An Interdisciplinary model for the representation of Scientific Studies and associated data holdings

Shoaib Sufi (CCLRC - Daresbury Laboratory)

Brian Mathews (CCLRC - Rutherford Appleton Laboratory)

Kerstin Kleese van Dam (CCLRC - Daresbury Laboratory)

A Common model for the representation of scientific study metadata does not exist, by proposing a model and an implementation, the adoption of such a system would aid interoperability of scientific information systems on the Grid, or at the very least the model will form a specification of the type and categories of metadata that studies should capture about their investigations and the data they produce. This will allows further exploitation of the Study, associated datasets, ease citation, facilitate collaboration and allow the easy integration of pre-Grid metadata into a common Grid based information platform.

1. CCLRC SCIENTIFIC METADATA MODEL

The CCLRC scientific metadata model (CSMD) is study-dataset orientated model and comprises of information pertaining to provenance, conditions of use, data description and location and related material, and includes indexing information. The main influences for developing the model were in-house facilities at CCLRC; specifically ISIS (Neutron Spallation at Rutherford Appleton Laboratory), SR Synchrotron Radiation source (at Daresbury Laboratory) and the British Atmospheric Database (BADC) at RAL.

The specific metadata formats which have influenced the design and ordering of the CSMD are CIP from Earth observation [1], DDI from social sciences [2], publication type metadata from the Dublin Core [3] and lower level 'Scientific Data Objects' metadata found in XSIL [4] as well as CERA [5] from the MPIM in Hamburg. The Dublin core was found to be too high level and not detailed enough whereas XSIL lower level and missed higher level entities, CERA was a close fit but was somewhat specific to the Earth Sciences and as a key feature of our meta data model was generality CSMD was developed.

2. PURPOSE OF A MODEL

The Model Specifies in a semi-structured way the types of metadata that need to be captured which will make studies easier to exploit, cite, groups to collaborate and allow a lowest common denominator for scientific study information integration within a Grid environment.

2.1 Implementation issues

One of the driving force applications has been the CCLRC e-Science DataPortal [6] technology in which an XML implementation of the model is the main data transport. Using the model in this way has acted as 'stress-test' of the model as well as the implementation; limitations have been identified and new requirements discovered which have lead to changes in the model and thus the implementation.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE CLRCMetadata SYSTEM "clrmetadata.dtd">
<CLRCMetadata><MetadataRecord metadataID="N000001">
        <Topic>
                <Discipline>Chemistry</Discipline>
                <Subject>Crystal Structure</Subject>
                <Subject>Copper</Subject>...
        <Experiment>
                <StudyName>Crystal Structure: Copper : Palladium:  :complex:
150K ...
                <Investigator><Name><Surname>Porter...<Institution>University
of  Peebles ...
                <Funding>EPSRC ...
                <TimePeriod><StartDate><Date>2 /04/1999....
                <Purpose><Abstract>
To study the structure of Copper and Palladium co-ordination complexes at n
150K.
                <DataManager><Name><Surname>Teat...
                <Instrument>SRS Station 9.8  BRUKER AXS SMART 1K...


<Condition>...Wavelength...<Units>Angstrom...<ParamValue>0.6890...
                <Condition>...Crystal-to-detector
distance<Units>cm...<ParamValue>5.00...
        <AccessConditions>The user has to be one of: Prof. E. Porter
```

Figure 1 - Example of XML representation of metadata model

There is also a relational implementation of the CSMD being used in the e-minerals projects [7] and also on the e-materials project [8].

3. MODEL BREAKDOWN

The following section gives a break down of the metadata stored in the CCLRC Scientific Metadata format. The cardinality of the pieces of metadata stored and issues relating to the allowable values of that data are discussed later.

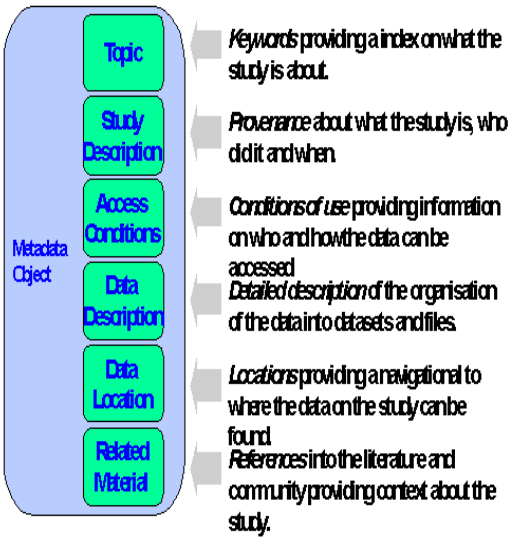

Figure 2 - Schematic of metadata model

3.1 The Study

The Study holds provenance (i.e. who did the study and where was it done etc.) information.

3.1.1 Study Name

This holds the complete Study name.

3.1.2 Study Institution

Institutions involved in the Study and their respective roles.

3.1.3 Investigator

The name, institutions, contacts details and roles of the individuals in the Study.

3.1.4 Study Information

Extended study information such as a Study abstract, funding,

start and end times of the study, the study status (e.g. in progress, complete), resources used by the study (e.g. facilities)

3.1.5 Investigation

Investigations carried out under this Study with further metadata for each investigation.

3.1.5.1 Name

The Investigation Name.

3.1.5.2 Investigation Type

An instance of the enumeration of the allowed types e.g. was the investigation an experiment, simulation, measurement, calculation etc.

3.1.5.3 Abstract

A short description of the investigation and why it was performed.

3.1.5.4 Resources

A description of the resources used in this investigation.

3.1.5.5 Data Holding

Holds a link to the Data Description and Data Location for this investigation.

3.1.6 Notes

Miscellaneous Study notes (could be reviewer's remarks for example).

3.2 Indexing by Topic

By topics we mean subjects and keywords. This is the main indexing method of the Study metadata and what can be searched on (in structured and user constructed unstructured free-word searches).

3.2.1 Keywords

3.2.1.1 Discipline

The area of science from which the keyword is referring (e.g. field in maths different from field in biology).

3.2.1.2 Keyword Source

A link to a domain dictionary from which this terms definition is stated.

3.2.1.3 Keyword

The actual keyword itself.

3.2.2 Subjects

3.2.2.1 Discipline

The scientific domain that the subject is referring to.

3.2.2.2 Subject Source

A link to a dictionary, restricted vocabulary, ontology etc, from which this terms definition is stated.

3.2.2.3 Subject

A hierarchical classification list of terms ending with one subject in the correct context:

(e.g. earth sciences/atmosphere/temperature/ air temperature)

3.3 Access Conditions

Either a list of users or groups who are allowed access to the metadata and data, or a pointer to an access control system which contains such data for this study.

3.4 Data Description

3.4.1 Data Name

The logical name of the data object (e.g. database BLOB, file-system file) or data collection.

3.4.2 Type of Data

The data format of the data object or general format information for the data collection.

3.4.3 Status

Whether the data is complete or in being added to.

3.4.4 Data Topic

Allows the attachment of Keyword and Subject indexing down to the data object level.

3.4.5 Logical Description

3.4.5.1 Parameter

A list of tuples containing information about parameters used to collect the data collection/data object. This includes information about how the parameter was derived (e.g. possible enumerations being measured, fixed). The units of the value and possible range is also stated. Ranges and margins of error can also be stated.

3.4.5.2 Timer Period

The creation and completion date-time of the data collection or data object.

3.4.5.3 Description

Textual description of the collection/object.

3.4.5.4 Facility Used

Facilities used to generate the collection/object data.

3.5 Data Location

3.5.1 Data Name

This is the logical name of the data as in the data description.

3.5.2 Locator

This holds specific information about the data object and metadata concerning it's location and retrieval method on a file system, web server, database system etc (it is usually a URL).

3.6 Related Material

One or many links and or textual descriptions of material related to this study e.g. earlier studies or parallel studies.

4. MODEL CONSTRAINTS

4.1 Mandatory and Optional

The mandatory or optional nature of a piece of metadata could be viewed as an 'implementation issue' however if no data is captured one could say this is still conformant as everything is optional – this is an obviously unacceptable situation.

Thus a base level of information that must be included for the metadata to be conformant is specified in the model including the number of occurrences i.e. cardinality.

4.2 Enumeration Issues

In a sense enumeration e.g. types of institution, roles etc are distinct from the model in the same way that the classifications system and controlled vocabularies used in the keywords and subjects specified in the Topic metadata are. Thus they are necessary but need a different source to specify them. Implementation e.g. the XML one used in the DataPortal project does specify institution roles, people roles, institution types and investigation types but these are not necessarily part of the model.

4.3 Cardinality

In some cases there should only be 1 instance of a particular item with its value In others 1 to many and in other 0 to many. These issues are sometimes the source of fierce debate and are best left to the implementation the model documentation gives an indication of what it thinks is a possible resolution to this issue.

e.g. should a study one have one full name or could it have more than one name ?; the model suggests one name but there is nothing stopping an implementation from having more than one name.

5. CONFORMANCE LEVEL

There is a lot of information to be stored per study/data holding for the metadata record to be complete. Additionally indexing issues and the level of indexing is an issue. Thus conformance levels are needed with higher numbers representing a more complete metadata record. Each level would indeed lead to an increase in processing needed to generate and maintain the metadata conversely each level would increase the metadata mining possibilities.

5.1 Conformance and Integration

There are different levels of conformance one can have to the model; if all the items specified in the model are captured but stored in a different way such that they could be mapped to the model then we can say that the metadata is conformant (to some degree) to the model even if they did not know this; e.g. the CERA metadata model is conformant to a certain degree. In practise an wrapper architecture [9] is used to convert from one format (e.g. the format of the data archive) to another (e.g. the CSMD format) in

e.g. the DataPortal Project ; but it is possible to do this because all the data is their in the source archive albeit in a different form from what the implementation of the CSMD expects.

At a workshop in October 2002 at NIEES on Metadata [10]. A discussion on the various formats was ensuing – the simple question was asked – are all NERC projects capturing the type of metadata which will make them useful in the future regards less of format – the simple answer 'no' was given. Thus the model could at the very least form a basis for the specification of the types of metadata that should be captured by scientific studies.

5.2 Conformance Levels

5.2.1 Level 1

In Level 1 information about the study and investigations is complete but there is no mention of data holding i.e. the data collections and data sets metadata is missing. Indexing is provided only at study level. This could also be considered 'library' level metadata.

5.2.2 Level 2

Level 2 consists of information about studies and data holding and indexing is provided only at study level.

5.2.3 Level 3

In Level 3 information about study, data holdings, related material and access conditions is available and indexing is done to data collection level.

5.2.4 Level 4

Level 4 is as Level 3 but with the additional constraint that the granularity of the indexing reaches the data object level and the data objects have full parameter level information

5.2.5 Level 5

Level 5 contains all information about each section i.e. full study, access conditions, related material, data description, logical description, parameter information and full indexing as well as details on facilities used and funding. It is not envisioned that existing study systems will hold this level of metadata and that only new system developed with the concepts outlines previously maybe able to reach this level.

5.3 Benefits of a conformance level

If a stated level of conformance is met this would allows for better clients to use systems based on the CSMD format. Richer data mining and presentation options would be inherit benefits to conforming to the model to a higher level.

6. MORE INFORMATION

The latest description of the CSMD model can be viewed at:

http://www-dienst.rl.ac.uk/library/2002/tr/dltr-2002001.pdf

The latest Relational implementation and XML Schema of the model with enumerations is available on request by e-mailing:

dataportal@dl.ac.uk with the subject containing [metadata model request]

For further information about the DataPortal project which is using the CSMD as it's data format:

http://www.e-science.cclrc.ac.uk/web/projects/dataportal

7. FURTHER WORK

It is hoped that the following items will be worked on:

1) Using standard enumerations for such things as institution roles, people roles, institution types, investigation types, parameter types etc.

2) Checking other emerging metadata standards for scientific information and incorporating new ideas from these models into the CSMD.

3) Updating the XML and Relational implementation of the model so they more closely track the model; perhaps offer examples also.

4) Language issues – should different languages be subject to direct translation only or is other support needed.

5) Internationalisation – i.e. will the different norms of scientific investigation affect the model and how will the terminology used in North America and Europe be harmonised.

8. REFERENCES

[1] CIP Metadata from Earth Observation. http://lcweb.loc.gov/z3950/agency/profiles/cip.html

[2] DDI Metadata from Social Sciences. http://www.icpsr.umich.edu/DDI

[3] Dublin Core Metadata Initiative. http://dublincore.org/

[4] Extensible Scientific Interchange Language. http://www.cacr.caltech.edu/SDA/xsil/

[5] Meta information on Geo-referenced Data. http://www.pik-potsdam.de/dept/dc/e/sdm/cera/

[6] Glen Drinkwater, Kertsin Kleese, Shoaib Sufi, Lisa Blanshard, Ananta Manandhar, Rik Tyer, Kevin O'Neill, Michael Doherty, Mark Williams, Andrew Woolf. The CCLRC DataPortal. http://www.e-science.clrc.ac.uk/web/projects/dataportal

[7] Environment from the molecular level (e-minerals) http://www.eminerals.org/ http://www.e-science.clrc.ac.uk/web/projects/eminerals

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

[9] Shoaib Sufi. XML Wrapper Architecture. http://www.e-science.clrc.ac.uk/documents/projects/dataportal/xmlwrapper.pps

[10] Workshop on data and metadata standards, 10–11 September 2002. http://www.niees.ac.uk/metadata_programme.html