A Metadata Catalogue

The logical structure of the metadata in the catalogue is based on the scientific metadata model developed in the project. This model exploits experience gained in developing general metadata models for other domains, such as the Data Documentation Initiative for social science (DDI) and has the overall structure of 6 major areas as shown in Figure 1. This structuring is influenced by the classification of metadata given in (Jeffery 2000). The study metadata corresponds to *associative descriptive metadata,* the access condition to *associative restrictive metadata,* data description to a form of *schema metadata* (describing how the data is laid out in the file structure), data location to *navigational metadata,* and related material to *associative supportive metadata.*

It is necessarily very generic to cater for a large range of differing types of data; specialisations of this metadata format will be used for each domain; generic queries can be then devised to search over the common views on the metadata. The model uses a hierarchical model of the structure of scientific research programmes, projects and studies, and also generic model of the organisation of data sets into collections and files. This allows a flexible structure to be developed, relating different data sets and their components together. For example related sets derived from one another from raw data through data reduction and analysis to a final result; alternative and failed analyses can also be recorded, as well as calibration data sets, against which results are measured.

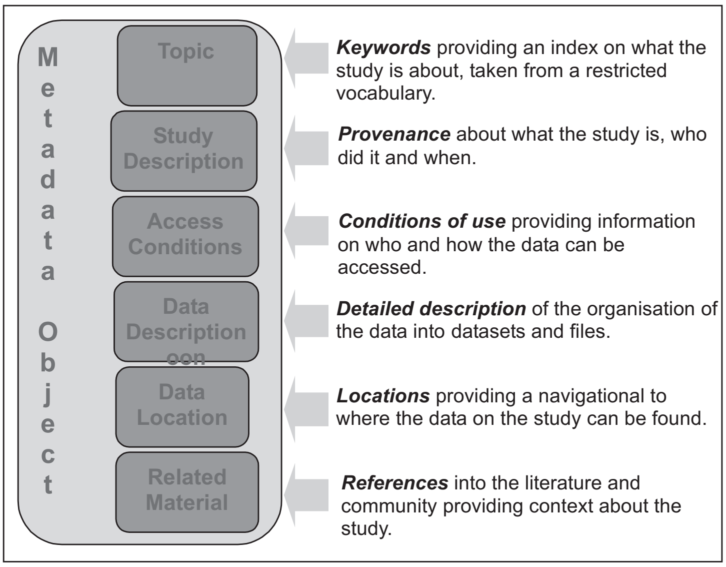
Metadata Object
Topic
Study Description
Access Conditions
Data Description
Data Location
Related Material
Keywords providing an index on what the study is about, taken from a restricted vocabulary.
Provenance about what the study is, who did it and when.
Conditions of use providing information on who and how the data can be accessed.
Detailed description of the organisation of the data into datasets and files.
Locations providing a navigational to where the data on the study can be found.
References into the literature and community providing context about the study.

Figure 1: Overall Metadata Structure

The metadata catalogue is implemented using a standard relational database. Once the specific data sets required by the user have been identified using the available metadata, the catalogue provides links to the files holding the actual data. Users can then use these links to access the data with their own applications for analysis as required.

4 The Metadata Structure

The metadata within the general metadata structure is 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.

4.1 Modelling Scientific Activity

The data model attempts to capture scientific activities at different levels: generically, all activities are called ***Studies***. Each study has an ***Investigator*** that describes who is undertaking the activity, and the ***Study Information*** that captures the details of this particular study. The general structure of the metadata is given as a UML diagram in Figure 2.

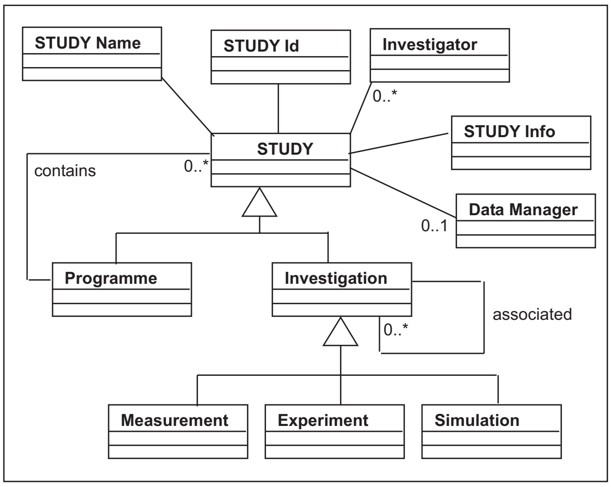

Figure 2: The UML model for the Study

Studies can be of different kinds, as represented by the subclass information in the UML diagram.

- ***Programmes:*** are studies that have a common theme, and usually a common source of funding, instigated by a principal investigator or institution. Programmes can be single projects (such as EPSRC projects, or application for beam time on ISIS), linked sequences of projects; for example an EPSRC Faraday project would have a set of linked projects. Each programme can thus be associated (linked) with a series of sub-investigations. Programmes are not expected to have direct links to data, but rather through the set of investigations within the programmes. ***Investigations:*** are studies that have links directly to data holdings. More specific types of investigations include experiments, measurements or simulations.
- ***Experiments:*** investigations into the physical behaviour of the environment usually to test an hypothesis, typically involving an instrument operating under some instrumental settings and environmental conditions, and generating data sets in files. E.g. the subjection of a material to bombardment by X-Rays of known frequency generated by the Synchrotron Radiation Source with the result diffraction pattern recorded.
- ***Measurements:*** investigations that record the state of some aspect of the environment over a sequence of point 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 a computer program with some initial parameters, and generate a dataset representing the result of the simulation. E.g. a computer simulation of fluid flow over a body using a specific program, with input parameters the shape of the body, and the velocity and viscosity of the fluid, generating a data set of fluid velocities

Each investigation has a particular purpose and uses a particular experimental set up of instruments or computer systems. Experiments may be organised within larger studies or projects, which themselves may be organised into programmes of linked studies.

Classes within the model have several fields. For example, within 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 the following table.

Study Description Class Fields	
Funding	Source of funds of the study, including grant-funding body.
Time	Date, time and duration of study. Can be either a point time and date, or a begin time and end time. We expect it to be in a standard format: dd/mm/yyyy for dates; hh:mm:ss for times.
Purpose	Description of purpose of study, including • Free text abstract of investigation • Keywords categorising subject of investigation – preferably selected from a controlled vocabulary. • Study type: a field that can be used to indicate the type of study being undertaken – such as a calibration run.
Status	Status of study, (*not-started*, *in progress, complete…*).
Resources	Statement of the resources being used, e.g. which facility.

4.2 Modelling scientific data holdings

The metadata format given here is designed for use on general scientific data holdings. These 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.

An investigation is a study that generates raw data. This raw data can then be processed via a set of tools, forming on the way intermediate data sets, which may or may not be held in the data holding. The final processing step generates the final analysed data set. At each stage of the data process stores data in a set of physical files with a physical location. It is possible that there may be different versions of the data sets in the holding. In a general data portal, all stages of the process should be held and available as reviewers of the data holdings may wish to determine the nature of the analysis performed, and other scientist may wish to use the raw data to perform different analyses. Thus each *data holding* takes the form of a hierarchy: one *investigation* generates a *sequence* of logical *data sets*, and each data set is instantiated via a *set* of physical files. The design of the metadata model is tailored to capture such an organisation of data holdings. A single metadata record in this model can provide sufficient metadata to access all the components of the data holding either all together or separately.

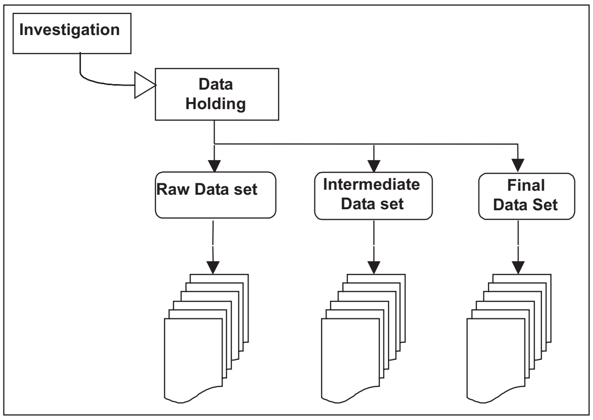

Figure 3: Model of the hierarchy of scientific data holdings

- ***Access control***: Access is controlled by the access entry in the metadata record; how this is actually done is dependent on the data holder. For example, there might be an access type, with settings such as "*open*", "*on application*", "*restricted*", "*commercial in confidence*". This may be given in conjunction with explicit instructions on how to access the data, and who to contact.
- ***Data Location***: The data location provides a mapping between the URI's used in the data definition component of the metadata model, and the actual URL's of the files. This can provide facilities for describing mirror location for the whole structure, and also for individual files.

5 Example

As an example of this scientific metadata model, consider the SXD information from the ISIS neutron spallation source. A *study* in this case is an application for beam-time, uniquely identified with an 'RB number', which covers a programme of investigations, and is described by a description of the purpose in the original study application. This programme is in turn broken down into a series of individual investigations, each of which are experiments on the SXD detector. Each investigation may have a sequence of *runs*, each generating a data set. Each run keeps the major parameters of the experiment the same (e.g. temperature of study), but alter some other parameter (e.g. orientation of the sample in the target).

For example, consider an investigation has name *Benzene*, *variable temperature study: 150K*. It should have a unique ID - this is not necessarily the RB number as that may relate to a programme of investigations, but it might be generated from it. It will have associated with it a set of RAW files, for example: files SXD10091, SXD10092, SXD10093, SXD10094, SXD10095: Benzene, variable temperature study: 150K. There may also be a set of intermediate SXD files, and also a set of processed final files in standard data formats for specific programs, such as .HKL, .INS and .RES files. The system keeps track of the relationships between files, and records which have been processed. We give a small sample of the fields in the metadata. We use *#classname* to represent cross-references between classes.

Experiment	
StudyID	SXD10091
Study Name	Benzene, variable temperature study: 150K
Investigator	*#investigator*
Study Information	*#study-information*
Data holder	*#data-holder*
Instrument	*#instrument*
Environmental Conditions	*#conditions*

The Investigator gives details of the people involved in the study.

Investigator	
Name	Anne X. Perimenter
Institution	University of Somewhere
Status	Lecturer
Role	Principal Investigator
Address	Dept of Organic Chemistry, Univ of Somewhere, UK.

Study information gives the information on this study.

Study Information	
Funding Source	EPSRC
Time	1/11/00, 11.45
Purpose	*#purpose*
Status	Complete
Resources	Beam time on ISIS using the SXD, for 1hr on 1/11/00

The Purpose itself may have several fields.

Purpose	
Abstract	To study the structure of Benzene at a temperature of 150K.
Keywords	Chemistry: organic: benzene: denatured benzene, C6H6

The data holder refers to the institution principally responsible for holding the data - this is not a locator in the sense of a URL.

Data Holder	
Institution	ISIS, CLRC Rutherford Appleton Laboratory

The conditions in this case just record the temperature under which the sample has been studied.

Conditions	
Temperature	150K

Files may also be in several different locations, separating out the identity of data sets from the location. Giving filetype/directory pairs does this:

Data location	
Data holding locations	ftp://ftp.isis.rl.ac.uk/SXD/ SXD1009/http://www.dooc.uos.ac.uk/~perimenter/bezene/
Data set Directories	(RAW, "raw/"), (Intermediate, "SXD/"), (HKL, "HKL/"),…

The data description would break down into a hierarchy of entries. Firstly the top-level entry, which contains references to the data sets of the study.

Data description	
Data Sets	*#raw, #intermediate, #processed*

Then the raw data set would have references to the metadata for each file (not the file itself):

Raw	
Dataset type	RAW
Files	#SXD10091.RAW,#SXD10092.RAW, …

Each file would have an entry, giving its URI:

SXD10091.RAW	
URI	SXD10091.RAW

There will also be a dataset entry for intermediate and processed files.

6 Mapping to Dublin Core

The data portal offers the potential for integrating the outputs of scientific research, thus producing a combined portal for literature and data. Thus a feature of this would be not only the linking of publications to the data set which they depend upon, but the use of literature to guide a more general search for appropriate data in the area, and also from data to appropriate literature which could be used for further analysis. Clearly, to enable this, the metadata formats of the two systems will have to be related to enable searches from one metadata system to be passed to the other. Clearly, there is much commonality between the generic science metadata used in the data portal with generic formats proposed for library systems, especially Dublin Core and CERIF.

The 15 standard elements of the Dublin Core all have their counterparts within the much more details structure used within the Data Portal, and through Dublin Core's "dumbing-down" principle can easily be abstracted, although potentially with little precision.

Mapping between Dublin Core and Data Portal Science Metadata

Dublin Core Element	***Science Metadata Class path and attribute***
Title	Study: Name
Creator	Study: Investigator: Name (Role is principle investigator)
Subject	Topic: Keyword
Description	Study: Study Information: Purpose
Publisher	Investigation: Data Manager
Contributor	Study: Investigator: Name ; Investigation: Data Manager
Date	Study: Study Information: Time
Resource Type	If a data holding is being referenced, this should be set to *Collection*; if a single data-set, then this should be set to *Dataset*.
Format	Data Description: File Format
Resource Identifier	Study: Study Id (for the whole study) Data description: File: URI (for individual data files).
Source	Data description: Data sets: Related Data sets Related Material: Related work
Language	*Not covered in the current metadata format; but an simple extension*
Relation	Related Material: Related work
Coverage	Data description: Logical Description: Coverage
Right Management	Access Conditions

Thus a common search can be set up between the CLRC Data Portal and Dublin Core enabled library catalogues. The more complex model provided by CERIF (CERIF) provides the opportunity for a more precise mapping of the provenance metadata, and a consequentially more better retrieval. For this to be enabled, a mapping would need to be established between the Data Portal's science metadata format's model of the scientific hierarchy (with programmes, projects, participants, studies and experiments) and CERIF's model using People and Project entities.

7 Project Status and Future Plans

This pilot project was completed at the end of March 2001 with the operation of a working prototype system. The longer-term goal is to extend the system to provide a common user interface to metadata for all the scientific data resources held in CLRC. Work in progress is taking the system embedding the system into the facilities and also extending the range of the portal, for example allowing access to earth observation data via the same portal; a new version has been released in April 2002 (CLRC Data Portal) and it is planned to extend the use of the system to materials science. In this process, the generic science metadata has proven remarkably robust, with only small changes needed.

Beyond this, the publication of scientific data as "grey literature" in its own right, together with its appropriate metadata affords the opportunity of it being curated as part of the "corporate memory" of the research organisation, treated and available as an important asset in its own right, rather than a disposable, and in the medium term, uninterpretable legacy of past activity.

Acknowledgements and Contacts

The CLRC Data Portal Project is part of the CLRC E-Science programme (http://www.e-science.clrc.ac.uk), within the UK Research Council's e-Science Initiative. We would like to thank the Data Portal team who has contributed extensively to the definition of the Science Data Model.

8 References

J V Ashby, J C Bicarregui, DR S Boyd, K Kleese van Dam, S C Lambert, B M Matthews, K D O'Neill. (2001a): The CLRC Data Portal British National Conference on Databases

J V Ashby, J C Bicarregui, D R S Boyd, K Kleese van Dam, S C Lambert, B M Matthews, K D O'Neill (2001b): A Multidisciplinary Scientific Data Portal HPCN 2001: International Conference on High Performance and Networking Europe Amsterdam

C. Baru, A. Gupta, V. Chu, B.Ludäscher, R. Marciano, Y. Papakonstantinou, P. Velikhov, (1999) XML-Based Information Mediation for Digital Libraries, Digital Libraries '99.
www.npaci.edu/DICE/Pubs/dl99-demo.pdf

CERIF: the Common European Research Information Format
www.cordis.lu/cerif/

CLRC Data Portal Project
www.escience.clrc.ac.uk/Activity/ACTIVITY=DataPortal

The Data Documentation Initiative
www.icpsr.umich.edu/DDI/

Dublin Core Metadata Initiative
www.dublincore.org/

H. Hoeck, H. Thiemann, M. Lautenschlager, I. Jessel, B Marx, M. Reinke (1995): The CERA Metadata Model Technical Report No. 9, DKRZ - German Climate Computer Centre,
www.dkrz.de/forschung/reports/report9/CERA.book.html

C. Houstis, S. Lalis, (2001): ARION: An Advanced Lightweight Software System Architecture for accessing Scientific Collections, Cultivate Interactive, no.4,
www.cultivate-int.org/issue4/arion/

K G Jeffery. (2000): Metadata Information Systems Engineering Sjaak Brinkkemper, Eva Lindencrona, Arne Solvberg (Eds), Lecture Notes in Computer Science, Springer Verlag ISBN 1-85233-317-0.

NESSTAR (Networked European Social Science Tools and Resources)
www.nesstar.org

J. Ryssevik, S. Musgrave (1999): The Social Science Dream Machine: Resource discovery, analysis and delivery on the Web, the IASSIST Conference, Toronto,
www.nesstar.org/papers/iassist_0599.html

XSIL: Extensible Scientific Interchange Language,
www.cacr.caltech.edu/SDA/xsil/

9 Contact Information

Brian Matthews
CLRC
Rutherford Appleton Laboratory
Didcot
OX11 0QX
UK

e-mail: b.m.matthews@rl.ac.uk; .d.wilson@rl.ac.uk; k.kleese@dl.ac.uk