CCLRC Technical Report: DL-TR-2004-001

CCLRC Scientific Metadata Model: Version 2

Shoaib Sufi[1], Brian Mathews[2]

[1] e-Science Centre, CCLRC, Daresbury Laboratory

[2] Business and Information Technology Department, CCLRC, Rutherford Appleton Laboratory

August 2004

A general model for the representation of scientific study metadata does not exist. The e-Science enablement of the data holdings of CCLRC requires such a model to allow access to the data resources of the facilities in a uniform way. By proposing a model and an implementation, the adoption of such a system would aid interoperability of scientific information systems in the organisation and perhaps form a specification of the type and categories of metadata that studies should capture about their investigations and the data they produce inside and outside of CCLRC. This will allows further exploitation of scientific Studies and associated datasets, ease citation, facilitate collaboration and allow the easy integration of pre-Grid metadata into a common Grid/e-Science enabled scientific information platform.

Acknowledgments

I would like to thank Kerstin Kleese-van-Dam (for writing the forward and general comments), Philip Couch (for discussions and providing the eCCP1 example in Appendix D), and Lisa Blanshard (for review and comments).

I would also like to thank the following with whom discussions and whose comments have proved fruitful for the production of 'CCLRC Scientific Metadata Model: Version 2': Kevin O'Neal, Glen Drinkwater, Frank Toussaint, Richard Tyer and Trudy Watson.

Foreword

The Council for the Central Laboratory of the Research Councils (CCLRC) in the UK is one of Europe's largest multidisciplinary research support organisations. From its three sites, CCLRC operates several large scale scientific facilities for the UK research and industrial community including accelerators, lasers, telescopes, satellites and supercomputers, which all create copious quantities of data. Currently CCLRC is holding data in access of 60 TB, deployed through a number of data centres including One World Data Centre, three National Data Centres and a range of facilities and instruments for particular user communities. However, it is expected that much more data will be collected in the future with the advent of new instruments, facilities (DIAMOND) and projects (Large Hadron Collidor at CERN), which will lead to data volumes in excess of several PB within the next three to four years.

The data held at CCLRC covers most major science areas e.g. Astronomy, Biology, Chemistry, Environmental Science and Physics. These data resources are stored in many file systems and databases physically distributed throughout the organisation 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, what is recorded for each dataset held and sometimes this information may not even be available on-line, only in experimenters' logbooks. This situation could potentially lead to serious under-utilisation of these data resources or to the wasteful regeneration of data. It could also hinder 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, a web-based data portal is being developed with the aim of offering a single method of browsing and searching the contents of all the CCLRC data resources. Central to the DataPortal is the philosophy that each facility/data centre is responsible for its own data and metadata and that ownership and residence are untouched by the Portal. Each facility/data centre is expected to have or develop its own metadata catalogue, which will provide at least a given set of common information held in a chosen local format. Interaction with the metadata catalogues and interchange between them is based on a metadata model for representing scientific data, which has been developed by the project and version 2 is described in this document. Version 2 improves the representation of indexing information, data hierarchy, data description and relationships between Studies amongst other things.

The preferred solution for the interchange metadata schema would have been the reuse or adaptation of an existing tool, however a distinguishing feature of CCLRC's requirement is the necessary *generality* of the Metadata model. Other metadata approaches are either usually closely associated with a particular scientific domain (e.g. CERA – Environmental Science), or else are metadata design frameworks (e.g. RDF, XMI). Some of the initiatives studied were XSIL from Caltech, the work of the OODT group and the activities surrounding the *Dublin Core*. The Extensible

Scientific Interchange Language (XSIL) is a flexible, hierarchical, extensible, transport language for scientific data objects. At a lower level than CCLRC's metadata model it allows the definition of the data array and the transport of those arrays: it is used on LIGO - Laser Interferometer Gravitational-Wave Observatory (http://www.cacr.caltech.edu/SDA/xsil/index.html). The Object Oriented Data Technology group (OODT) at the Jet Propulsion Lab http://oodt.jpl.nasa.gov is also producing a generic framework for connecting XML based metadata profiles, and uses a CORBA based OO-system to provide a distributed resource location service. A good deal of activity also surrounds the *Dublin Core* metadata http://dublincore.org/. This provides a basic set of elements (15 in the original definition), but is unfortunately not detailed enough for CCLRC's purposes. Elements of CCLRC's model could be mapped onto the Dublin Core - an important feature for interoperability, especially with Digital Libraries. Therefore CCLRC decided to develop its own metadata schema for scientific data.

In the context of scientific data and this paper, we consider metadata to be all the information, additional to the raw data itself, which a potential user of the data would need to know to be able to make full and accurate use of the data in a subsequent scientific analysis.

The raw data may have been collected or generated in many ways, such as by measurements or observation of the environment, by carrying out an analytical experiment or by running a computer simulation. When developing our metadata model, we aimed to provide a high-level generic model, which can be specialized to specific scientific disciplines. The resulting model has been implemented as an XML schema and a relational database model and is currently in use in the prototype implementation of the CCLRC DataPortal (http://www.escience.clrc.ac.uk/web/projects/dataportal).

Kerstin Kleese van Dam

1 INTRODUCTION

This document is an attempt to provide a generic metadata model to describe scientific data holdings from the perspective of Studies.

The CCLRC Scientific Metadata Model (CSMDM) described here provides a high-level generic model, which can be specialised to specific scientific disciplines.

Based on discussions with ISIS, SR and BADC departments at CCLRC and initially tailored to the needs of the data holdings of those facilities. Nevertheless, the model will abstract away from the specific requirements of these facilities in order to capture scientific data from any discipline.

Other influences come from the CIP metadata catalogue for Earth Observation and the DDI metadata description for Social Science data and the CERA data model for meta-information on geo-referenced data[1].

2 APPROACH

As a framework, we follow the categorisation of metadata from Keith Jeffery[2], dividing metadata into three main divisions and several subdivisions beneath that.

- **Schematic**

  The data model: a (logical) description of the structure of the resource, the relationship between the elements of the resource, and any constraints.

  - **Concrete**

    Provides a machine view close to physical representation - data formats, fields, strings, code interfaces.

  - **Abstract**:

    Provides a user view, near the real-world - abstract entities and relationships between them

  - **Mappings**

    Capturing the relationships between levels and domains; one abstract to many concrete.

- **Navigational**

  Provides the information on where resources and their sub-components are located. Often tends to be mixed up in other metadata. Ideally, kept separate from other information.

- **Associative**

All other information about a resource.

- **Descriptive**:

  what the resource is about and where it comes from.

- **Restrictive**:

  how the resource can be used.

- **Supportive**:

  the context in which the resource sits.

The metadata is structured into categories which correspond to this framework.

This provides a set of top-level categories. To specialise to a particular domain of interest, we use an *object inheritance* mechanism, providing an initial class hierarchy of metadata objects in this document.

We shall use UML class diagrams to illustrate the relationships between the components of the metadata model.

We shall also supply an XML Schema binding of the metadata model with example instance documents.

3 THE DOMAIN OF INTEREST

3.1 Modelling Scientific Activity

The data model attempts to capture scientific activities at different levels: At the top level we have *Policies* which are enacted by initiating & maintaining *Programmes* which consist of one or more generic activities called *Studies*. Each *Study* has one or more *Investigations* which can be of different types (e.g. Measurement, Simulation, Experiment etc).

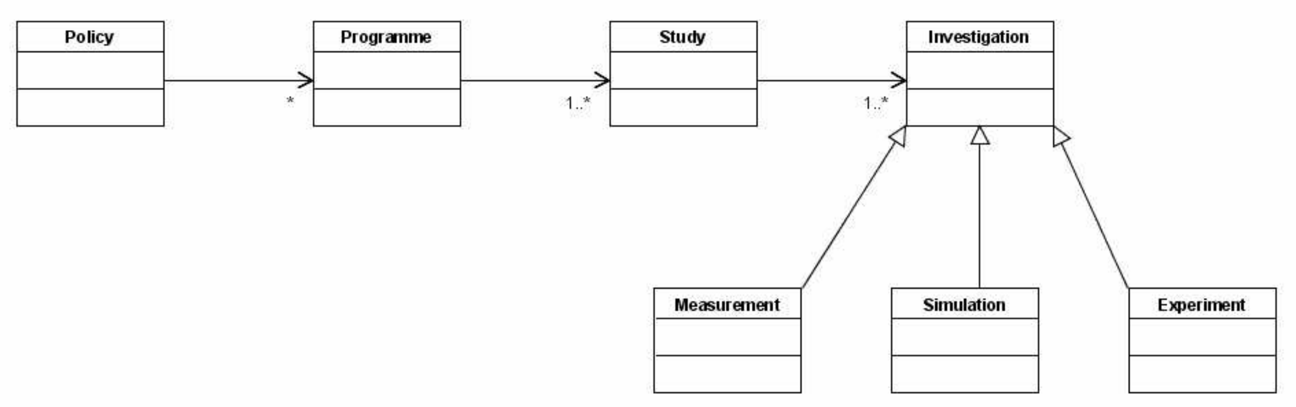

Figure 1: Policy Programmes & Studies

The various entities in Figure 1 are explained in detail below:

Policy[*]: are company or government policies which initiate Programmes of work.

Programmes: are related studies that have a commone theme which are usually funded and resourced directly or with an intermediary organisation under the rubrick of the programme. The UK e-Science Programme[3] is an example of this.

Studies *(sometimes referred to as Projects)*: Studies investigate some aspect of science and have a Principal Investigator and/or institution, co-investigators and are usally funded. e.g. single projects such as EPSRC projects, or application for beam time on ISIS.

[*] Note Policy aggregation is beyond the scope of the model and falls under the realm of knowledge engineering in the social sciences.

Studies consist of investigations and these can be of different types:

Investigations: are sub parts of 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 know frequency generated by the SR source at Daresbury, 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. 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

Virtual Studies: although this is not shown in Figure 1 this is a logical collection of Studies which are linked in someway - i.e. preceding ones and followon ones. This models when a Person or Organisation carries out related work which is funded by different organisation and may have been carried out under different programmes but is linked in some way (e.g. subject matter).

3.2 The Data Holdings

The metadata format given here is designed for use on general scientific data holdings. These data holdings have various layers: the Scientific Programme(if applicable), the Study, the Investigation, the logical data, and the physical data.

Each investigation (experiment, measurement or simulation) has a particular purpose and uses a particular experimental set up of instruments or computer systems.

[*] Further types of investigation may be added. For example, extending the model to social science data may well introduce the additional investigation Survey representing a study comprising of a questionnaire completed by a sample of the population

Experiments may be organised within larger studies or projects, which themselves may be organised into programmes of linked studies.

An investigation generates raw data. E.g. in the ISIS/SR context this raw data can then be processed via set processing tools, forming on the way intermediate stages, which may or may not be held in the data holding. The final processing step generates the final data set.

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.

Thus each *data holding* takes the form of a hierarchy: one *investigation* generates a sequence or hierarchy of logical data collections, and each data collection is instantiated via a set of physical files (or database references).

The design of the metadata model is tailored to capture such an organisation of data holdings; i.e. hierarchical file system type organisation with a top level *Data Holding* and nested *Data Collection* holding references to many *Atomic Data Objects* (i.e. physical files and database queries (e.g. BLOBs)).

A single metadata record in this model can provide sufficient metadata to access all the components of the data holding either all together (e.g. root of the file system) or separately (addressing each file directly).

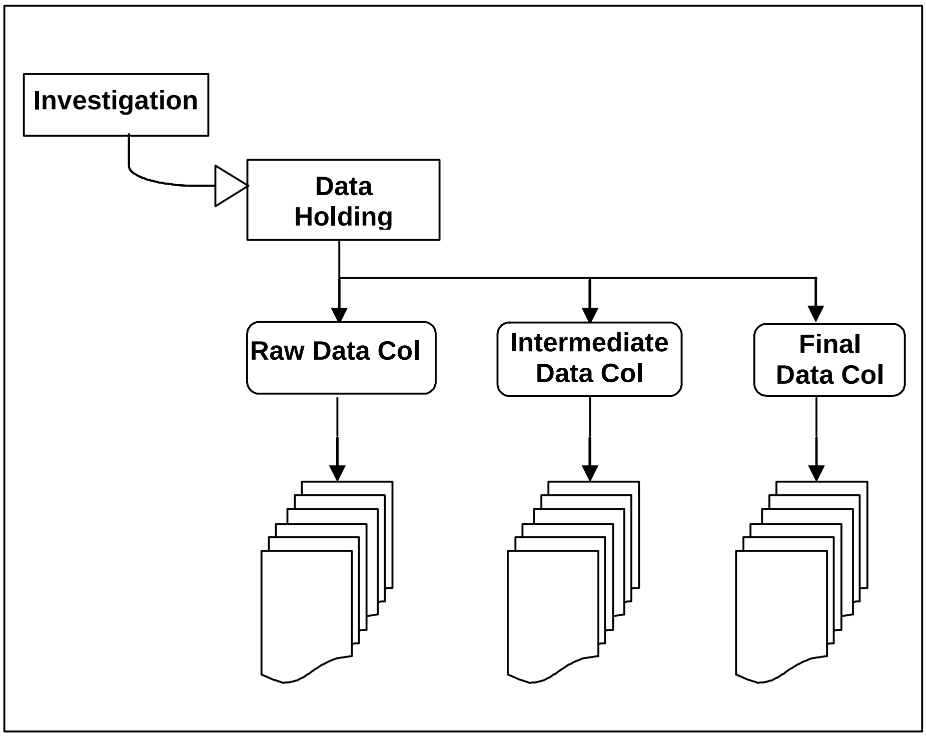

Figure 2: Scientific Data Holdings Example

As an example of this scientific metadata model, consider the SXD information from ISIS.

The *study* in this case would be an application for beam-time, uniquely identified with an 'RB number', which covers a study of investigations, and is described by a description of the purpose in the original study application. This study 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). This information will need to be preserved in the metadata model.

For example an investigation with name *Benzene, variable temperature study: 150K*, would have user, purpose and date and time information associated with it. It should have a unique ID – this is not necessarily the RB number as that may relate to a programme of studies, but it might be generated from it. It will have associated with it a set of RAW files, for example:

Raw File	**Study Name**
SXD10091	Benzene, variable temperature study: 150K
SXD10092	Benzene, variable temperature study: 150K
SXD10093	Benzene, variable temperature study: 150K
SXD10094	Benzene, variable temperature study: 150K
SXD10095	Benzene, variable temperature study: 150K

It may also have a set of intermediate SXD files, and it also may have set of processed final files in standard data formats for specific programs, such as .HKL, .INS and .RES files. The system should keep track of the relationship between files, and record which have been processed and which not.

4 SOME CONVENTIONS

There are some conventions in the way that the following tables describing the fields in the metadata model are expressed.

4.1 Conformance Levels

Level	Description
1	Dublin Core[4] style metadata – essentially Study and Investigations metadata with indexing at the Study level
2	As above and in addition data description and data location metadata
3	As above and in addition Related Material, Access condition, indexing to data collection level (i.e Taxonomy information and Keyword information specific to the data collection level of data organisation)
4	As above and in addition indexing to Atomic Data Object (ADO) Level and ADO parameter information
5	As above and in addition all other metadata about the study; including but not limited to funding, resources used, facilities used etc.

4.2 Type Information

Type Name	**Description**
Complex	Contains other element references and/or a set of simple types; it may have some associated data itself – e.g. an ID of some type and a short name and or description
Enumeration	A value taken from a restricted list/hierarchy/controlled vocabulary
Simple	An atomic type e.g. a string or number
Nested	Contains two values a Simple type and a link to an element of the same type thus supporting classification formation

4.3 Cardinality

Symbol	Description
1	Exactly one occurrence
+	Minimum of 1 occurrence and unbounded maximum occurrences
*	Minimum of 0 occurrences and unbounded maximum occurrences
?	Minimum of 0 occurrences and maximum of 1 occurrence
n…n	A range of occurrences e.g.: 2…4 specifies minimum 2 occurrences and maximum 4 occurrences (note that the range is inclusive) 0…unbounded specifies a minimum of 0 occurrences and an unlimited maximum number of occurrences

4.4 CTS heading

In the table describing each of the Metadata items we use a heading called CTS this stands for:

C	-	Cardinality
T	-	Type
S	-	Suggested Conformance Level

Where Cardinality, Type and Suggested Conformance Level values are as suggested above.

Values are delimited by a colon, a possible example value is 1:Complex:+ - for an entity which is to be populated in conformance level 1 is of Complex type and can have 1 to many occurrences.

This scheme is used to save space.

4.5 Examples / Related Restricted Vocabulary Reference

Note that examples and links to related restricted vocabulary references are only relevant for simple or enumerated types they are not generally applicable for complex types and are therefore usually left blank.

5 CCLRC METADATA RECORD

A single metadata record in the CCLRC Metadata record will have the following format. This will relate all the information for a single *Study*, as defined in Figure 1. A top level *Programme, or Policy* may or not be present. What we present here is how the *Study* metadata is recorded. The Study information in encapsulated in a Scientific Study Metadata Record (this will also be referred to as just Metadata Record). This is show in below in Figure 3.

Each metadata record is provided with its unique identifier within the set of metadata records, which each facility/data archive usually having one set.

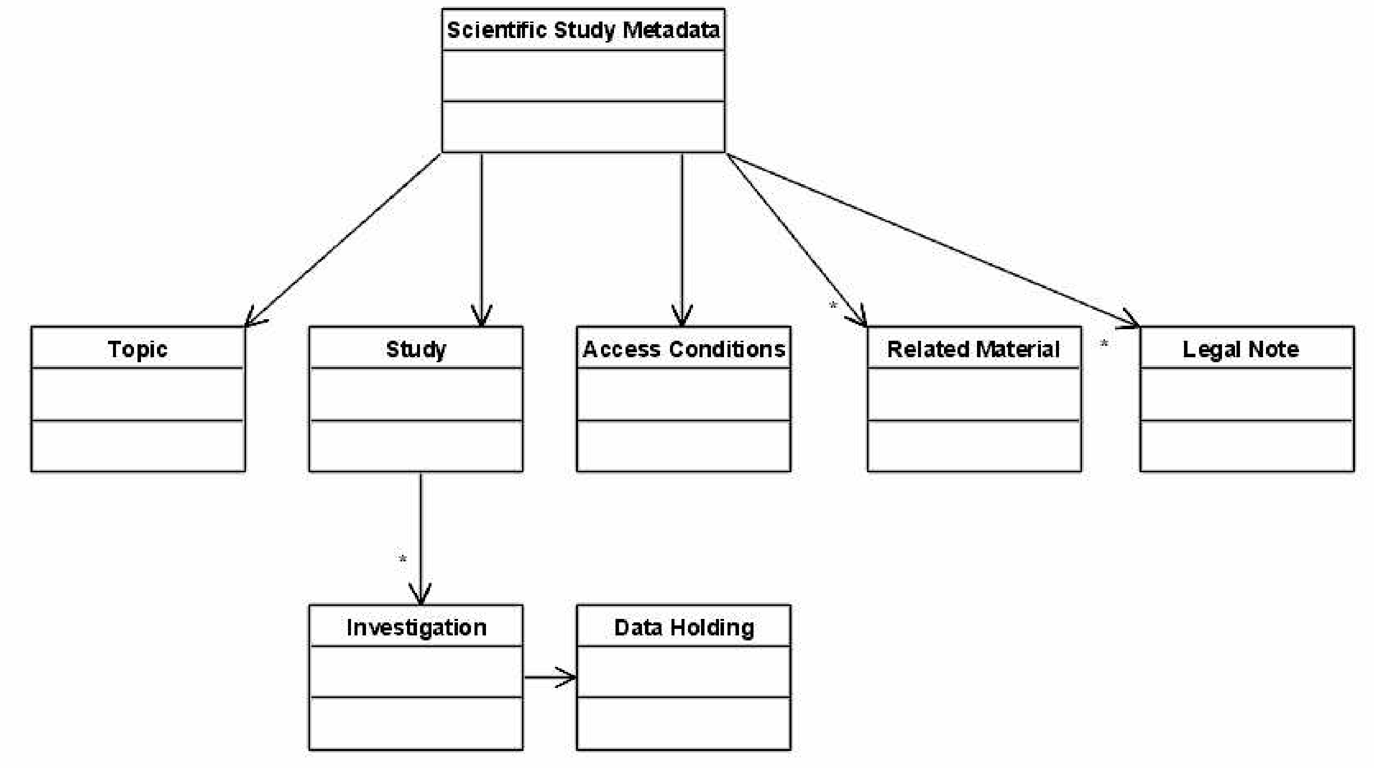

Figure 3: Scientific Study Metadata Hierarchy

The CCLRC Scientific metadata description will contain seven major data areas, forming the top-level categorisation of the metadata. These are as follows:

Metadata Category: **Scientific Study Metadata Record**	**Description**	**CTS**	**Example**	**Example restricted vocabulary reference**
Topic	Keywords from restricted vocabulary giving the hierarchical subject categorisation within each discipline; the discipline and the categorisation within the discipline are taken from controlled vocabularies.	1:Complex:1		
Study	The type of entry which this metadata description is capturing. Description of the study within which the data has been generated. Includes investigator, experimental conditions, and purpose. This also aggregates Investigations and their Data Holdings produced during this study.	1:Complex:1		
Access Conditions	Access rights and conditions on the data referred to within this entry. Includes ownership and access control information.	3:Complex:1		
Related Material	Contextual information: domain definitional information; links to literature, related studies, user communities	3:Comples:*		
Legal Note	A Legal note referring to any Copyright, Patent, Licenses issues regarding this Study and the	4:Complex:*		

	data it contains			

We break down these top-level categories further in the following sections and chapters.

5.1 Scientific Study Metadata Record Meta Information

This is needed to have additional information pertaining to each metadata record to allow easy attribution of the source of the metadata record and uniqueness information. The Metadata Record captured data about the data – i.e. information about the data produced and the Study itself; this Meta Information is needed to describe the Metadata Records themselves such they are easily identifiable, attributable and understandable. Essentially this is a short piece of information which gives the whole records context.

Metadata Category: Metadata Record Meta Information	**Description**	**CTS**	**Example**	**Example restricted vocabulary reference**
Metadata Source	The name of the Metadata Archive Serving the metadata	1:Simple:1		
Metadata ID	A key for the metadata record which is unique in the archive – it Is assumed that this will probably be the same as the Study/Study ID value but it can be different. This could also be a PURL[5]	1:Complex:1		
Metadata ID Scheme	The scheme of the metadata ID e.g. it	1:Enumeration:1		

	is PURL based or just local			
Metadata Conformance	The conformance level of the document	1:Enumeration:1		
Metadata Schema	A pointer to the schema for this metadata record – this could be an XML Schema or a database schema	1:Simple:1		

6 TOPIC

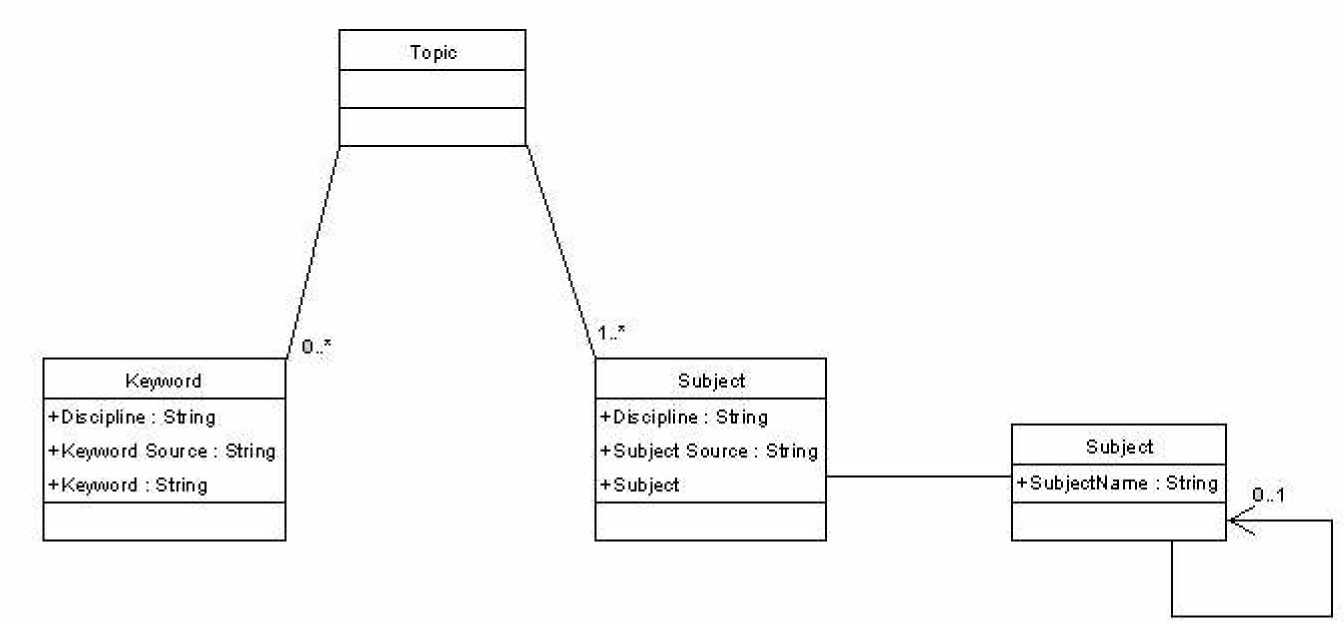

Figure 4

The Topic gives a set of keywords and Subjects relevant to the particular study. Each of the Keyword and Subject entries also states the Discipline with which it is concerned. The Topic thus forms the “encyclopaedia” categorisation of the study.

Keywords will usually come from a particular domain specific restricted vocabulary, and correspond with some named thesaurus or glossary of terms. Thus in the Topic, we allow for not only the keywords, but also the discipline and source of the term. Thus for *each* keyword, we have three fields (the first two are optional):

Metadata Category: Topic	**Description**	**CTS**	**Example**	**Example restricted vocabulary reference**
Keywords	Keywords defined for this Study	2:Complex:*		
Subjects	Subject categorisations for this Study	1:Complex:+		

Metadata Category: Keywords	**Description**	**CTS**	**Example**	**Example restricted vocabulary reference**
Discipline	The Keywords domain (e.g. chemistry, astronomy, ecology etc).	2:Simple:1		
Keyword Source	A pointer (such as URL) to a reference work providing the definition of the restricted vocabulary of which the Keyword list is a subset.	2:Simple:1		
Keyword	List of Keywords which are relevant to this study	2:Simple:+		

The Subject descriptions follow a similar pattern of definition to keywords, however they are more concerned with categorisation of elements of the study than free standing words (e.g. a subject could be /earth sciences/atmosphere/atmospheric temperature/temperature).

Metadata Category: Subjects	Description	CTS	Example	Example restricted vocabulary reference
Discipline	The Keywords domain (e.g. chemistry, astronomy, ecology etc).	1:Simple:1		
Subject Source	A pointer (such as URL) to a reference work providing the definition of the hierarchy of terms for this Discipline of which the Subject hierarchy is an example of a path through.	1:Simple:1		
Subject	A hierarchy of terms which form a classification for this study.	1:Nested:1		

This approach should help overcome inappropriate hits being returned to the user. For example, the term *Field* has quite distinct meanings in Mathematics (an algebraic structure), Physics (the region of influence of some physical phenomenon), and Geography (a region of farmed land). Searches can be qualified by discipline to prevent results in one domain being returned in response to a query in another.

Controlled vocabularies may be arranged in hierarchies or more complicated ontologies. This structure currently does not attempt to capture such structures or reason over them in the search mechanism. This would be the subject of further development of this model.

7 STUDY DESCRIPTION

The study description describes the current investigation being undertaken.

Metadata Category: Study	**Description**	**CTS**	**Example**	**Example restricted vocabulary reference**
Study Name	Full name of the Study	1:Simple:1		
Study Id	One or more identifier of the Study provided as a reference number by the user; e.g. a grant number, or a run number. Could be qualified by the source of the identifier; e.g. EPSRC, ISIS.	1:Simple:1		
Study Institution	Institutions involved in the study and their roles	3:Complex:*		
Investigator	One or more people involved in the study.	1:Complex:+		
Study Information	A description of the study being undertaken	1:Complex:1		
Notes	Any additional miscellaneous notes that may be added to the study.	3:Complex:?		
Investigation	The set of investigations i.e. experiment, measurement, simulation etc involved in this study	1:Complex:+		

7.1 Study Institution

The 'Study Institution' metadata contains information about which Institutions took part in this Study and what their roles were, e.g. data providers, instrument providers.

Metadata Category: Study Institution	**Description**	**CTS**	**Example**	**Example restricted vocabulary reference**
Name	Full name of the Institution	3:Simple:1	CCLRC	
Role	Role of the Institution in the Study	3:Enumeration:1	Data Manager, Data Holder	
Type	The type of institution	3:Enumeration:1	Academic, Government, Commercial	
Id	A key that uniquely identifies this Institution	3:Simple:1		

7.1.1 More on Enumerations

Enumerations (i.e. Standard Terms taken ideally from domain dependent restricted vocabularies, taxonomies, or domain dictionaries) are necessary for instances of the model to work although not being part of the model themselves. e.g. *Role in Study* as in 7.1 above

e.g. consider the following in a collection of Roles and their definitions:

Data manager	A description of the primary organisation(s) responsible for generating, curating and/or holding the data. Not to be confused with a pure data archive, though they may be the same. E.g. the *Data Manager* for meteorological data held at BADC may be the Met Office; BADC is represented with the with the *Data Holder* role.

7.2 The Investigator

The investigator category defines *one or more* people involved in the study; one person can be distinguished as the principle investigator in the study. Others may include experimenters, students, contact staff at the facility, data processors, technicians etc.

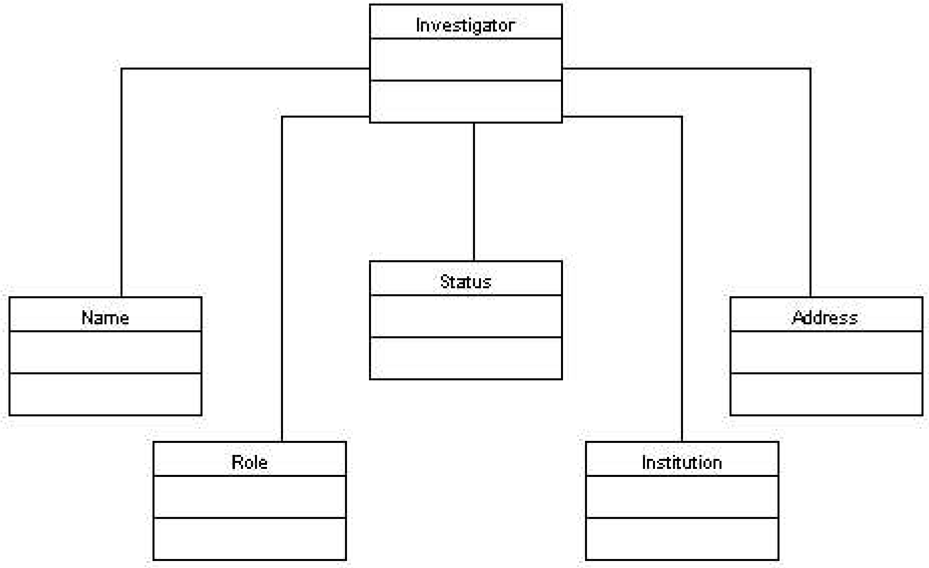

Figure 5: The Investigator and its relations

Each person has the following fields:

Metadata Category: Investigator	**Description**	**CTS**	**Example**	**Example restricted vocabulary reference**
Name	Full name of the Investigator	3:Simple:1		
Institution Id	The Id of the Institution that this person belongs to.	3:Simple:1		
Contact Details	Address and other contact information	3:Complex:1		
Role In Study	Role of the person in this study	3:Enumeration:1	Principal Investigator, Co-Investigator, Experimenter, Data Manager, Data Holder	
Role In Institution	Position of the person within the institution	3:Simple:1	Professor, PDRA, Team Leader	

7.2.1 *Contact Details*

The contact details can be further broken down the following is a suggestion used in the model

Metadata Category: Contact Details	**Description**	**CTS**	**Example**	**Example restricted vocabulary reference**
Address	Full address of the Person	3:Simple:1		
Phone Number	Phone Number and Type of number (e.g. Direct line, Switchboard, Fax)	3:Complex:+		
E-mail Address	e-mail address and type (e.g. work, home)	3:Complex:+		
Web Page	URL of personal web page/pages	3:Comples:+		

7.3 The Study Information

The information about this study has the following fields:

Metadata Category: Study Information	**Description**	**CTS**	**Example**	**Example restricted vocabulary reference**
Funding Source	Source of funds of the study, including	1:Simple:+		

	grant-funding body, and reference number			
Time Line	Start Data, End Date and any other relevant milestone of Study with a note on how long the Study is funded for.	1:Complex:+		ISO 8601[6]
Purpose	Description of purpose of study, including • Free text abstract of investigation • Study type: a field which can be used to indicate the type of study being undertaken.	1:Complex:1	calibration run	
Status of Study	The status of the study.	1:Enumeration:1	Not-started, In Progress, Complete	
Resources	Statement of the resources being used, e.g. which facility.	3:Complex:+		

8 INVESTIGATION

Investigations are carried out in the context of a Study (e.g. many experiments in on particular Study) although sometimes one Study only has one major experiment, or an Monte Carlo simulation could be modelled as one Study with one Investigation (i.e. simulation).

Although Figure 2 shows only Measurements, Simulation and Experiment as sub types of Investigation there could indeed be many more sub-classes of Investigation (e.g. Calculation) and Measurement, Simulation and Experiment (MSE) could also be sub classed; the benefits of such sub-classing maybe more relevant to an object oriented implementation of the model.

In all cases we have facilities that are used, fixed parameters (often called Conditions) and variable parameters that are measured.

The way CSMDM models parameters in a generic way hopes to suffice the differences between whether the facility was an Instrument (for Experiments) or a High Performance Computer (for a Simulation) and whether the fixed parameters were particular experimental conditions (for Experiments) or input data (for Simulations).

We would expect further subclasses of MSE to be defined for particular facilities that are tailored to the special conditions applying in those cases; e.g. ISIS Experiment, SR Experiment, or possibly experiments using particular instruments/hardware within those facilities.

The model aims to be general therefore when different subclasses of conditions for different classes of experiments are needed these can be expressed in a general way e.g. for ISIS[7] we would have a field for the temperature of the sample would be needed while for LHC[8] we might have a field for the beam-energy of the accelerator these can all be captured by the generic parameter class described in section 9.1.1.1.1.1.

Metadata Category: Investigation	**Description**	**CTS**	**Example**	**Example restricted vocabulary reference**
Name	The Name of the Investigation	1:Simple:1		
Investigation Type	The subclass of the investigation i.e. is this an Experiment, Simulation, Measurement etc	1:Simple:1		
Abstract	A short description of the type of Investigation being carried out and why	1:Simple:1		
Resources	A list of resources – mainly facilities used for this Investigation	1:Simple:+		
Data Holding	A logical hierarchy of the Data Collections and Atomic Data Objects and their directory style grouping. The Data Holding can be considered as the 'root' of the data file/object system.	3:Complex:1		

Note:

An Atomic Data Object is an item considered to be a single entity by the individual generating the data; e.g. this is often a file or a database BLOB but can be a named select on a database (essentially a query on a database with a handle).

9 DATA HOLDING

There is one Data Holding (DH) related to each Investigation in the Study. The Data Holding contains the Data Collections & Atomic Data Object (DC&ADO's). The DH Hierarchy of DC's & ADO's very much mimics a File System hierarchy.

The DH each DC & ADO can have its own type marker (e.g. 'raw', 'intermediate', and 'final') information. DH/DC/ADO can also have their own Data Description (such that indexing and parameter information can be set for each level of information in the hierarchy) with each level inheriting from the prior level. Also each level support a Related Reference type which can store relation information between Studies/Investigations/DH/DC/ADO and the type of relation (e.g. if the current ADO uses an ADO from another Study for it's parameter information or if this Study/Investigation depends upon a previous calibration Study/Investigation).

A file is mapped to an ADO and this would have it's own set of metadata relating to logical names and URI's and parameter settings for this file. E.g. in the Birkbeck[9] protein crystallography archive each image (to all intents and purposes a different file) could have different settings for the Crystal-to-detector distance.

The hierarchy for data layout is show in Figure 6 below.

The different levels of data information (Data Holding, Data Collection and ADO) each have Related Reference Information and Data Description information; this is show in Figure 7 below. This allows archives that do not have ADO level parameter, topic or keyword indexing to be searched for relevant ADO (or Data Collections) based on the metadata stored at higher level Data Collection and Data Holding levels.

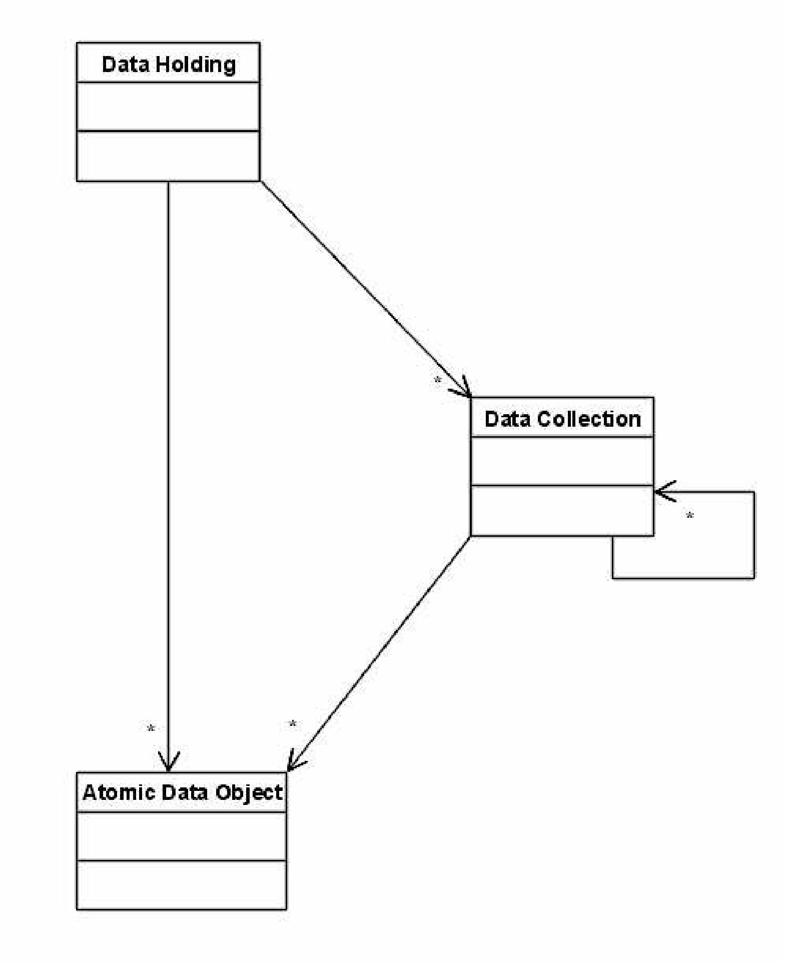

Figure 6: Data Holding Hierarchy

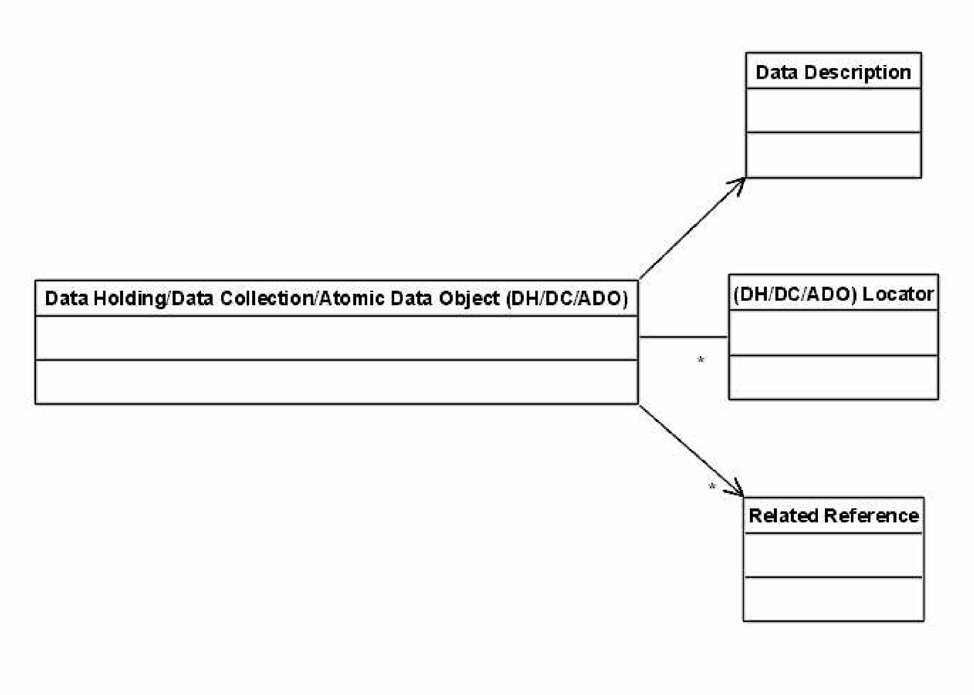

Figure 7: Data Entity Associations

The model supports Logical names with mapping to many physical Locators (e.g. URI/L's) for DHs/DCs/ADOs. Although this allows a basic form of replica location there are other systems such as SRB[10], PURL, and Globus RLS[11] which are more suited to brokering between logical names and physical location and these should be used where such features are required. Hence in terms of physical location we suggest URL to uniquely identify a particular resource (in the case of files) and connection/query information in the case of databases, file replica locating should be deferred to a more specialist service ideally.

The Benefits of using logical names which are resolved by another service are:

- Physical locations of resources can move, while their identity remains the same. By using a storage broker and maintaining its mapping in one place, we can simply maintain the consistency of the metadata.
- A resource identified with one URI can have more than one physical manifestation in the form of copies, caches and mirrors, located using URLs. The map can maintain the correspondence to all of these and thus offer the user a choice of data source.

Thus for each data holding, we give a set of fields:

Metadata Category: Data Holding	**Description**	**CTS**	**Example**	**Example restricted vocabulary reference**
Data Description	A description of the data kept in this data holding from the data archive perspective	3:Complex:1		
Data Collection	Data Collections in the hierarchy of data organisation used in this Investigation; much like directories in a file system and they can be nested.	3:Complex:*		
Atomic Data Object	Atomic Data Objects (files, blobs, named selects etc)	3:Complex:*		
Related Reference	Other Studies/Investigations related to this Data Holding and their type or relationship; e.g. derived from or used by	4:Complex:*		
Data Holding Locator	A locator for addressing the overall Data Holding.	3:Complex:?	URI of top level directory or data	

9.1.1.1 Data Description

Metadata Category: Data Description	Description	CTS	Example	Relevant Reference
Data Name	The logical name of the Data Entity; i.e. Data Holding, Data Collection or Atomic Data Object (ADO)	1:Simple:1		
Type of Data	Is the data a collection or If the data is an ADO then what type of data is it	2:Simple/Enumeration:1	application/ postscript	MIME types[12]
Status	What is the status of the Data Holding, Data Collection or ADO	1:Enumeration:1	Complete, Partial, Under Review, Abrogated, Mirror	
Data Quality	A quality rating for this data – self certified or certified by others with a link (URI) to such certification	5:Complex:*		
Data Topic	As in section 7. Topic this holds taxonomy and keyword indexing information for the DH/DC or ADO.	3:Complex:1		
Logical Description	Reference to a set of logical description fields (e.g. parameter information)	3:Complex:1		

	concerning the DH, DC or ADO			
Software	A description of the software used to produce, analyse, visualise and convert this data produced with reference to program names, web pages describing the software, the relevant Operating Systems and OS versions, and machine Architectures that the software is relevant for.	4:Complex:?		

9.1.1.1.1 Logical Description

Metadata Category: Logical Description	Description	CTS	Example	Example restricted vocabulary reference
Parameter	Parameter information relating to DHs, DCs and ADOs. No difference is made between conditions and measured items; i.e. they are all parameters but have a parameter type qualifier (e.g. fixed parameter & measured in this case). Parameters can be nested allowing parameter aggregation	4:Complex:1		
Time Period	The start and end time associated with this DH, DC or ADO; this could be used synonymously with creation and	4:Comples:1		ISO 8601[6]

	modification times of said data entities. Note if this is the start and end date for a condition of the data i.e. a temporal span covering the simulation output of a data run then this is not what is meant by Time Period and this should be coded as a separate Parameter of the Data Holding / Data Collection / ADO in question (again for date-time we recommend ISO 8601[6])			
Description	A free text (keyword index-able) description of what the DH, DC or ADO contains	4:Simple:1		
Facility Used	A description of which facility and resources at that facility were used to generate the data in the DH, DC or ADO	4:Complex:1		

9.1.1.1.1.1 Parameter

Metadata Category: Parameter	Description	CTS	Example	Relevant Reference
Param Name	The Name of the Parameter or aggregation	4:Complex:1	Temperature, Rate of data capture	
Param Id	The Id of the Parameter such that related parameters can reference each other (e.g. Longitude ranges can reference each other); note they only need be unique for this set of parameters and dictionary order could be used for ordering.	4:Simple:1		
Param Ref	The Id of any related parameters	4:Simple:*		
Param Class	Is the parameter a quantitative measure or a qualitative measure – i.e. are we dealing with values with units or information pertaining to which material this run applies to, which catalyst was used etc.			
Derivation	Is this a fixed, measured, computed (i.e. derived) value etc	4:Enumeration :1		
Units	What are the unit name, unit acronym and	4:Complex:1		

	unit system associated with the value			
Param Value	The value of the parameter or a marker if the parameter denotes an aggregation, although their need not be a value if it is just a description of the Data Collection/ADO	4:Simple:?	30 degrees C, Reading per Day	
Facility Used	A text description of the facility used to generate this value e.g. the instrument in an experiment	4:Simple:1		
Range	The type of limit of this value and the margin of error associated with the value	4:Enumeration :?	Upper, lower	
Parameter	The parameter may consist of sub parameters or be a parameter aggregation which may itself consist of parameters with sub parameters. This self reference allows parameter nesting & hierarchies to be formed.	4:Complex:*		

9.1.1.2 Data Collection

Metadata Category: Data Collection	**Description**	**CTS**	**Example**	**Example restricted vocabulary reference**
Data Description	A description of the data as per section 9.1.1.1	3:Complex:1		
Atomic Data Object	One or more Atomic Data Objects (files, blobs, streams, named selects) etc	3:Complex:*		
Data Collection	This allows nesting of Data Collections	3:Complex:*		
Related Reference	A link to another data set with a description about how it is linked, a collection can have many of these and they are meant to support workflow. Relations types can be taken from enumeration such as *Derived from* Data Collection to show where data was derived from and *Used by* Data Collection to show in which studies data is being used.	4:Complex:*		
Data Collection Locator	Contains the Data Collections name and a locator to reference it; this could either be an absolute locator to the physical data directory or a relative one which would need the Data Holding Locator (and any parent Data Collections) pre-pended to it for	3:Complex:*		

	correct resolution.			

9.1.1.2.1 Atomic Data Objects

The smallest addressable item of storage in the data organisation is the Atomic Data Object (ADO). This is both individually addressable and individually retrievable.

The logical metadata (i.e. name, description, relationship to other ADOs, Derivation from other ADO's) are all the same for different subclasses of Atomic Data Objects.

The types of Atomic Data Object Locators are show in the diagram below (however this is an example and the ADO Locator class can be sub-classed in an implementation dependent manner):

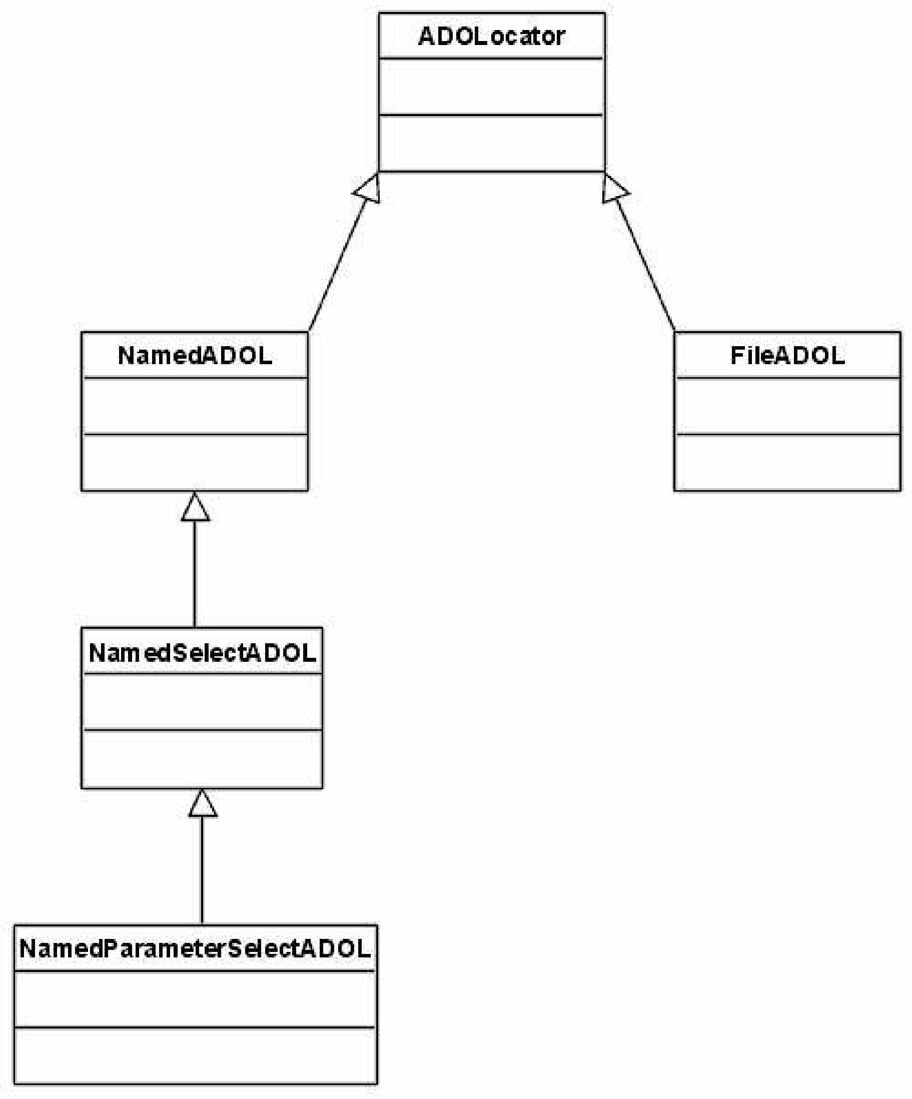

Figure 8: Atomic Data Object Locator Hierarchy

Metadata Category: Atomic Data Objects	**Description**	**CTS**	**Example**	**Example restricted vocabulary reference**
Name	The logical name of the ADO	3:Simple:1		
Data Description	Description of the ADO see 9.1.1.1	3:Complex:1		
Related Reference	Holds a link to another ADO and the relation this has with that ADO taken from an Enumeration; this could be used to track things like where the ADO is being used and where it was derived from	3:Complex:*		
ADO Locator	Contains information to determine the physical location of files, BLOBS in databases and Selects from databases (which have been given a name – equivalent to dynamically generated views).	3:Complex:+		

9.1.1.3 Related Reference

Metadata Category: Related Reference	**Description**	**CTS**	**Example**	**Example restricted vocabulary reference**
Type	The relation between the source of the reference and the referred to item.	3:Simple/Enumeration:1	Derived From, Used By, Prior Study, Follow On Study, Parent Study	
Direction	The logical direction of the reference (e.g. From being a derivation, and To being a usage, Peer being bidirectional)	3:Enumeration :1	From, To, Peer	
Referred to Item	The type of the reference i.e. whether to a Study, Investigation, Data Collection or ADO	3:Enumeration :1	Study, Investigation Data Collection	

			ADO	
Method	A description of the link between the source of the relation and the referred item.	3:Complex:1		
Reference Location	A description of the Physical resources and logical names used to identify the resource	3:Complex:1		

9.1.1.3.1 Reference Location

While a service based referencing system is suggested it could be implemented otherwise.

Metadata Category: Reference Location	**Description**	**CTS**	**Example**	**Example restricted vocabulary reference**
Server	The Name of the host where the Study, Investigation, Data Collection or ADO can be accessed	3:Simple:1		
Port	The host port used for access	3:Simple:?		
Service	The name of the service on the Server at the particular port through which different aspects of Model instances can be accessed e.g. Data Portal based Query and Reply Web service endpoint	3:Simple:1		
Archive Name	Then Name of the Archive which the	3:Simple:1		

	Study/Investigation/Data Collection/ADO metadata belongs to			
Archive Id	The archive id for the metadata being accessed	3:Simple:?		
Study Name	Then Name of the Study being accessed	3:Simple:1		
Study Id	The Id of the Study being accessed	3:Simple:1		
Investigation Name	The Name of the Investigation being accessed	3:Simple:?		
Investigation Id	The Id of the Investigation being accessed	3:Simple:?		
Data Collection	The Name of the Data Collection being accessed	3:Simple:?		
Data Collection Id	The Id of the Data Collection being accessed	3:Simple:?		
ADO Name	The ADO Name being accessed	3:Simple:?		
ADO Id	The ADO Id being accessed	3:Simple:?		

Note

If a Data Collection is being referenced then it is mandatory to have the Investigation Name and/or Investigation Id filled in; i.e. more coarse grained information needs populating if linking to a finer grained entity.

9.1.1.4 ADO Locator

Metadata Category: ADO Locator	**Description**	**CTS**	**Example**	**Example restricted vocabulary reference**
URI	URI pointing to the physical data, this can be relative or absolute (if this is relative then the parent data collection/holding resolved URI is pre-pended to this relative URI for full resolution). Relative URI's only makes sense for file based access.	3:Simple:1		
Size	The size of the data in bytes	4:Simple:1		
Offset	The offset for binary files & blobs taken with Size the actual data section can be extracted	4:Simple:?		

The ADO Locator has various subclasses (see Figure 8) here is the structure of two of the following File ADOL and NamedSelect ADOL.

9.1.1.4.1 File ADOL

This contains the metadata necessary to resolve a file reference. Additional fields stored are:

Metadata Category: File ADOL	**Description**	**CTS**	**Example**	**Example Reference**
URI Type	Whether the URI is relative or absolute	3:Enumeration:?	Relative,	

			Absolute	
Media	A description of the media the file is stored on – e.g. disk, tape and whether this is online, near-line or offline and gives us an indication of the time taken to get the data	4:Complex:?		
File Type	The type of the data in the file	4:Simple:1		MIME types [12]

9.1.1.4.2 Named Select ADOL

This contains the metadata necessary to physically locate and retrieve data from a database with a Select type statement (could equally apply to Relational Databases such as Oracle and XQuery enabled XML databases such as eXist). This would be the common way to extract BLOB information although it is not limited to extracting such information; for anything more than single column/element selects, more metadata is needed about the format of the result type (column names/types or XML Schema) such that results returned from the select could be correctly interpreted; semantic concerns would perhaps require a reference as the data format is essentially dynamically generated.

Metadata Category: Named Select ADOL	Description	CTS	Example	Example Reference
Database Type	The type of database where the data is stored	4:Enumeration :1	Relational, Hierarchical, Object Oriented	
Database Product Name	The name of the database product	3:Simple:1	Oracle, eXist, Xindice, Tamino,	

			PostgresQL, MySQL	
Database Product Version	The version of the database; this is needed as connectivity middleware often requires this information.	3:Simple:1		
Host	The hostname of the machine	3:Simple:1		The DNS name
Instance	The instance id of the database	3:Simple:?		
Database	The name of the database	3:Simple:?		
Port	The IP port on which the database sits	3:Simple:?		
Query	The Query String, this could be in ANSI SQL or W3C XQUERY	3:Simpe:1		
Query Type	The type of the Query	3:Enumeration :1	SQL, XQUERY	
Encoding	The character encoding system of the database	5:Simple:1	UTF-8, UTF-16, ISO 10646, ISO 8859-1	
Result Data Format	The result of the data if this is a BLOB then this can use the same type reference system that files do	3:Simple:?		MIME types[12]
Result Schema	If Query is a complex select (i.e. more than just one column/element) then this is the schema (relation, XML Schema etc) of the	3:Complex:?		

	result and a link to the format used for this description (e.g. the XML Schema standard or a particular databases Data Definition Language).			

10 ACCESS CONDITIONS

A description of the conditions that must be satisfied before a data set can be accessed. Currently, this is left as a place-holder, but would potentially include access control lists, statements on access polices, conditions of use, and information on pricing and payment. Here we suggest one way in which this may be represented, as in Figure 9 below.

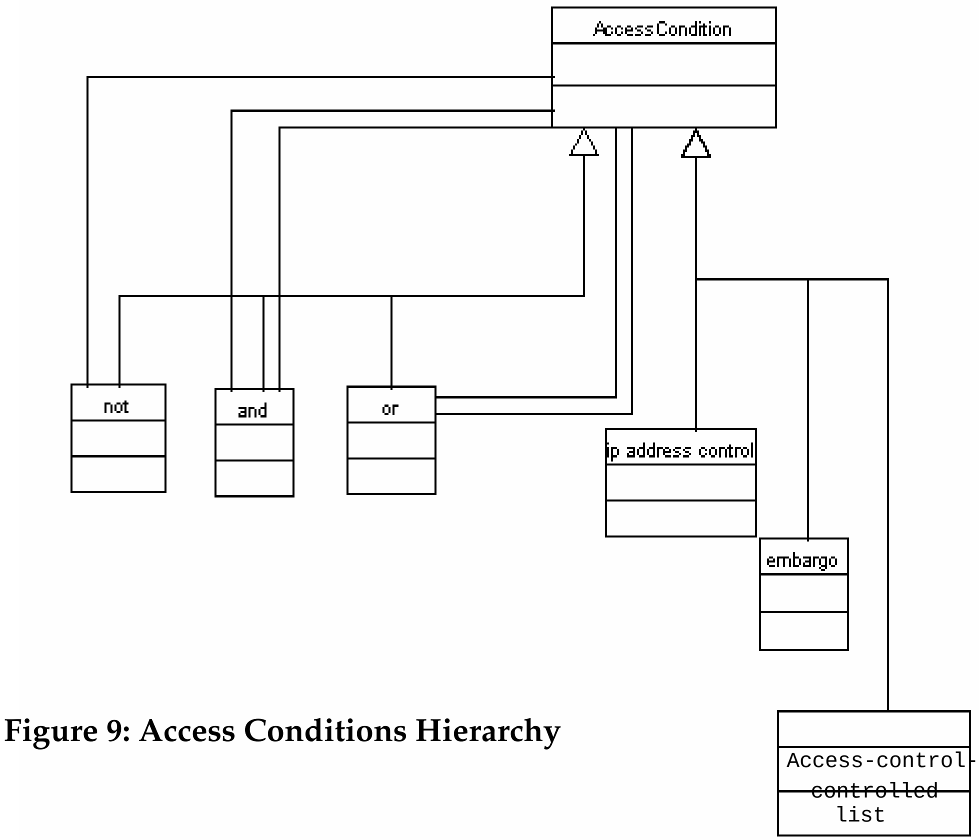

Figure 9: Access Conditions Hierarchy

The model provides generic classes for providing access control features within the metadata model. The *Access Condition* class, defines the access control policy being used for this particular study. Some common access control policies, such as embargo until a fixed date, or testing the IP number of the user, or a lookup on an access-control list. Also some constructors (*and*, *or*, *not*) are provided for combining access conditions. Further access conditions can be provided in specific user packages, such as conditions specified by the archive e.g. ISIS, or SRS. These access control policies can include restricting access to view the metadata itself as well as the data holding.

In many instances it is suggested that this be a link to an Access Control/Authorisation service for this data. Although the granularity of access control could be at applied to each element, the model suggests Topic, Study, DataHolding and Related Material have their own access control. A suggested Authorisation Architecture would be the CCLRC Data Portal Security Architecture[13].

11 RELATED MATERIALS

This category would include contextual information associated with the resource being described.

This could include the following types of information (note the CTS for this is 4:Complex/Simple:* as more information would be needed for related Publications and less for Community Information):

Related Material Metadata Types	Description
Publications	Publications reporting results derived from this Study, this would include information such as Title, Authors and ISBN and possibly URI style location information.
References	References to literature and standards of overall significance to the Study
Community	References to the wider community that is working in this area.

12 EXAMPLES

The appendices contain example schema and instance documents showing implementation and usage of the CSMD Model described in this Document.

- Appendix A - contains an XML Schema based implementation of the CCLRC Scientific Metadata Model
- Appendix B - show an instance document based on an experiment at the ISIS[7] facility which holds its data in files
- Appendix C - shows an instance document based on Measurements done at the MPIM[14] (Meteorology) facility which holds its data in a database.
- Appendix D - shows an instance document based on the eCCP1 (Quantum Chemistry) project which holds its data in a service based XML database with results held in individual documents.

13 CONCLUSION AND FUTURE DIRECTION

The CCLRC Scientific Metadata Model hopes to address the needs for a generic metadata model concerning scientific information; specifically scientific studies and their data holding and supporting their efficient and useful indexing. What we have presented here is hopefully comprehensive and representative of the type of information needed to make studies and the data they produce searchable and useful for reference and perhaps interdisciplinary research. It is hoped that the CSMDM is used as a template for the types of information that need to be stored if not as an actual model.

Currently the model is being used as a template for metadata on the eCCP1[21] project for Quantum Chemistry experiments with relational implementation on the E-Minerals[15] and E-Materials[16] projects, ISIS[7] and externally to the MPIM Meteorology archive in Hamburg. The ISIS 20 year back catalogue data ingestion metadata base is heavily based on CSMDM Version 2. The MyGrid[17] project whose PI is Professor Carol Goble of Manchester University also use the CSMDM in the provenance aspect of their work.

There are other ways in which we hope to improve and extend the model. Efficient ways of supporting synonyms (e.g. Topic Maps), internationalisation and localisation issues are being investigated. We also hope to provide a more fully featured relational implementation compatible with Oracle, PostgreSQL and MySQL databases. Further development and updates to the model and supporting tool set can be monitored by referencing the Multi-disciplinary Scientific Metadata Management e-Science Page (http://www.e-science.cclrc.ac.uk/web/projects/scientific_metadatamgnt). Issues pertaining to long term storage of metadata & quality issues (as related to digital curation e.g. the work of the DCC[18]) and other relevant metadata standard also need to be incorporated into this model to make it more general but hopefully not unwieldy.

Ultimately the model will need to a more semantically rich way of expressing the elements and concepts in scientific metadata management and move toward an ontological representation with example OWL implementation or RDFS/RDF implementation. This will allow for more sophisticated Semantic Web data agent search tools which will allow the exposition of relationships which were previously hard to expose or uncover facilitated by inference engines and the use of description logic.

It is clear that the CSMDM has already proved itself as useful to data management in Earth Sciences, Meteorology, Bioinformatics, Quantum Chemistry, Neutron Spallation and Material Science. It is hoped that further e-Science enablement of CCLRC facilities and its use on other projects will allow the CSMDM to become more general, useful and viewed as standard in Scientific Information Management.

APPENDIX A

This is an XML Schema[19] Implementation of the CCLRC Scientific Metadata Model which is used on the DataPortal[20] project:

```
<?xml version="1.0" encoding="UTF-8"?>

<xsd:schema targetNamespace="http://www.escience.clrc.ac.uk/schemas/scientific"
            xmlns:xsd="http://www.w3.org/2001/XMLSchema"
            xmlns:cmd="http://www.escience.clrc.ac.uk/schemas/scientific"
            elementFormDefault="qualifyied"
            attributeFormDefault="unqualified">

   <xsd:annotation>
      <xsd:documentation xml:lang="en">
         CCLRC Scientific Metadata Model Version 2 - XML Schema implementation
         for CCLRC e-Science DataPortal Project (http://www.e-
science.cclrc.ac.uk/web/projects/dataportal), author: Shoaib Sufi.
      </xsd:documentation>
   </xsd:annotation>

<!-- enumerations defined : -->

   <xsd:simpleType name="AccessControlSystemTypes">
      <xsd:restriction base="xsd:string">
         <xsd:enumeration value="On Application"/>
         <xsd:enumeration value="Digital Access Control System"/>
         <xsd:enumeration value="Other"/>
      </xsd:restriction>
   </xsd:simpleType>



   <xsd:simpleType name="InvestigationTypes">
     <xsd:restriction base="xsd:string">
       <xsd:enumeration value="Experiment"/>
       <xsd:enumeration value="experiment"/>
        <xsd:enumeration value="Measurement"/>
        <xsd:enumeration value="measurement"/>
       <xsd:enumeration value="Simulation"/>
       <xsd:enumeration value="simulation"/>
        <xsd:enumeration value="other"/>
     </xsd:restriction>
   </xsd:simpleType>



   <xsd:simpleType name="institutionTypes">
```

```
  <xsd:restriction base="xsd:string">
     <xsd:enumeration value="academic"/>
     <xsd:enumeration value="research"/>
     <xsd:enumeration value="government"/>
     <xsd:enumeration value="military"/>
     <xsd:enumeration value="commercial"/>
     <xsd:enumeration value="nonprofit"/>
     <xsd:enumeration value="other"/>
  </xsd:restriction>
</xsd:simpleType>



<xsd:simpleType name="TitleTypes">
  <xsd:restriction base="xsd:string">
     <xsd:enumeration value="professor"/>
     <xsd:enumeration value="Professor"/>
     <xsd:enumeration value="Prof"/>
    <xsd:enumeration value="doctor"/>
    <xsd:enumeration value="Doctor"/>
    <xsd:enumeration value="Dr"/>
    <xsd:enumeration value="Mr"/>
    <xsd:enumeration value="Mrs"/>
    <xsd:enumeration value="Ms"/>
     <xsd:enumeration value="other"/>
  </xsd:restriction>
</xsd:simpleType>



<xsd:simpleType name="StudyRoleTypes">
  <xsd:restriction base="xsd:string">
    <xsd:enumeration value="Post Doctoral Research Assistant"/>
    <xsd:enumeration value="pdra"/>
    <xsd:enumeration value="PDRA"/>
    <xsd:enumeration value="PI"/>
    <xsd:enumeration value="Principal Investigator"/>
    <xsd:enumeration value="Co-Investigator"/>
    <xsd:enumeration value="Data Holder"/>
    <xsd:enumeration value="Data Manager"/>
     <xsd:enumeration value="other"/>
  </xsd:restriction>
</xsd:simpleType>



<xsd:simpleType name="InstitutionRoleTypes">
  <xsd:restriction base="xsd:string">
    <xsd:enumeration value="Professor"/>
    <xsd:enumeration value="Senior Lecturer"/>
    <xsd:enumeration value="Lecturer"/>
    <xsd:enumeration value="PDRA"/>
    <xsd:enumeration value="Post Doctoral Research Assistant"/>
    <xsd:enumeration value="PG"/>
    <xsd:enumeration value="Post Graduate"/>
    <xsd:enumeration value="Undergraduate"/>
     <xsd:enumeration value="other"/>
  </xsd:restriction>
</xsd:simpleType>


<xsd:simpleType name="PathTypes">
   <xsd:restriction base="xsd:string">
      <xsd:enumeration value="absolute"/>
      <xsd:enumeration value="file_absolute"/>
      <xsd:enumeration value="relative"/>
      <xsd:enumeration value="file_relative"/>
      <xsd:enumeration value="database"/>
      <xsd:enumeration value="other"/>
   </xsd:restriction>
</xsd:simpleType>


<xsd:simpleType name="FormatSystems">
   <xsd:restriction base="xsd:string">
      <xsd:enumeration value="MIME"/>
      <xsd:enumeration value="custom"/>
      <xsd:enumeration value="database"/>
      <xsd:enumeration value="file"/>
      <xsd:enumeration value="other"/>
   </xsd:restriction>
</xsd:simpleType>


<xsd:simpleType name="AuthorTypes">
   <xsd:restriction base="xsd:string">
      <xsd:enumeration value="primary"/>
      <xsd:enumeration value="co"/>
      <xsd:enumeration value="other"/>
   </xsd:restriction>
</xsd:simpleType>


<xsd:simpleType name="UniqueIdentifierSystem">
   <xsd:restriction base="xsd:string">
```

```
         <xsd:enumeration value="ISBN"/>
         <xsd:enumeration value="SAN"/>
         <xsd:enumeration value="ISMN"/>
      </xsd:restriction>
   </xsd:simpleType>



   <xsd:simpleType name="DerivationTypes">
      <xsd:restriction base="xsd:string">
        <!-- i.e. experimental conditions e.g. date, location : -->
        <xsd:enumeration value="condition"/>
        <!-- i.e. what was actually measured e.g. temperature : -->
        <xsd:enumeration value="measured"/>
        <!-- i.e. a derived parameter from measured/fixed ones e.g. average temperature : -->
        <xsd:enumeration value="calculated"/>
        <!-- i.e. a characteristic of the environment e.g. compiler version  -->
        <xsd:enumeration value="environment"/>
         <xsd:enumeration value="other"/>
      </xsd:restriction>
   </xsd:simpleType>



   <xsd:simpleType name="boundTypes">
      <xsd:restriction base="xsd:string">
         <xsd:enumeration value="upper"/>
        <xsd:enumeration value="lower"/>
         <xsd:enumeration value="other"/>
      </xsd:restriction>
   </xsd:simpleType>

   <xsd:simpleType name="TypesOfData">
      <xsd:restriction base="xsd:string">
        <xsd:enumeration value="Collection"/>
        <xsd:enumeration value="File"/>
        <xsd:enumeration value="BLOB"/>
        <xsd:enumeration value="Database Select"/>
        <xsd:enumeration value="Named Select"/>
         <xsd:enumeration value="other"/>
         <xsd:enumeration value="other"/>
      </xsd:restriction>
   </xsd:simpleType>



<!-- complex types defined : -->
```

```
   <xsd:element name="CLRCMetadata" type="cmd:CLRCMetadataType"/>



   <xsd:complexType name="CLRCMetadataType">
      <xsd:sequence>
         <xsd:element ref="cmd:MetadataRecord" minOccurs="0" maxOccurs="unbounded"/>
      </xsd:sequence>
   </xsd:complexType>



   <xsd:element name="MetadataRecord" type="cmd:MetadataRecordType"/>



   <xsd:complexType name="MetadataRecordType">
      <xsd:sequence>
         <xsd:element name="Topic" type="cmd:TopicType" />
         <xsd:element name="Study" type="cmd:StudyType"/>
         <xsd:element name="AccessConditions" minOccurs="0">
            <xsd:complexType>
               <xsd:simpleContent>
                  <xsd:extension base="xsd:string">
                     <xsd:attribute name="acsystem" type="cmd:AccessControlSystemTypes"/>
                  </xsd:extension>
               </xsd:simpleContent>
            </xsd:complexType>
         </xsd:element>
        <xsd:element name="RelatedPublication" type="cmd:PublicationType" minOccurs="0"
maxOccurs="unbounded"/>
        <xsd:element name="OtherRelatedMaterial" type="cmd:RelatedMaterialTypes" minOccurs="0"
maxOccurs="unbounded"/>
      </xsd:sequence>
      <xsd:attribute name="MetadataID" type="xsd:ID" use="required"/>
      <xsd:attribute name="Facility" type="xsd:string" use="required"/>
   </xsd:complexType>



   <!-- MetadataID uniquely identified one record - logically it consists of the -->
   <!-- name of the data archive (which has to be unique in the dataportal )    -->
   <!-- and something which identifies unique records in the data archive         -->



   <xsd:complexType name="TopicType">
      <xsd:sequence>
         <xsd:element name="Keywords" type="cmd:KeywordsType" minOccurs="0"
maxOccurs="unbounded" />
        <xsd:element name="Subjects"    type="cmd:SubjectsType" minOccurs="1"
maxOccurs="unbounded" />
```

```
      </xsd:sequence>
   </xsd:complexType>



   <!-- perhaps there is a cleverer way to do this by restriction but for now to control
optionality we just use another top level element -->



   <xsd:complexType name="DataTopicType">
      <xsd:sequence>
         <xsd:element name="Keywords" type="cmd:KeywordsType" minOccurs="0"
maxOccurs="unbounded" />
        <xsd:element name="Subjects"    type="cmd:SubjectsType" minOccurs="0"
maxOccurs="unbounded" />
      </xsd:sequence>
   </xsd:complexType>



   <xsd:complexType name="KeywordsType">
      <xsd:sequence>
        <xsd:element name="Discipline" type="xsd:string"/>
        <xsd:element name="KeywordSource" type="xsd:string" minOccurs="0"/>
        <xsd:element name="Keyword" type="xsd:string" maxOccurs="unbounded" />
      </xsd:sequence>
   </xsd:complexType>



   <xsd:complexType name="SubjectsType">
      <xsd:sequence>
        <xsd:element name="Discipline" type="xsd:string" />
        <xsd:element name="SubjectSource" type="xsd:string" minOccurs="0"/>
        <xsd:element name="Subject" type="cmd:SubjectType" />
      </xsd:sequence>
   </xsd:complexType>



   <xsd:complexType name="SubjectType">
      <xsd:sequence>
        <xsd:element name="SubjectName" type="xsd:string"/>
        <xsd:element name="Subject" type="cmd:SubjectType" minOccurs="0"/>
      </xsd:sequence>
   </xsd:complexType>



   <xsd:complexType name="StudyType">
      <xsd:sequence>
         <xsd:element name="StudyName"        type="xsd:string"/>
        <xsd:element name="StudyInstitution"     type="cmd:InstitutionType" minOccurs="0"
maxOccurs="unbounded"/>

         <xsd:element name="StudyPerson"     type="cmd:StudyPersonType" minOccurs="1"
maxOccurs="unbounded"/>

         <xsd:element name="StudyInformation" type="cmd:StudyInformationType"/>

         <xsd:element name="Notes"            type="xsd:string" minOccurs="0" maxOccurs="1"/>

        <!-- contains a link to related studies etc -->

         <xsd:element name="RelatedReference" type="cmd:RelatedReferenceType" minOccurs="0"
maxOccurs="unbounded"/>

        <xsd:element name="Investigation"    type="cmd:InvestigationType" minOccurs="1"
maxOccurs="unbounded"/>

      </xsd:sequence>

      <xsd:attribute name="StudyID" type="xsd:ID" use="required"/>

   </xsd:complexType>



   <xsd:complexType name="InvestigationType">

      <xsd:sequence>

         <xsd:element name="Name" type="xsd:string"/>

        <xsd:element name="InvestigationType" type="cmd:InvestigationTypes"/>

        <xsd:element name="Abstract" type="xsd:string"/>

        <xsd:element name="Resources" type="xsd:string" minOccurs="0" maxOccurs="unbounded"/>

         <xsd:element name="RelatedReference" type="cmd:RelatedReferenceType" minOccurs="0"
maxOccurs="unbounded"/>

        <xsd:element name="DataHolding" type="cmd:DataHoldingType" minOccurs="0"/>

      </xsd:sequence>

      <xsd:attribute name="InvestigationID" type="xsd:ID" use="required"/>

   </xsd:complexType>



   <xsd:complexType name="InstitutionType">

      <xsd:sequence>

         <xsd:element name="Name" minOccurs="0" maxOccurs="1">

         <!-- adding attributes to an element -->

            <xsd:complexType>

               <xsd:simpleContent>

                  <xsd:extension base="xsd:string">

                     <xsd:attribute name="institutionID" type="xsd:string" use="optional"/>

                     <xsd:attribute name="institutiontype" type="cmd:institutionTypes"
use="required"/>

                  </xsd:extension>

               </xsd:simpleContent>

            </xsd:complexType>

         </xsd:element>

        <xsd:element name="Role" type="xsd:string" minOccurs="0"/>

      </xsd:sequence>

   </xsd:complexType>
   <xsd:simpleType name="institutionsType">
      <xsd:list itemType="cmd:institutionTypes"/>
   </xsd:simpleType>



  <xsd:complexType name="PersonType">
     <xsd:sequence>
        <xsd:element name="Name" type="cmd:NameType"/>
        <xsd:element name="InstitutionAffiliatedTo"    type="cmd:ContactDetailsType"
minOccurs="0"/>
        <xsd:element name="ContactDetails" type="cmd:ContactDetailsType"
maxOccurs="unbounded"/>
     </xsd:sequence>
  </xsd:complexType>



  <xsd:complexType name="StudyPersonType">
     <xsd:complexContent>
        <xsd:extension base="cmd:PersonType">
           <xsd:sequence>
              <xsd:element name="RoleInStudy" type="cmd:StudyRoleTypes"/>
              <xsd:element name="RoleInInstitution" type="cmd:InstitutionRoleTypes"
minOccurs="0" />
           </xsd:sequence>
        </xsd:extension>
     </xsd:complexContent>
  </xsd:complexType>



  <xsd:complexType name="NameType">
     <xsd:sequence>
        <xsd:element name="Surname" type="xsd:string"/>
        <xsd:element name="MiddleInitials" type="xsd:string" minOccurs="0"/>
        <xsd:element name="Forename" type="xsd:string"/>
        <xsd:element name="Title" type="cmd:TitleTypes" minOccurs="0" maxOccurs="1"/>
     </xsd:sequence>
  </xsd:complexType>



  <xsd:complexType name="ContactDetailsType">
     <xsd:sequence>
        <xsd:element name="Address"   type="cmd:AddressType"/>
        <xsd:element name="DirectLine" type="xsd:string" minOccurs="0"/>
        <xsd:element name="Switchboard" type="xsd:string"/>
        <xsd:element name="Fax"       type="xsd:string" minOccurs="0" maxOccurs="1"/>
        <xsd:element name="Email"     type="xsd:string" minOccurs="0" maxOccurs="1"/>
        <xsd:element name="WebPage"     type="xsd:string" minOccurs="0" maxOccurs="1"/>
     </xsd:sequence>
  </xsd:complexType>


  <xsd:complexType name="AddressType">
     <xsd:sequence>
        <xsd:element name="Addressline1" type="xsd:string"/>
        <xsd:element name="Addressline2" type="xsd:string" minOccurs="0" />
        <xsd:element name="Addressline3" type="xsd:string" minOccurs="0" />
        <xsd:element name="Addressline4" type="xsd:string" minOccurs="0" />
        <xsd:element name="Town"         type="xsd:string"/>
        <xsd:element name="Region"       type="xsd:string" minOccurs="0" maxOccurs="1"/>
        <xsd:element name="Postcode"     type="xsd:string" minOccurs="0" maxOccurs="1"/>
        <xsd:element name="Country">
           <xsd:complexType>
              <xsd:simpleContent>
                 <xsd:extension base="xsd:string">
                   <!-- country abbreviation candidate for enumeration perhaps -->
                    <xsd:attribute name="countryabbrev" type="xsd:string" use="optional"/>
                 </xsd:extension>
              </xsd:simpleContent>
           </xsd:complexType>
        </xsd:element>
     </xsd:sequence>
  </xsd:complexType>


  <!-- expanding on the study information type -->


   <xsd:complexType name="StudyInformationType">
      <xsd:sequence>
         <xsd:element name="Funding" type="xsd:string" minOccurs="0" maxOccurs="1"/>
         <xsd:element name="TimePeriod" type="cmd:TimePeriodType"/>
         <xsd:element name="Purpose" type="cmd:PurposeType"/>
         <xsd:element name="StudyStatus" type="xsd:string"/>
         <xsd:element name="Resources" type="xsd:string" minOccurs="0" maxOccurs="unbounded"/>
      </xsd:sequence>
   </xsd:complexType>


   <xsd:complexType name="TimePeriodType">
      <xsd:sequence>
         <xsd:element name="StartDate" type="cmd:DateTimeType" minOccurs="0"/>
         <xsd:element name="EndDate"   type="cmd:DateTimeType" minOccurs="0"/>
      </xsd:sequence>
   </xsd:complexType>



   <xsd:complexType name="DateTimeType">
      <xsd:sequence>
<!-- this format will be CCYY-MM-DD as this is the lexical format for the 'date' -->
<!-- datatype in XMLSchema  -->
<!-- the 'date' type also supports an optional timezone component  -->
<!-- the 'date' type adheres to the ISO8601 -->
<!-- standard on representing dates- however the right-truncated format of the date -->
<!-- appears to be only a subset of what is allowable in ISO8601 -->
         <xsd:element name="Date" type="xsd:date"/>
<!-- the format of the time attribute is hh:mm:ss.sss with an optional timezone information
section -->
         <xsd:element name="Time" type="xsd:time" minOccurs="0"/>
      </xsd:sequence>
   </xsd:complexType>



   <xsd:complexType name="PurposeType">
      <xsd:sequence>
         <xsd:element name="Abstract" type="xsd:string" minOccurs="0" maxOccurs="1"/>
      </xsd:sequence>
   </xsd:complexType>



   <xsd:complexType name="DataHoldingType">
      <xsd:sequence>
         <xsd:element name="DataDescription"    type="cmd:DataDescriptionType"/>
         <xsd:element name="DataHoldingLocator" type="cmd:CollectionLocatorType"/>
         <xsd:element name="RelatedReference"   type="cmd:RelatedReferenceType" minOccurs="0"
maxOccurs="unbounded"/>
         <xsd:element name="DataCollection"     type="cmd:DataCollectionType" minOccurs="0"
maxOccurs="unbounded"/>
        <xsd:element name="AtomicDataObject"   type="cmd:ADOType" minOccurs="0"
maxOccurs="unbounded"/>
      </xsd:sequence>
      <xsd:attribute name="InvestigationID" type="xsd:IDREF" use="required"/>
   </xsd:complexType>



   <xsd:complexType name="ParameterType">
      <xsd:sequence>
         <xsd:element name="ParamName" type="xsd:string"/>
        <xsd:element name="Derivation" type="cmd:DerivationTypes"/>
         <xsd:element name="Units" type="cmd:UnitsType" minOccurs="0" maxOccurs="1"/>
         <xsd:element name="ParamValue" type="xsd:string" minOccurs="0" maxOccurs="1"/>
         <xsd:element name="Range" type="cmd:RangeType" minOccurs="0" maxOccurs="1"/>
         <xsd:element name="Parameter" type="cmd:ParameterType" minOccurs="0"
maxOccurs="unbounded"/>
      </xsd:sequence>
   </xsd:complexType>



   <xsd:complexType name="UnitsType">
      <xsd:sequence>
        <xsd:element name="UnitName" type="xsd:string" minOccurs="0"/>
        <xsd:element name="UnitAcronym" type="xsd:string" minOccurs="0"/>
        <xsd:element name="UnitSystem" type ="xsd:string" minOccurs="0"/>
        <xsd:element name="UnitFormat" type ="xsd:string" minOccurs="0"/>
      </xsd:sequence>
   </xsd:complexType>



   <xsd:complexType name="RangeType">
      <xsd:sequence>
         <xsd:element name="Limit" minOccurs="1" maxOccurs="unbounded">
            <xsd:complexType>
               <xsd:simpleContent>
                  <xsd:extension base="xsd:string">
                     <xsd:attribute name="bound" type="cmd:boundTypes" use="optional"/>
                  </xsd:extension>
               </xsd:simpleContent>
            </xsd:complexType>
        </xsd:element>
        <xsd:element name="MarginOfError" type="xsd:string" minOccurs="0"/>
      </xsd:sequence>
   </xsd:complexType>



   <xsd:complexType name="LogicalDescriptionTypes">
      <xsd:sequence>
         <xsd:element name="Parameter" type="cmd:ParameterType" minOccurs="0"
maxOccurs="unbounded"/>
        <xsd:element name="TimePeriod" type="cmd:TimePeriodType" minOccurs="0"/>
        <xsd:element name="Description" type="xsd:string" minOccurs="0"/>
        <xsd:element name="FacilityUsed" type="cmd:FacilityType" minOccurs="0"/>
      </xsd:sequence>
   </xsd:complexType>



   <xsd:complexType name="FacilityType">
      <xsd:sequence>
```

```
         <xsd:element name="FacilityName" type="xsd:string" />
         <xsd:element name="Resource" type="xsd:string" minOccurs="0" maxOccurs="unbounded"/>
      </xsd:sequence>
   </xsd:complexType>



   <xsd:complexType name="DataDescriptionType">
      <xsd:sequence>
         <xsd:element name="DataName"   type="xsd:string"/>
        <xsd:element name="TypeOfData" type="cmd:TypesOfData" minOccurs="0" />
         <xsd:element name="Status"     type="xsd:string" minOccurs="0" />
         <xsd:element name="DataTopic"  type="cmd:DataTopicType" minOccurs="0" />
         <xsd:element name="LogicalDescription" type="cmd:LogicalDescriptionTypes" 
minOccurs="0" />
        <xsd:element name="Software" type="cmd:SoftwareType" minOccurs="0" />
      </xsd:sequence>
   </xsd:complexType>



   <xsd:complexType name="SoftwareType">
      <xsd:sequence>
         <xsd:element name="Production" type="cmd:ProgramType" minOccurs="0" 
maxOccurs="unbounded"/>
         <xsd:element name="Anlaysis" type="cmd:ProgramType" minOccurs="0" 
maxOccurs="unbounded"/>
         <xsd:element name="Conversion" type="cmd:ProgramType" minOccurs="0" 
maxOccurs="unbounded"/>
         <xsd:element name="Visualisation" type="cmd:ProgramType" minOccurs="0" 
maxOccurs="unbounded"/>
         <xsd:element name="MultiPurpose" type="cmd:ProgramType" minOccurs="0" 
maxOccurs="unbounded"/>
         <xsd:element name="other" type="cmd:ProgramType" minOccurs="0" maxOccurs="unbounded"/>
      </xsd:sequence>
   </xsd:complexType>



   <xsd:complexType name="ProgramType">
      <xsd:sequence>
         <xsd:element name="LongName" type="xsd:string" minOccurs="0" />
         <xsd:element name="ProgramName" type="xsd:string" />
         <xsd:element name="Version" type="xsd:string"/>
        <!-- URI should contain a reference to a description of this software e.g. sourceforge 
link for some oss works -->
         <xsd:element name="URI" type="xsd:string"/>
         <xsd:element name="OperatingSystem" type="xsd:string" minOccurs="0"/>
         <xsd:element name="OperatingSystemVersion" type="xsd:string" minOccurs="0" />
         <xsd:element name="Architecture" type="xsd:string" minOccurs="0"/>
      </xsd:sequence>
   </xsd:complexType>
   <xsd:complexType name="DataCollectionType">
      <xsd:sequence>
         <xsd:element name="DataDescription" type="cmd:DataDescriptionType"/>
         <xsd:element name="DataCollectionLocator"    type="cmd:CollectionLocatorType"
minOccurs="0" />
         <xsd:element name="RelatedReference" type="cmd:RelatedReferenceType" minOccurs="0"
maxOccurs="unbounded"/>
        <xsd:element name="AtomicDataObject"            type="cmd:ADOType" minOccurs="0"
maxOccurs="unbounded"/>
         <xsd:element name="DataCollection"           type="cmd:DataCollectionType"
minOccurs="0" maxOccurs="unbounded"/>
      </xsd:sequence>
      <xsd:attribute name="dataid" type="xsd:ID" use="required"/>
   </xsd:complexType>



   <xsd:complexType name="RelatedReferenceType">
      <xsd:sequence>
         <xsd:element name="Type" type="xsd:string"/>
        <xsd:element name="Direction" type="xsd:string" minOccurs="0" />
        <xsd:element name="ReferredToItem" type="xsd:string"/>
        <xsd:element name="Method" type="xsd:string"/>
        <xsd:element name="ReferenceLocation" type="cmd:ReferenceLocationType"
maxOccurs="unbounded"/>
      </xsd:sequence>
   </xsd:complexType>



   <!--service oriented reference location resolution-->
   <xsd:complexType name="ReferenceLocationType">
      <xsd:sequence>
         <xsd:element name="Server" type="xsd:string" minOccurs="0"/>
        <xsd:element name="Port" type="xsd:string" minOccurs="0" />
        <xsd:element name="Service" type="xsd:string" minOccurs="0" />
        <xsd:element name="Archive" type="xsd:string"/>
        <xsd:element name="ArchiveId" type="xsd:IDREF" minOccurs="0"/>
        <xsd:element name="StudyName" type="xsd:string"/>
        <xsd:element name="StudyId" type="xsd:IDREF" minOccurs="0"/>
        <xsd:element name="InvestigationName" type="xsd:string" minOccurs="0"/>
        <xsd:element name="InvestigationId" type="xsd:IDREF" minOccurs="0"/>
        <xsd:element name="DataCollection" type="xsd:string" minOccurs="0"/>
        <xsd:element name="DataCollectionId" type="xsd:IDREF" minOccurs="0"/>
        <xsd:element name="ADOName" type="xsd:string" minOccurs="0"/>
        <xsd:element name="ADOId" type="xsd:IDREF" minOccurs="0"/>
        <!-- for a none service based reference a URI-esque (if it is an absolute) location can
also be stored -->
```

```
         <xsd:element name="Locator" minOccurs="0" maxOccurs="unbounded">
            <xsd:complexType>
               <xsd:simpleContent>
                  <xsd:extension base="xsd:string">
                     <xsd:attribute name="pathtype" type="cmd:PathTypes"/>
                  </xsd:extension>
               </xsd:simpleContent>
            </xsd:complexType>
         </xsd:element>
      </xsd:sequence>
   </xsd:complexType>



   <xsd:complexType name="ADOType">
      <xsd:sequence>
         <xsd:element name="DataDescription" type="cmd:DataDescriptionType"/>
         <xsd:element name="ADOLocator"        type="cmd:ADOLocatorType" minOccurs="0"
maxOccurs="unbounded"/>
         <xsd:element name="RelatedReference" type="cmd:RelatedReferenceType" minOccurs="0"
maxOccurs="unbounded"/>
      </xsd:sequence>
      <xsd:attribute name="dataid" type="xsd:ID" use="required"/>
   </xsd:complexType>



      <!-- how to use the specialisation parent/child class/subclass 'feature' : -->
      <!--
      <root_element
         xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
         ...
      >
      note this assumed that the cmd schema is the default one
      ...

      ...<ADOLocator xsi:type="FileADOL">

      ...

      -->
   <!-- note ADOL stands for Atomic Data Object Locator -->

   <xsd:complexType name="FileADOL">
      <xsd:complexContent>
         <xsd:extension base="cmd:ADOLocatorType">
            <xsd:sequence>
              <xsd:element name="Media" type="xsd:string" minOccurs="0" />
              <xsd:element name="Filetype" type="xsd:string" minOccurs="0"/>
            </xsd:sequence>
         </xsd:extension>
      </xsd:complexContent>
   </xsd:complexType>



   <xsd:complexType name="SelectNamedADOL">
      <xsd:complexContent>
         <xsd:extension base="cmd:ADOLocatorType">
            <xsd:sequence>
               <xsd:element name="DatabaseType" type="xsd:string"/>
               <xsd:element name="DatabaseProductName" type="xsd:string"/>
               <xsd:element name="DatabaseProductVersion" type="xsd:string"/>
               <xsd:element name="Host" type="xsd:string"/>
               <xsd:element name="Instance" type="xsd:string" minOccurs="0"/>
               <xsd:element name="Database" type="xsd:string"/>
               <xsd:element name="Port" type="xsd:string" minOccurs="0"/>
               <xsd:element name="Query" type="xsd:string"/>
               <xsd:element name="Encoding" type="xsd:string" minOccurs="0"/>
               <xsd:element name="DataFormat" minOccurs="0" maxOccurs="unbounded">
                  <xsd:complexType>
                     <xsd:simpleContent>
                        <xsd:extension base="xsd:string">
                           <xsd:attribute name="formatsystem" type="cmd:FormatSystems"
use="optional"/>
                        </xsd:extension>
                     </xsd:simpleContent>
                  </xsd:complexType>
               </xsd:element>
               <xsd:element name="ResultSchema" type="xsd:string" minOccurs="0"/>
            </xsd:sequence>
         </xsd:extension>
      </xsd:complexContent>
   </xsd:complexType>



   <xsd:complexType name="CollectionLocatorType">
      <xsd:sequence>
         <xsd:element name="DataName" type="xsd:string"/>
         <xsd:element name="Locator" minOccurs="0" maxOccurs="unbounded">
            <xsd:complexType>
               <xsd:simpleContent>
                  <xsd:extension base="xsd:string">
                     <xsd:attribute name="pathtype" type="cmd:PathTypes"/>
                  </xsd:extension>
               </xsd:simpleContent>
            </xsd:complexType>
         </xsd:element>
      </xsd:sequence>
   </xsd:complexType>



   <xsd:complexType name="ADOLocatorType">
      <xsd:sequence>
         <xsd:element name="Locator" minOccurs="0" maxOccurs="unbounded">
            <xsd:complexType>
               <xsd:simpleContent>
                  <xsd:extension base="xsd:string">
                     <xsd:attribute name="pathtype" type="cmd:PathTypes"/>
                  </xsd:extension>
               </xsd:simpleContent>
            </xsd:complexType>
         </xsd:element>
         <xsd:element name="AccessMethod" minOccurs="0" maxOccurs="unbounded">
            <xsd:complexType>
               <xsd:simpleContent>
                  <xsd:extension base="xsd:string">
                     <xsd:attribute name="authenticationtype" type="xsd:string"/>
                  </xsd:extension>
               </xsd:simpleContent>
            </xsd:complexType>
         </xsd:element>
         <xsd:element name="Size" type="xsd:string" minOccurs="0"/>
         <xsd:element name="offset"  type="xsd:string" minOccurs="0"/>
         <xsd:element name="length"  type="xsd:string" minOccurs="0"/>
      </xsd:sequence>
   </xsd:complexType>



   <xsd:complexType name="PublicationType">
      <xsd:sequence>
         <xsd:element name="PublicationName" type="xsd:string"/>
         <xsd:element name="Author" minOccurs="1" maxOccurs="unbounded">
            <xsd:complexType>
               <xsd:simpleContent>
                  <xsd:extension base="xsd:string">
                     <xsd:attribute name="authortype" type="cmd:AuthorTypes"/>
                  </xsd:extension>
               </xsd:simpleContent>
            </xsd:complexType>
         </xsd:element>
         <xsd:element name="Identifier" minOccurs="0" maxOccurs="unbounded">
```

```
            <xsd:complexType>
               <xsd:simpleContent>
                  <xsd:extension base="xsd:string">
                     <xsd:attribute name="identsystem" type="cmd:UniqueIdentifierSystem"/>
                  </xsd:extension>
               </xsd:simpleContent>
            </xsd:complexType>
         </xsd:element>
         <xsd:element name="URI" type="xsd:string" minOccurs="0" maxOccurs="unbounded"/>
      </xsd:sequence>
   </xsd:complexType>



   <xsd:simpleType name="References">
      <xsd:restriction base="xsd:string"/>
   </xsd:simpleType>



   <xsd:simpleType name="CommunityInformation">
      <xsd:restriction base="xsd:string"/>
   </xsd:simpleType>



   <xsd:simpleType name="RelatedMaterialTypes">
      <xsd:union memberTypes="xsd:string cmd:References cmd:CommunityInformation"/>
   </xsd:simpleType>



</xsd:schema>
```

APPENDIX B

This is an XML instance document showing data from an ISIS experiment encoded in accordance with the XML Schema in Appendix A. Data in these experiments are organised in files.

```
<?xml version="1.0" encoding="utf-8"?>
<CLRCMetadata xmlns="http://www.escience.clrc.ac.uk/schemas/scientific"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.escience.clrc.ac.uk/schemas/scientific
C:\schemas\clrcmetadata.xsd">


   <MetadataRecord MetadataID="ISIS-SXD1009" Facility="ISIS">
      <Topic>
         <Keywords>
            <Discipline>Chemistry</Discipline>
            <Keyword>Organic</Keyword>
            <Keyword>Benzene</Keyword>
            <Keyword>Denatured Benzene</Keyword>
            <Keyword>C6H6</Keyword>
            <Keyword>C6D6</Keyword>
         </Keywords>
         <Subjects>
            <Discipline>Chemistry</Discipline>
            <Subject>
               <SubjectName>Chemsitry</SubjectName>
               <Subject>
                  <SubjectName>Organic</SubjectName>
                  <Subject>
                     <SubjectName>Benzene</SubjectName>
                  </Subject>
               </Subject>
            </Subject>
         </Subjects>
      </Topic>
      <Study StudyID="SXD1009">
         <StudyName>Benzene, variable temperature study: 150K</StudyName>
         <StudyInstitution>
            <Name institutionID="ISIS" institutiontype="academic">ISIS</Name>
            <Role>Study Institution</Role>
         </StudyInstitution>
         <StudyPerson>
            <Name>
               <Surname>Perimentor</Surname>
               <MiddleInitials>X</MiddleInitials>
               <Forename>Anne</Forename>
               <Title>Professor</Title>
            </Name>
            <ContactDetails>
               <Address>
                  <Addressline1>Department of Organic Chemistry</Addressline1>
                  <Addressline2>University of SomeWhere</Addressline2>
                  <Town>Some Town</Town>
                  <Region>Some Region</Region>
                  <Postcode>S12 4WE</Postcode>
                  <Country countryabbrev="UK">United Kindom</Country>
               </Address>
               <DirectLine>01922 222 2222</DirectLine>
               <Switchboard>01922 222 2000</Switchboard>
               <Fax>01922 222 1111</Fax>
               <Email>a.x.perimentor@chem.someuni.ac.uk</Email>
               <WebPage>http://www.somuni.ac.uk/axperimentor</WebPage>
            </ContactDetails>
            <RoleInStudy>Principal Investigator</RoleInStudy>
            <RoleInInstitution>Professor</RoleInInstitution>
         </StudyPerson>
         <StudyPerson>
            <Name>
               <Surname>Wilson</Surname>
               <Forename>Chick</Forename>
               <Title>Doctor</Title>
            </Name>
            <ContactDetails>
               <Address>
                  <Addressline1>ISIS</Addressline1>
                  <Addressline2>Rutherford Appleton Laboratory</Addressline2>
                  <Addressline3>Chilton</Addressline3>
                  <Town>Didcot</Town>
                  <Region>Oxon</Region>
                  <Postcode>OX11 0QX</Postcode>
                  <Country countryabbrev="UK">United Kindom</Country>
               </Address>
               <DirectLine>01111 111 1111</DirectLine>
               <Switchboard>01111 111 1000</Switchboard>
               <Fax>01111 111 2111</Fax>
               <Email>chick.wilson@rl.ac.uk</Email>
               <WebPage>http://www.rl.ac.uk/wilsonc</WebPage>
            </ContactDetails>
            <RoleInStudy>Data Holder</RoleInStudy>
         </StudyPerson>
         <StudyInformation>
            <Funding>EPSRC</Funding>
```

```
      <TimePeriod>
         <StartDate>
            <Date>2000-11-01</Date>
            <Time>09:00:00-00:00</Time>
         </StartDate>
         <EndDate>
            <Date>2001-11-01</Date>
            <Time>17:30:00-00:00</Time>
         </EndDate>
      </TimePeriod>
      <Purpose>
         <Abstract>To study the structure of Benzene at a temperature of
         150K</Abstract>
      </Purpose>
      <StudyStatus>Complete</StudyStatus>
      <Resources>Beam time on ISIS using the SXD for 1 hour on
      11/01/2001</Resources>
   </StudyInformation>
   <Investigation InvestigationID="INV-SXD1009">
      <Name>Benzene, variable temperature study: 150K</Name>
      <InvestigationType>Experiment</InvestigationType>
      <Abstract>To study the structure of Benzene at a temperature of
      150K</Abstract>
      <Resources>Beam time on ISIS using the SXD for 1 hour on
      11/01/2001</Resources>
      <DataHolding InvestigationID="INV-SXD1009">
         <DataDescription>
            <DataName>INV-SXD1009</DataName>
            <TypeOfData>Collection</TypeOfData>
            <Status>Complete</Status>
            <LogicalDescription>
               <Parameter>
                  <ParamName>Temperature</ParamName>
                  <Derivation>condition</Derivation>
                  <Units>
                     <UnitName>Kelvin</UnitName>
                     <UnitAcronym>K</UnitAcronym>
                  </Units>
                  <ParamValue>150</ParamValue>
               </Parameter>
               <TimePeriod>
                  <StartDate>
                     <Date>2000-11-01</Date>
                     <Time>09:00:00-00:00</Time>
                  </StartDate>
                  <EndDate>
                     <Date>2001-11-01</Date>
                     <Time>17:30:00-00:00</Time>
                  </EndDate>
               </TimePeriod>
               <FacilityUsed>
                  <FacilityName>ISIS</FacilityName>
                  <Resource>ISIS SXD</Resource>
               </FacilityUsed>
            </LogicalDescription>
         </DataDescription>
         <DataHoldingLocator>
            <DataName>INV-SXD1009</DataName>
            <Locator pathtype="absolute">
            ftp://ftp.isis.rl.ac.uk/SXD/SXD1009/</Locator>
            <Locator pathtype="absolute">
            http://www.dooc.uos.ac.uk/~perimentor/benzene/</Locator>
         </DataHoldingLocator>
         <DataCollection dataid="INV-SXD1009-DC-RAW">
            <DataDescription>
               <DataName>RAW</DataName>
               <TypeOfData>Collection</TypeOfData>
               <Status>Complete</Status>
            </DataDescription>
            <DataCollectionLocator>
               <DataName>RAW</DataName>
               <Locator pathtype="relative">raw/</Locator>
            </DataCollectionLocator>
            <AtomicDataObject dataid="INV-SXD1009-DC-RAW-ADO-1">
               <DataDescription>
                  <DataName>SXD10091.RAW</DataName>
                  <TypeOfData>File</TypeOfData>
                  <Status>Complete</Status>
               </DataDescription>
               <ADOLocator xsi:type="FileADOL">
                  <Locator pathtype="relative">SXD10091.RAW</Locator>
               </ADOLocator>
            </AtomicDataObject>
            <AtomicDataObject dataid="INV-SXD1009-DC-RAW-ADO-2">
               <DataDescription>
                  <DataName>SXD10092.RAW</DataName>
                  <TypeOfData>File</TypeOfData>
                  <Status>Complete</Status>
               </DataDescription>
               <ADOLocator xsi:type="FileADOL">
                  <Locator pathtype="relative">SXD10092.RAW</Locator>
               </ADOLocator>
```

```
                  </AtomicDataObject>
                  <AtomicDataObject dataid="INV-SXD1009-DC-RAW-ADO-3">
                     <DataDescription>
                        <DataName>SXD10093.RAW</DataName>
                        <TypeOfData>File</TypeOfData>
                        <Status>Complete</Status>
                     </DataDescription>
                     <ADOLocator xsi:type="FileADOL">
                        <Locator pathtype="relative">SXD10093.RAW</Locator>
                     </ADOLocator>
                  </AtomicDataObject>
               </DataCollection>
               <DataCollection dataid="INV-SXD1009-DC-INT">
                  <DataDescription>
                     <DataName>Intermediate</DataName>
                     <TypeOfData>Collection</TypeOfData>
                     <Status>Complete</Status>
                  </DataDescription>
                  <DataCollectionLocator>
                     <DataName>Intermediate</DataName>
                     <Locator pathtype="relative">SXD/</Locator>
                  </DataCollectionLocator>
                  <RelatedReference>
                     <Type>Derived</Type>
                     <Direction>From</Direction>
                     <ReferredToItem>Collection</ReferredToItem>
                     <Method>Software Processor</Method>
                     <ReferenceLocation>
                        <Archive>ISIS</Archive>
                        <StudyName>Benzene, variable temperature study:
                        150K</StudyName>
                        <InvestigationName>Benzene, variable temperature study:
                        150K</InvestigationName>
                        <DataCollection>RAW</DataCollection>
                        <DataCollectionId>INV-SXD1009-DC-RAW</DataCollectionId>
                        <Locator pathtype="relative">../raw/</Locator>
                        <Locator pathtype="absolute">
                        ftp://ftp.isis.rl.ac.uk/SXD/SXD1009/raw/</Locator>
                     </ReferenceLocation>
                  </RelatedReference>
                  <AtomicDataObject dataid="INV-SXD1009-DC-INT-ADO-1">
                     <DataDescription>
                        <DataName>SXD10091.SF</DataName>
                        <TypeOfData>File</TypeOfData>
                        <Status>Complete</Status>
                     </DataDescription>
                     <ADOLocator xsi:type="FileADOL">
                        <Locator pathtype="relative">SXD10091.SF</Locator>
                     </ADOLocator>
                  </AtomicDataObject>
                  <AtomicDataObject dataid="INV-SXD1009-DC-INT-ADO-2">
                     <DataDescription>
                        <DataName>SXD10092.SF</DataName>
                        <TypeOfData>File</TypeOfData>
                        <Status>Complete</Status>
                     </DataDescription>
                     <ADOLocator xsi:type="FileADOL">
                        <Locator pathtype="relative">SXD10092.SF</Locator>
                     </ADOLocator>
                  </AtomicDataObject>
               </DataCollection>
            </DataHolding>
         </Investigation>
      </Study>
      <AccessConditions acsystem="On Application">The user must be a registered
      user of the ISIS facility. To register, apply to ISIS, CCLRC, Rutherford
      Appleton Lab, UK (http://www.isis.rl.ac.uk/)</AccessConditions>
   </MetadataRecord>
</CLRCMetadata>
```

APPENDIX C

This is an XML instance document showing data from an MPIM experiment encoded in accordance with the XML Schema in Appendix A. Data in these experiments are stored in a database.

```
<?xml version="1.0" encoding="utf-8"?>
<CLRCMetadata xmlns="http://www.escience.clrc.ac.uk/schemas/scientific"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.escience.clrc.ac.uk/schemas/scientific
C:\schemas\clrcmetadata.xsd">


   <MetadataRecord MetadataID="MPIM-ECHAM3_T42_22032AMIP" Facility="MPIM">
      <Topic>
         <Subjects>
            <Discipline>Earth Sciences</Discipline>
            <SubjectSource>
            http://www.pik-
potsdam.de/dept/dc/e/sdm/cera/Descriptions/TABLES/TOPIC_DKRZ.html</SubjectSource>
            <Subject>
               <SubjectName>ATMOSPHERE</SubjectName>
               <Subject>
                  <SubjectName>Atmospheric Temperature</SubjectName>
                  <Subject>
                     <SubjectName>Temperature Anomalies</SubjectName>
                  </Subject>
               </Subject>
            </Subject>
         </Subjects>
      </Topic>
      <Study StudyID="ECHAM3_T42_22032AMIP">
         <StudyName>T42_ECHAM3_22032AMIP</StudyName>
         <StudyInstitution>
            <Name institutionID="MPIM" institutiontype="research">Max-Planck
            Institute for Meteorology</Name>
            <Role>Study Institution</Role>
         </StudyInstitution>
         <StudyPerson>
            <Name>
               <Surname>Bengtsson</Surname>
               <Forename>Lennart</Forename>
            </Name>
            <ContactDetails>
               <Address>
                  <Addressline1>Max Planck Institute for
                  Meteorology</Addressline1>
                  <Addressline2>Bundesstra&#223;e 53</Addressline2>
                  <Town>20146 Hamburg</Town>
                  <Country>Germany</Country>
               </Address>
               <Switchboard>(+49 40) 41173 - 0</Switchboard>
               <Fax>(+49 40) 41173 - 298</Fax>
               <Email>bengtsson@dkrz.de</Email>
               <WebPage>http://www.mpimet.mpg.de/~bengtsson.lennart/</WebPage>
            </ContactDetails>
            <RoleInStudy>Principal Investigator</RoleInStudy>
            <RoleInInstitution>Professor</RoleInInstitution>
         </StudyPerson>
         <StudyInformation>
            <Funding>EPSRC</Funding>
            <TimePeriod>
               <StartDate>
                  <Date>2001-11-26</Date>
                  <Time>19:19:25+01:00</Time>
               </StartDate>
            </TimePeriod>
            <Purpose>
               <Abstract>The purpose of the experiment was to evaluate the long
               term atmospheric response to pre- scribed SST anomoaly patterns
               over the period 1979-1992 by carrying out a series of simula-
               tions with the same atmospheric model but starting from
               different atmospheric initial states The ECHAM3 model was
               essentially finalized in late 1992 and has since then been
               extensively used in different climate experiments. Together with
               a large number of other modelling groups, the ECHAM3 model has
               also taken part in the so called AMIP investigations ("1992:
               AMIP:The atmospheric model intercomparison project, Bull. Am.
               Met. Soc., 73 (12), 1962-1970," Gates, 1992). The model has
               mainly been used at low and medium horizontal resolutions, T21
               and T42, in addition shorter integrations of 5 years duration
               have been done at a T106 resolution. Five different integrations
               were than carried out all using different atmospheric initial
               states. The initial data were all obtained from control runs
               with climatological SST starting several month beforehand. The
               original archive was generated with a time resolution of 12
               hours; the majority of the investigations in this study has been
               using a special archive of seasonally averaged data. The central
               objective of this study is to explore the impact and importance
               of SST anomalies of the low frequency circulation of the
               atmosphere."(Climate predictability experiments with a general
               circulation model, Max-Planck_Report No. 145, Bengtsson, Arpe,
               Roeckner, Schulzweida). This is the control run which produces
               the initial states for the five experiments:
               ECHAM3_T42_22056HMBG, ECHAM3_T42_22062LLNL,
               ECHAM3_T42_23101AMIP, ECHAM3_T42_23102AMIP,
               ECHAM3_T42_23103AMIP. The URL address of the ECHAM3 model
               description is http://www.dkrz.de/forschung/reports-eng.html
               Report No. 6 ECHAM3-Atmospheric General Circulation
               Model.</Abstract>
            </Purpose>
            <StudyStatus>Complete</StudyStatus>
         </StudyInformation>
         <Investigation InvestigationID="INV-ECHAM3_T42_22032AMIP">
            <Name>ECHAM3_T42_22032AMIP</Name>
            <InvestigationType>Simulation</InvestigationType>
            <Abstract>As Study Abstract</Abstract>
            <DataHolding InvestigationID="INV-ECHAM3_T42_22032AMIP">
               <DataDescription>
                  <DataName>ECHAM3_T42_22032AMIP</DataName>
                  <TypeOfData>Collection</TypeOfData>
                  <Status>Complete</Status>
                  <LogicalDescription>
                     <Parameter>
                        <ParamName>Location</ParamName>
                        <Derivation>condition</Derivation>
                        <ParamValue>World (general)</ParamValue>
                     </Parameter>
                     <FacilityUsed>
                        <FacilityName>MPIM</FacilityName>
                     </FacilityUsed>
                  </LogicalDescription>
               </DataDescription>
               <DataHoldingLocator>
                  <DataName>ECHAM3_T42_22032AMIP</DataName>
                  <Locator pathtype="database">Database</Locator>
               </DataHoldingLocator>
               <DataCollection dataid="ECHAM3_T42_22032AMIP_T2M">
                  <DataDescription>
                     <DataName>ECHAM3_T42_22032AMIP_T2M</DataName>
                     <TypeOfData>Collection</TypeOfData>
                     <Status>Complete</Status>
                     <LogicalDescription>
                        <Parameter>
                           <ParamName>latitude</ParamName>
                           <Derivation>condition</Derivation>
                           <Units>
                              <UnitName>degrees</UnitName>
                           </Units>
                           <Range>
                              <Limit bound="upper">87.8638</Limit>
                              <Limit bound="lower">-87.8638</Limit>
                           </Range>
                        </Parameter>
                        <Parameter>
                           <ParamName>longitude</ParamName>
                           <Derivation>condition</Derivation>
                           <Units>
                              <UnitName>degrees</UnitName>
                           </Units>
                           <Range>
                              <Limit bound="upper">357.1875</Limit>
                              <Limit bound="lower">0.0</Limit>
                           </Range>
                        </Parameter>
                        <Parameter>
                           <ParamName>altitude</ParamName>
                           <Derivation>condition</Derivation>
                           <Units>
                              <UnitName>Hectopascal</UnitName>
                           </Units>
                           <ParamValue>2.0</ParamValue>
                        </Parameter>
                        <Parameter>
                           <ParamName>Location</ParamName>
                           <Derivation>condition</Derivation>
                           <ParamValue>World (general)</ParamValue>
                        </Parameter>
                        <Parameter>
                           <ParamName>start date</ParamName>
                           <Derivation>condition</Derivation>
                           <Units>
                              <UnitSystem>ISO8601</UnitSystem>
                              <UnitFormat>YYYY-MM-DD</UnitFormat>
                           </Units>
                           <ParamValue>1911-01-01</ParamValue>
                        </Parameter>
                        <Parameter>
                           <ParamName>end date</ParamName>
                           <Derivation>condition</Derivation>
                           <Units>
                              <UnitSystem>ISO8601</UnitSystem>
                              <UnitFormat>YYYY-MM-DD</UnitFormat>
                           </Units>
                           <ParamValue>1941-01-01</ParamValue>
```

```
         </Parameter>
      </LogicalDescription>
   </DataDescription>
   <DataCollectionLocator>
      <DataName>ECHAM3_T42_22032AMIP_T2M</DataName>
      <Locator pathtype="database">database</Locator>
   </DataCollectionLocator>
   <AtomicDataObject dataid="DH-ECHAM3_T42_22032AMIP_T2M-ADO1">
      <DataDescription>
         <DataName>ECHAM3_T42_22032AMIP_T2M</DataName>
         <TypeOfData>Named Select</TypeOfData>
         <Status>Complete</Status>
      </DataDescription>
      <ADOLocator xsi:type="SelectNamedADOL">
         <AccessMethod authenticationtype="GSI">
         OGSA-DAI</AccessMethod>
         <DatabaseType>Relational</DatabaseType>
         <DatabaseProductName>Oracle</DatabaseProductName>
         <DatabaseProductVersion>
         9.2.0.1</DatabaseProductVersion>
         <Host>cldb.dkrz.de</Host>
         <Instance>mpim</Instance>
         <Database>cera4</Database>
         <Port>5651</Port>
         <Query>Select Data From
         ECHAM3_T42_22032AMIP_T2M</Query>
         <DataFormat formatsystem="MIME">
         application/wmo-grib</DataFormat>
      </ADOLocator>
   </AtomicDataObject>
</DataCollection>
<!-- -->
<DataCollection dataid="ECHAM3_T42_22032AMIP_U10M">
   <DataDescription>
      <DataName>ECHAM3_T42_22032AMIP_U10M</DataName>
      <TypeOfData>Collection</TypeOfData>
      <Status>Complete</Status>
      <LogicalDescription>
         <Parameter>
            <ParamName>latitude</ParamName>
            <Derivation>condition</Derivation>
            <Units>
               <UnitName>degrees</UnitName>
            </Units>
            <Range>
               <Limit bound="upper">87.8638</Limit>
               <Limit bound="lower">-87.8638</Limit>
            </Range>
         </Parameter>
         <Parameter>
            <ParamName>longitude</ParamName>
            <Derivation>condition</Derivation>
            <Units>
               <UnitName>degrees</UnitName>
            </Units>
            <Range>
               <Limit bound="upper">357.1875</Limit>
               <Limit bound="lower">0.0</Limit>
            </Range>
         </Parameter>
         <Parameter>
            <ParamName>altitude</ParamName>
            <Derivation>condition</Derivation>
            <Units>
               <UnitName>Hectopascal</UnitName>
            </Units>
            <ParamValue>10.0</ParamValue>
         </Parameter>
         <Parameter>
            <ParamName>Location</ParamName>
            <Derivation>condition</Derivation>
            <ParamValue>World (general)</ParamValue>
         </Parameter>
         <Parameter>
            <ParamName>start date</ParamName>
            <Derivation>condition</Derivation>
            <Units>
               <UnitSystem>ISO8601</UnitSystem>
               <UnitFormat>YYYY-MM-DD</UnitFormat>
            </Units>
            <ParamValue>1911-01-01</ParamValue>
         </Parameter>
         <Parameter>
            <ParamName>end date</ParamName>
            <Derivation>condition</Derivation>
            <Units>
               <UnitSystem>ISO8601</UnitSystem>
               <UnitFormat>YYYY-MM-DD</UnitFormat>
            </Units>
            <ParamValue>1941-01-01</ParamValue>
         </Parameter>
      </LogicalDescription>
   </DataDescription>
   <DataCollectionLocator>
      <DataName>ECHAM3_T42_22032AMIP_U10M</DataName>
      <Locator pathtype="database">database</Locator>
   </DataCollectionLocator>
   <AtomicDataObject dataid="ECHAM3_T42_22032AMIP_U10M-ADO1">
      <DataDescription>
         <DataName>ECHAM3_T42_22032AMIP_U10M</DataName>
         <TypeOfData>Named Select</TypeOfData>
         <Status>Complete</Status>
      </DataDescription>
      <ADOLocator xsi:type="SelectNamedADOL">
         <AccessMethod authenticationtype="GSI">
         OGSA-DAI</AccessMethod>
         <DatabaseType>Relational</DatabaseType>
         <DatabaseProductName>Oracle</DatabaseProductName>
         <DatabaseProductVersion>
         9.2.0.1</DatabaseProductVersion>
         <Host>cldb.dkrz.de</Host>
         <Instance>mpim</Instance>
         <Database>cera4</Database>
         <Port>5651</Port>
         <Query>Select Data From
         ECHAM3_T42_22032AMIP_U10M</Query>
         <DataFormat formatsystem="MIME">
         application/wmo-grib</DataFormat>
      </ADOLocator>
   </AtomicDataObject>
</DataCollection>
<!-- -->
<DataCollection dataid="ECHAM3_T42_22032AMIP_W900">
   <DataDescription>
      <DataName>ECHAM3_T42_22032AMIP_W900</DataName>
      <TypeOfData>Collection</TypeOfData>
      <Status>Complete</Status>
      <LogicalDescription>
         <Parameter>
            <ParamName>latitude</ParamName>
            <Derivation>condition</Derivation>
            <Units>
               <UnitName>degrees</UnitName>
            </Units>
            <Range>
               <Limit bound="upper">87.8638</Limit>
               <Limit bound="lower">-87.8638</Limit>
            </Range>
         </Parameter>
         <Parameter>
            <ParamName>longitude</ParamName>
            <Derivation>condition</Derivation>
            <Units>
               <UnitName>degrees</UnitName>
            </Units>
            <Range>
               <Limit bound="upper">357.1875</Limit>
               <Limit bound="lower">0.0</Limit>
            </Range>
         </Parameter>
         <Parameter>
            <ParamName>altitude</ParamName>
            <Derivation>condition</Derivation>
            <Units>
               <UnitName>Hectopascal</UnitName>
            </Units>
            <ParamValue>900.00</ParamValue>
         </Parameter>
         <Parameter>
            <ParamName>Location</ParamName>
            <Derivation>condition</Derivation>
            <ParamValue>World (general)</ParamValue>
         </Parameter>
         <Parameter>
            <ParamName>start date</ParamName>
            <Derivation>condition</Derivation>
            <Units>
               <UnitSystem>ISO8601</UnitSystem>
               <UnitFormat>YYYY-MM-DD</UnitFormat>
            </Units>
            <ParamValue>1911-01-01</ParamValue>
         </Parameter>
         <Parameter>
            <ParamName>end date</ParamName>
            <Derivation>condition</Derivation>
            <Units>
               <UnitSystem>ISO8601</UnitSystem>
               <UnitFormat>YYYY-MM-DD</UnitFormat>
            </Units>
            <ParamValue>1941-01-01</ParamValue>
         </Parameter>
      </LogicalDescription>
   </DataDescription>
   <DataCollectionLocator>
```

```
                     <DataName>ECHAM3_T42_22032AMIP_W900</DataName>
                     <Locator pathtype="database">database</Locator>
                  </DataCollectionLocator>
                  <AtomicDataObject dataid="ECHAM3_T42_22032AMIP_W900-ADO1">
                     <DataDescription>
                        <DataName>ECHAM3_T42_22032AMIP_W900</DataName>
                        <TypeOfData>Named Select</TypeOfData>
                        <Status>Complete</Status>
                     </DataDescription>
                     <ADOLocator xsi:type="SelectNamedADOL">
                        <AccessMethod authenticationtype="GSI">
                        OGSA-DAI</AccessMethod>
                        <DatabaseType>Relational</DatabaseType>
                        <DatabaseProductName>Oracle</DatabaseProductName>
                        <DatabaseProductVersion>
                        9.2.0.1</DatabaseProductVersion>
                        <Host>cldb.dkrz.de</Host>
                        <Instance>mpim</Instance>
                        <Database>cera4</Database>
                        <Port>5651</Port>
                        <Query>Select Data From
                        ECHAM3_T42_22032AMIP_W900</Query>
                        <DataFormat formatsystem="MIME">
                        application/wmo-grib</DataFormat>
                     </ADOLocator>
                  </AtomicDataObject>
               </DataCollection>
            </DataHolding>
         </Investigation>
      </Study>
      <AccessConditions acsystem="On Application">The user must be a registered
      user of the MPIM facility. To register, apply to MPIM, Bundesstra&#223;e
      53, 20146 Hamburg, Germany
      (http://http://www.mpimet.mpg.de/en/web/)</AccessConditions>
   </MetadataRecord>
</CLRCMetadata>
```

APPENDIX D

This is an XML instance document showing metadata for the eCCP1[21] project; This Study is encoded essentially using the XML Schema in Appendix A. Data is stored in an XML database accessed via http with document level ADO organisation.

```
<?xml version="1.0" encoding="utf-8"?>
<!-- by Philip Couch (CCLRC) -->
<CLRCMetadata>
   <MetadataRecord MetadataID="MID00001" Facility="CCLRC">
      <Topic>
         <Keywords>
            <Discipline>quantum chemistry</Discipline>
            <KeywordSource>http://www.eccp.org/keywords</KeywordSource>
            <Keyword>organic</Keyword>
            <Keyword>alcohol</Keyword>
            <Keyword>structure</Keyword>
         </Keywords>
         <Subjects>
            <Discipline>quantum chemistry</Discipline>
            <SubjectSource>http://www.eccp.org/subjects</SubjectSource>
            <Subject>
               <SubjectName>quantum chemistry</SubjectName>
            </Subject>
         </Subjects>
      </Topic>
      <Study StudyID="SID00001">
         <StudyName>Calculation of the lowest energy structures of organic
         alcohols</StudyName>
         <StudyInstitution>
            <Name institutionID="ABCD" institutiontype="government" />
            <Role>Research</Role>
         </StudyInstitution>
         <StudyPerson>
            <Name>
               <Surname>Chemist</Surname>
               <MiddleInitials>A</MiddleInitials>
               <Forename>John</Forename>
               <Title>Dr</Title>
            </Name>
            <ContactDetails>
               <Address>
                  <Addressline1>Room no.</Addressline1>
                  <Addressline2>Department</Addressline2>
                  <Addressline3>Institution</Addressline3>
                  <Addressline4>Suberb</Addressline4>
                  <Town>Town</Town>
                  <Region>County</Region>
                  <Postcode>AB1 2CD</Postcode>
                  <Country countryabbrev="UK" />
               </Address>
               <DirectLine>+44(0)1234567890</DirectLine>
               <Switchboard>+44(0)1234567000</Switchboard>
               <Fax>+44(0)1234567891</Fax>
               <Email>chemistja@institution.org</Email>
               <WebPage>http://www.institution.org/~chemistja</WebPage>
            </ContactDetails>
            <RoleInStudy>Principal Investigator</RoleInStudy>
            <RoleInInstitution>Professor</RoleInInstitution>
         </StudyPerson>
         <StudyInformation>
            <Funding>EPSRC</Funding>
            <TimePeriod>
               <StartDate>
                  <Date>2003-11-01</Date>
                  <Time>14:20:00-05:00</Time>
               </StartDate>
               <EndDate>
                  <Date>2005-11-01</Date>
                  <Time>14:20:00-05:00</Time>
               </EndDate>
            </TimePeriod>
            <Purpose>
               <Abstract>Calculate the lowest energy structures of a series of
               aromatic alcohols</Abstract>
            </Purpose>
            <StudyStatus>in progress</StudyStatus>
         </StudyInformation>
         <Investigation InvestigationID="INV-eCCP001">
            <Name>structure of phenol</Name>
            <InvestigationType>simulation</InvestigationType>
            <Abstract>To calculate the lowest energy structure of
            phenol</Abstract>
            <DataHolding InvestigationID="INV-eCCP001">
               <DataDescription>
                  <DataName>calculation of the lowest enery structure of
                  phenol</DataName>
                  <TypeOfData>File</TypeOfData>
                  <Status>complete</Status>
                  <LogicalDescription>
                     <TimePeriod>
                        <StartDate>
                           <Date>2003-11-01</Date>
                           <Time>14:20:00-05:00</Time>
                        </StartDate>
                        <EndDate>
                           <Date>2003-11-01</Date>
                           <Time>15:20:00-05:00</Time>
                        </EndDate>
                     </TimePeriod>
                  </LogicalDescription>
               </DataDescription>
               <DataHoldingLocator>
                  <DataName>HID00001</DataName>
                  <Locator pathtype="absolute">
                  http://localhost:8080/exist/servlet/db</Locator>
               </DataHoldingLocator>
               <DataCollection dataid="CID00001">
                  <DataCollectionLocator>
                     <DataName>eCCP data repository</DataName>
                     <Locator pathtype="relative">/eccp/data</Locator>
                  </DataCollectionLocator>
                  <AtomicDataObject dataid="AID00001">
                     <DataDescription>
                        <DataName>structure of phenol</DataName>
                        <TypeOfData>File</TypeOfData>
                        <Status>complete</Status>
                        <DataTopic>
                           <Keywords>
                              <Discipline>quantum chemistry</Discipline>
                              <KeywordSource>
                              http://www.eccp.org/keywords</KeywordSource>
                              <Keyword>structure</Keyword>
                              <Keyword>phenol</Keyword>
                              <Keyword>organic</Keyword>
                              <Keyword>alcohol</Keyword>
                           </Keywords>
                           <Subjects>
                              <Discipline>quantum chemistry</Discipline>
                              <SubjectSource>
                              http://www.eccp.org/subjects</SubjectSource>
                              <Subject>
                                 <SubjectName>quantum chemistry</SubjectName>
                              </Subject>
                           </Subjects>
                        </DataTopic>
                        <LogicalDescription>
                           <Parameter>
                              <ParamName>atomic basis set</ParamName>
                              <Derivation>condition</Derivation>
                              <ParamValue>6-31G*</ParamValue>
                           </Parameter>
                           <Parameter>
                              <ParamName>method</ParamName>
                              <Derivation>condition</Derivation>
                              <ParamValue>DFT</ParamValue>
                           </Parameter>
                           <Parameter>
                              <ParamName>functional</ParamName>
                              <Derivation>condition</Derivation>
                              <ParamValue>b3lyp</ParamValue>
                           </Parameter>
                           <TimePeriod>
                              <StartDate>
                                 <Date>2003-11-02</Date>
                                 <Time>15:00</Time>
                              </StartDate>
                              <EndDate>
                                 <Date>2003-11-02</Date>
                                 <Time>16:00</Time>
                              </EndDate>
                           </TimePeriod>
                           <FacilityUsed>
                              <FacilityName>UCL computing
                              facility</FacilityName>
                              <Resource>UCL condor pool</Resource>
                           </FacilityUsed>
                        </LogicalDescription>
                        <Software>
                           <Production>
                              <LongName>molpro</LongName>
                              <ProgramName>molpro</ProgramName>
                              <Version>2002.6</Version>
                              <URI>http://www.molpro.net</URI>
                              <OperatingSystem>Redhat Linux</OperatingSystem>
                              <OperatingSystemVersion>
                              9.0</OperatingSystemVersion>
                              <Architecture>Intel</Architecture>
                           </Production>
                        </Software>
                     </DataDescription>
                     <ADOLocator>
                        <Locator pathtype="relative">phenol.xml</Locator>
```

```
                        <AccessMethod authenticationtype="none">
                        http</AccessMethod>
                     </ADOLocator>
                  </AtomicDataObject>
               </DataCollection>
            </DataHolding>
         </Investigation>
      </Study>
   </MetadataRecord>
</CLRCMetadata>
```

REFERANCES

[1] http://www.pik-potsdam.de/dept/dc/e/sdm/cera/, The CERA Central Page

[2] http://www.itd.clrc.ac.uk/Person/K.G.Jeffery, Professor Keith Jeffery

[3] http://www.rcuk.ac.uk/escience/, UK e-Science Programme

[4] http://dublincore.org/documents/dces/, Dublin Core Metadata Element Set Reference Description

[5] http://www.purl.org/, Persistent Uniform Resource Locator

[6] http://www.w3.org/TR/NOTE-datetime, Date and Time Formats, W3C profile of ISO 8601

[7] http://www.isis.rl.ac.uk/, ISIS Pulsed Neutron & Muon Source

[8] http://lhc-new-homepage.web.cern.ch/lhc-new-homepage/, The Large Hadron Collider

[9] http://www.cryst.bbk.ac.uk/pdb/pdb.html, Bikbecks Protein Data Bank Mirror

[10] http://www.npaci.edu/DICE/SRB/, Storage Resource Broker

[11] http://www.globus.org/rls/, Globus Replica Location Service

[12] http://www.iana.org/assignments/media-types/ , MIME Media Types

[13] http://www.e-science.clrc.ac.uk/documents/projects/dataportal/security.pdf, Grid Authorisation Framework for CCLRC Data Portal, Ananta Manandhar et al.

[14] http://www.mpimet.mpg.de/en/web/, Max Planck Institute for Meteorology

[15] http://eminerals.org/, 'Environment from the Molecular Level' A NERC eScience testbed project

[16] http://www.e-science.clrc.ac.uk/web/projects/complexmaterials, Simulation of complex materials

[17] http://www.mygrid.org.uk/, MyGrid Project

[18] http://www.dcc.ac.uk/, Digital Curation Centre

[19] http://www.w3.org/XML/Schema, XML Schema

[20] http://www.e-science.cclrc.ac.uk/web/projects/dataportal, CCLRC DataPortal Project

[21] http://tyne.dl.ac.uk/twiki/bin/view/ECCP/WebHome, ECCP Web home