Enhancing the Core Scientific Metadata Model to Incorporate Derived Data

Erica Yang[1,]

Bodleian Libraries, University of Oxford, Oxford OX2 0EW, U. K.

Brian Matthews, Michael Wilson[1,]

STFC e-Science, Rutherford Appleton Laboratory, HSIC, Didcot, OX11 0QX, U.K.

Abstract

Much of the value in scientific data is provided not only in the raw data but through the analysis of that data to derive published results. A study of the data analysis process for structural science has shown that various data sets derived from the raw data are of use to scientists and should be stored with the raw data. The Core Scientific MetaData model (CSMD) is used by a number of large scientific facilities to catalogue scientific data. The current version provides support to experimental scientists to access their raw data, facility managers for accounting for facility usage and other scientists who wish to re-use raw experimental data. In this paper, extensions to the CSMD are presented to describe the analysis process so that the provenance of the derived data can be captured. A pilot implementation incorporating derived data through this extended CSMD model has been trialled with experimental scientists. Remaining challenges to the adoption of CSMD and tools it supports are considered.

Keywords: data management, information management, derived data management, data analysis, data provenance, large scale facilities, neutron sources, scientific process.

☆

Email addresses: erica.yang@bodleian.ox.ac.uk (Erica Yang), Brian.Matthews, Michael.Wilson@stfc.ac.uk (Brian Matthews, Michael Wilson)

1. Introduction

Increasing quantities of the raw experimental data generated using large scientific facilities, such as large-scale photon and neutron sources, are being made available in a systematic and secure way. This data is intended for three main users: the experimental scientists who undertook the study need access to the raw data from their universities in order to analyse it further; the facilities managers who need access to data to manage the use of their facilities; and other scientists who may be able to access the data for re-analysis, either to verify the published results, or to derive new scientific results without the cost of repeating the original experiment, possibly in combination with results from elsewhere.

The Core Scientific MetaData model (CSMD) [17, 10] has been designed to capture information about experiments and the data they produce in what are broadly known as the "structural sciences", such as chemistry or earth science, which consider the molecular structure of matter. It is used by the data cataloguing system ICAT [5] which is used by the ISIS neutron source[1] and the Diamond Light Source (DLS)[2], both operated at the Harwell Science and Innovation Campus in the UK. The DLS synchrotron generates brilliant beams of light, from infra-red to X-rays, which are used in a wide range of applications, from structural biology through fundamental physics and chemistry to cultural heritage. The ISIS source generates beams of neutrons and muons used to investigate the properties of materials at the scale of atoms for research into subjects ranging from clean energy and the environment, pharmaceuticals and health care, through to nanotechnology, materials engineering and IT. The two target stations of the ISIS neutron source host 30 beamlines with their associated instruments, while DLS currently hosts 13 instruments on separate beamlines. The use of these facilities is not limited to a small coterie of specialists, but between them these instruments are used by many thousands of experimental scientists each year from around the world. As similar large facilities are developed in other countries the data sets they create are becoming more common, and it becomes more urgent to capture that data, and to ensure that all stages of its analysis are accurately recorded. Consequently, facilities such as the Institut Laue-Langevin (ILL)[3]

[1] http://www.isis.stfc.ac.uk
[2] http://www.diamond.ac.uk
[3] http://www.ill.eu

are also adopting the ICAT infrastructure, and the PANDATA initiative[4] is developing best practice in data management across facilities internationally.

Data cataloguing systems support access to scientific data, but the present ICAT only catalogues the raw data produced by the facility, while derived data is managed locally by the scientist carrying out the analysis at the facility or in their home institution. This is on an ad hoc basis, and these intermediary derived data sets are not archived for other purposes. Thus the support for the intended users is partial.

In order to improve the support offered by the facilities data management tool such as ICAT, its underlying data model, CSMD needs to be extended. Currently, it does not support access to the derived data produced during analysis, nor does it allow the provenance of data supporting the final publication to be traced through the stages of analysis to the raw data.

Bioscientists have used workflow tools to capture and automate the flow of analyses and the production of derived data for many years [12] and can now automatically run many computational workflows [20]. In other structural sciences, such as chemistry and Earth sciences, the management of derived data is less mature, workflows are not standardised and can less readily be automatically enacted. Rather the data needs to be captured as the analysis proceeds so that scientists do not lose track of what has been done. A data management solution is required to capture the data trails that are generated during analysis, with the aim of making the methodologies used by one group of researchers available to others.

Further, the accurate recording of the process so that results can be replicated is essential to the scientific method. However, when data are collected from large facilities, the expense of operating the facility means that the raw data collection effectively cannot be repeated. Therefore tests to replicate results has to come from re-anlysis of raw data as much as repetition of the data capture in experiments.

In order to provide support for the analysis undertaken by the experimental scientists; to permit the tracing of the provenance of published data; and to allow access to derived data for secondary analysis, it is necessary to extend the CSMD to account for derived data and to record the analysis process sufficiently for the needs of each of these use cases. In terms of data

[4]PANDATA Photon and Neutron Data Infrastructure. http://www.pan-data.eu/Main_Page

provenance [8], the current CSMD approach identifies the source provenance of the resultant data product, but it needs to be extended to describe the transformation provenance as well.

In this paper, after a summary of the existing CSMD, an example scientific process will be described to motivate the extensions to the CSMD. Section 4 will then detail extensions to the CSMD to meet these requirements, before a pilot implementation of the extended CSMD is described using the ICAT data catalogue system. Finally the limitations of the proposed extensions, practical limitations on the adoption of the data catalogue system and future work will be considered.

2. Core Scientific MetaData model

The Core Scientific MetaData model (CSMD) [17] is an extensible model of metadata originally designed to capture a common set of information about the data produced by experiments, measurements, and simulations in facilities science. The model is the result of an analysis of science practice over a number of years and a range of projects, and has proved a robust system.

CSMD was developed primarily to allow facility operators, such as STFC, to introduce a systematic approach to manage their data assets across the heterogeneous scientific facilities. Although operators may produce data files of different formats and content resulting from different equipment, experiments, or disciplines, there are commonalities features of the context of the data that can be captured. They include:

1. the description of the data production process (e.g. where/when/by who/how);
2. the format, type, owner, and identifier of the data;
3. the parameters in which the data should be interpreted;
4. the relationships between data.

Having a standardised metadata model underpinning the data management infrastructure that an operator uses, supports a common strategy towards maintaining, searching, and discovering data assets, reducing the overall operating cost. This is important to both facility providers who host a wide range of scientific facilities and to users who utilize multiple facilities. Metadata are also crucial for scientists other than the ones who design the equipment or run the experiment, to interpret, understand and make use of the data.

The model as it currently stands aims to describe the physical raw data files (binary, images, or text containing numeric values) produced by the data acquisition software of a detector within an instrument. These files have formats which depends on the equipment, the facility, or the program that the data is produced from. The Network Common Data Format (netCDF) [14] and Hierarchical Data Format (HDF) [6] are well defined formats used by many laboratories, while NeXus [9], derived from HDF5, is a common data format targetted at neutron, X-ray, and muon sciences which several facilities have adopted to different degrees: not all the data files produced within these communities use this format since many instruments still produce older non-standard formats.

In CSMD data files are grouped into *datasets*, where a dataset is an abstract notion referring to a set of related data files. How the files are related is determined by the context. For example, if an experiment produces 10 files in a run, which is repeated 100 times in different temperatures, 100 datasets can be created, each with the 10 files produced under a specific temperature. This dataset concept is essential for experiments that produce a large number of files in each run.

Datasets are then grouped into *investigations*, where an investigation - which can be an experiment, a set of measurements, or a simulation - is defined as a data generation activity. For example an investigation may represent a particular allocation of time on an instrument to a scientist for the analysis of a sample of a material, which may generate a number of data sets each collected at a different experimental parameter setting. Like the dataset, an investigation is not a concept referring to an object of physical presence, but rather an abstract notion referring to a set of related datasets generated from the same data generation activity.

Investigations are further grouped into *studies*, where a study is also an abstract notion referring to a set of related investigations, in other words, a set of related data generation activities. For example, two investigations, an experiment on a sample and a related computer simulation of the experiment, could be grouped together to form a study of the sample.

The CSMD has been implemented and deployed in STFC to support scientific data cataloguing and management for its major international facilities. The current production implementation of CSMD, ICAT 3.3[5], is based

[5]http://code.google.com/p/icatproject/

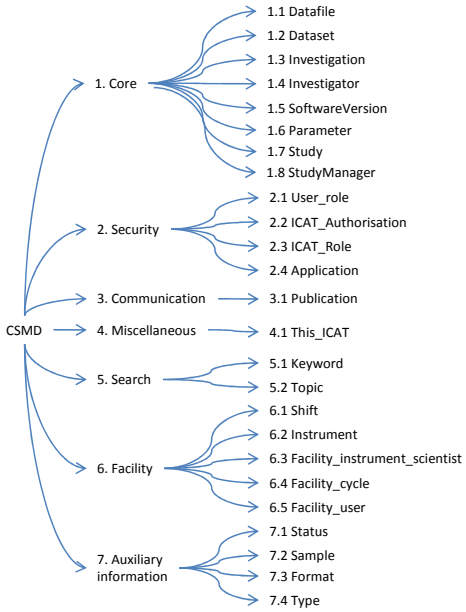

Figure 1: A classification of the concepts in CSMD

on the CCLRC Scientific Metadata Model v2 [17] with extensions. This model forms the core of the ICAT infrastructure to catalogue, manage and distribute data for facilities users.

Although CSMD was originally intended to accommodate data collection and processing a much wider context of scientific studies from raw data collection to downstream data analysis, it is currently only *being used* to support raw data cataloging. In order to focus on the key data management issues throughout the data production pipeline and to clarify the extensions needed for derived data, we identify the core and optional concepts in the model. The concepts in CSMD can be classified into seven categories (see Figure 1):

Core. The concepts which are central to scientific data management. Capturing the data outputs involve four data objects: datafile, dataset, investigation, and study. A datafile corresponds to a *physical data object* that is stored on physical storage disks, while datasets, investigations, and studies are *abstract data objects* that encapsulate other (physical or abstract) data objects as described above. Other core concepts include the Investigator and StudyManager, representing people associated with an investigation

and a study, respectively. The Software concept[6] captures an activity that produces or consumes data objects, while the Parameter concept captures some value which provides context to the data production process, such as environmental characteristics, instrument settings, or measured quantities.

Search. Classifiers which can be assigned to data to facilitate the search and discovery of core concepts.

Communication. Concepts which link between data and other research outputs so that the provenance of a research publication can be traced back to the data holdings.

Security. Concepts which are used to enforce access control policies on the data holdings. These may vary according to the security context of the facility.

Miscellaneous. Meta-entities which identify the specific instance of ICAT metadata catalogue.

Facility. Concepts related to facilities are introduced to capture the contextual information associated with the (raw) data collection process, such as which facility and instrument was used, which cycle, shift or run of the facility, the instrument-scientist (a specialised role in a large-scale facility) was involved, additional safety information. These concepts are specialised to facility usage, although there are analogues in other experimental contexts, such as university laboratory experiments.

Auxiliary Information. Specific information associated with data holdings. It is currently being used to store information related to *raw* data files, such as the sample under analysis, further parameters (e.g. temperature, humidity), and file format. But it should be possible to extend or adapt these concepts to store any information related to data holdings produced along data analysis pipelines.

Two types of information are left out from Figure 1: links between the concepts within a category; and those between the concepts across categories. We address the former in the rest of this paper. The latter does not directly relate to the paper, and we shall not expand on that further.

[6]In CSMD 2.0 and ICAT 3.3, the concept Software is referred as SoftwareVersion.

3. Derived Data in the Analysis Process

In this section we study in detail an example data analysis pipeline from the raw data gathered at a facility to the final scientific findings suitable for publication.

Along the pipeline, three concepts, raw, derived, and resultant data, are often used to differentiate the roles of data in different stages of the analysis and to capture the temporal nature of the processes involved. *Raw data* are the data acquired directly from the instrument hosted by an facility, in the format support by the detector. *Derived data* are the result of processing (raw or derived) data by one or more computer programs. *Resultant data* are the final findings of an analysis, for example, the structure and dynamics of a new material being studied in an experiment.

3.1. Background

We initially performed a desk study of three experiments involving two different types of facilities: neutron and synchrotron facilities, in the UK. One experiment is in the domain of Chemistry using the Diamond synchrotron and the UK National Crystallography Service (NCS) [4] to determine the structure of atoms in solids using X-ray diffraction. The other two experiments aim to determine the structure of atoms of matters (e.g. liquids or solids) using neutron techniques: one uses the neutron diffraction[7] provided by the GEM instrument[8] and the other small angle neutron scattering[9] offered by the Sandals instrument[10]. Both instruments are located at the ISIS neutron spallation source.

The NCS analysis workflow is the most prescriptive among the three experiments because the processes involved are standard and the data formats used are well established [4]. The analysis workflows for the other two experiments are more complicated but the nature of the analysis is similar and both workflows involve

- computationally intensive programs, and
- intensive human oriented activities that demand significant experience and knowledge to direct the programs.

[7] `http://www.isis.stfc.ac.uk/instruments/neutron-diffraction2593.html`
[8] `http://www.isis.stfc.ac.uk/instruments/gem/gem2467.html`
[9] `http://www.isis.stfc.ac.uk/instruments/small-angle-scattering2573.html`
[10] `http://www.isis.stfc.ac.uk/instruments/sandals/sandals6929.html`

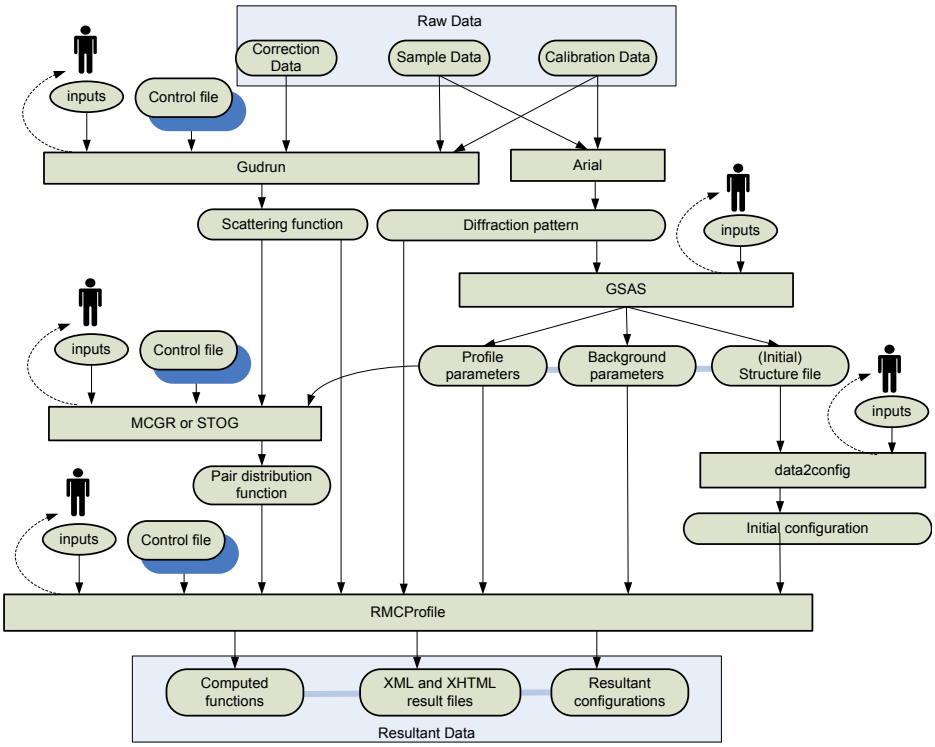

Figure 2: The RMC data analysis flow diagram

In practice, it can take months from the point that a scientist collects the raw data to the point where the resultant data are obtained. Both workflows overlap in their data correction process as they use the same set of programs to correct the raw data obtained from the instruments (e.g. to identify the data resulting from malfunctioning detectors), though this represents only a small part of the respective workflow.

Given these similarities we shall focus on the details of the data analysis flow of the neutron scattering experiment using the GEM instrument to study derived data problem, although hierarchical task anlaysis [16] has been applied to all the studies and the abstractions do generalise across instruments, techniques, programmes and disciplines.

3.2. Data Analysis

Data analysis is the crucial step transforming raw data into research findings. In a neutron experiment, the objective of the analysis is to determine

the structure or dynamics of materials under controlled conditions of temperature and pressure. Figure 2 illustrates a typical flow for analysing raw data generated from the GEM instrument using Reverse Monte Carlo (RMC) based modelling [19]. The RMC method is probabilistic, which means that a) it can only deliver an approximated answer and b) in theory, there is always scope to improve the results obtained earlier using the same method.

In the figure, rectangles represent the programs used for the analysis; rounded rectangles without shadow represent the data files generated by computer programs; rounded rectangles with shadow represent data files handwritten by scientists as inputs to the programs; ovals represent human inputs from scientists to drive the programs; solid lined arrows represent the information flow from files to programs, from programs to files, or from human to programs; and the dashed lined arrows are included to highlight the human oriented nature of these programs demanding significant expertise. This is an iterative process that takes considerable human effort.

3.2.1. Data reduction

Three types of raw data are input into the data analysis pipeline: sample, correction, and calibration data. They are first subject to a data reduction process which is facilitated by two programs: `Gudrun`, a Fortran program with a Java GUI, and `Arial`, a IDL program. The outputs from `Gudrun`[11] are a set of scattering functions, one for each bank of detectors. For `Arial`[12], the outputs are a set of diffraction patterns, again, one per bank of detectors.

With Gudrun, the human has to subtract any noise in the data going from scattering function to pair distribution function (through the `MCGR` or `STOG` program). Noise can arise from several sources, e.g. errors in the program, or noise due to the statistics on the data. In other words, when the other programs use the derived data generated by `Gudrun`, human expertise is required to steer the way the data is used.

3.2.2. Initial structural model generation

The next step is the process of generating the initial configuration of the structure model that will be used as the input to the rest of the RMC workflow. This step requires three programs (i.e. `GSAS`, `MCGR` or `STOG`, and

[11] `http://www.isis.rl.ac.uk/disordered/Manuals/gudrun/gudrun_GEM.htm`

[12] `http://www.isis.stfc.ac.uk/instruments/osiris/data-analysis/ariel-manual9033.pdf`

`data2config`) to transform the reduced data into structure models that best fit the experimental data. To do this requires determining the structural parameters (e.g. atom positions), illustrated as the sets of data files under `GSAS`, for all the crystalline phases present, which are: profile parameters, background parameters, and (initial) structure file.

Most neutron and synchrotron experiments use the Rietveld regression analysis method to refine crystal structures. Rietveld analysis, implemented in `GSAS`, is performed to determine the structural parameters as well as to fit the crystal structure to the diffraction patterns using regression methods. Like all regression methods, it needs to be steered to prevent it following a byway. Some values in the pair distribution functions produced from `MCGR` or `STOG` are compared with their counterparts in the scattering functions to ensure that they are consistent. If they are not, the scientist repeats the analysis.

The `data2config` program takes the configurations generated from `GSAS`, or from crystal structure databases to determine the configuration size of the initial structure model.

3.2.3. Model fitting

All the derived data generated up to this point represents an initial configuration of the atoms, random or crystalline, which is fed into the `RMCProfile` [18] program implementing the RMC method to refine models of matter that are mostly consistent with experimental data. It is the final step in the analysis process to search for a set of parameters that can best describe experimental data given a defined scope of the search space and computational capacity. This is a compute-intensive activity which is likely to take several days of computer time. It is also a human-oriented activity because human inputs are required to "steer" the refinement of the model.

3.3. Discussion

The scientific process under consideration passes through the main phases of sample preparation, raw data collection, data analysis and result gathering. The overall data analysis process described above passes through the three phases of data reduction, initial structural model generation, and model fitting. This hierarchical structure is common to the different processes analysed. However, as the detailed example above illustrates, within each of these phases there are many different programs involved (with potentially different versions), with varying numbers of input and output objects. Because the

analysis method is probabilistic, there is always scope for further improvements to the results so variations on the analysis can always be undertaken.

Throughout the analysis, many of the intermediate results are useful both for the scientists who perform the original experiment and others in the scientific community. The investigators or others can, for example: use them for reference; revisit them when better resources (more powerful computers, better analysis methods, programs or algorithms) are available; and revise them when better knowledge about the program behaviours are available.

The scientists consulted are thus not only motivated to publish their final results but also the raw and derived data generated along the analysis flow. This is especially true for new analysis methodologies, such as the RMC method described in this paper which is a relatively new method in the neutron scattering community which those who use it wish to have accepted more widely. In this case, scientists are highly motivated to publish the *entire data trail* along the analysis pipeline and publicise the *methodology* that is used to derive the resultant data. Making their data available potentially can lead to: more citations to their published papers and results; awareness and adoption of their methodology; and the discovery of better atomic models built on the models they have derived.

Data archiving is also of interest to the facilities operators because of the potential of derived data reuse by other researchers who would add more value to the initial experimental time. However, apart from the raw data, neither the ICAT infrastructure nor the CSMD model capture derived data whose management is currently left to the experimental scientist. In the next section we will propose extensions to the CSMD model to capture the derived data on the basis of an abstraction of the detailed workflow described here.

4. An Enhanced CSMD Model

This section presents how the CSMD model is extended to describe the data analysis process so that the provenance of derived data can be captured. Several factors are important for capturing data provenance, including:

- the *data objects* involved;
- the *programs* that produce or consume data objects;
- the *ordering* of the programs;

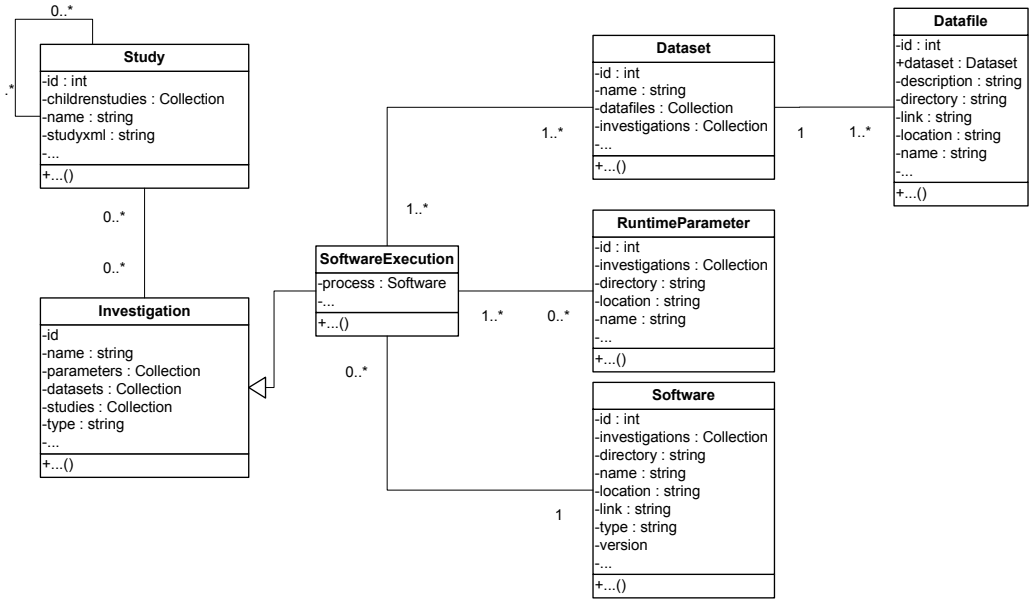

Figure 3: An Extended CSMD Object Model for Supporting Derived Data

- the *parameters* to the programs; and
- the *people* who drive the programs.

In a production system, the people element is as important, if not more important, than the others, for accountability, security, attribution, and archival reasons. However, because this element has been well captured in the current CSMD model and implemented in ICAT version 3.3, we shall not include it in the presentation of the extended model.

Figure 3 is a (concise) object model depicting the extensions and modifications to the core of the existing CSMD model to support derived data[13]. In order to keep it digestible, the connections between the entities presented here and those in the current model[14] are omitted. For example, in the

[13]For a complete set of the attributes and operations of each object, readers are referred to the sourceforge website: `http://icatlite.sourceforge.net/`.

[14]ICAT 3.3 Database schema: `http://icatproject.googlecode.com/svn/icat3_api/trunk/icat3-database/IcatDB/jdeveloper/icat/schemadiagrams/model/icat_v3.png`

current model, the entities - `datafile` and `dataset`, are linked to the entity `parameter` (via `datafileParameter` and `datasetParameter`) to describe the set of instrument parameter settings related to them. These links are not included in this Figure.

An object model is an *abstraction* of the objects involved and the relationship between the objects. In real world systems, the object model is manifested[15] as data models, which can be implemented in all kinds of object-oriented programming languages (e.g. Java, C++ or C#), relational databases (e.g. MySQL, Oracle), RDF, XML Schemas, and even an XML databases (e.g. eXist). A data model can also be implemented in scripting and interpretive programming languages, such as Perl, PHP, or even Javascript, in a non-object oriented fashion, although this is not recommended because it will lose the benefits of object-orientation.

However, this highlights two important points: 1) an object model can be mapped into various data models; and 2) the data captured in one data model can be transformed into another one, as long as they both conform to the same object model. A major benefit of doing so is to enable the interoperation among the implementations built upon the data models. In practical terms, this means that the derived data provenance captured in one data model, for example, implemented as a J2EE persistence data layer, can be pushed to a persistent data repository implemented in another data model, for example, implemented as a relational database. Conversely, the provenance stored in a database can be presented in another data model, for example, in a XML schema. The transformation between the data models can be facilitated by, for example, the Java programming language.

In the next section, we shall present our implementation of such an example, showcasing how these mappings can be realised and are used to the benefit of capturing derived data provenance trials.

Specifically, the extensions and modifications are introduced to the model underpinning ICAT 3.3 along the following directions:

- adding a *SoftwareExecution* subclass of investigation;
- linking program to a software execution;
- linking software executions with datasets;

[15] Hereafter, we use the following terms interchangeably: manifest and map.

- associating parameters with a software execution;
- re-introducing the *study*; and
- introducing study nesting.

We shall now describe these extensions and the rationales behind them.

4.1. Adding a SoftwareExecution Investigation Type

As discussed in Section 2, an investigation models a data handling activity, which, in the current model, means three types of activities: measurements, experiments, and simulations [10]. None relates to the data handling activities in an data analysis process.

A new type of investigation, *SoftwareExecution* is introduced to model the *executions* of *one data analysis task* in the process. In modern research, a task is typically the running of a piece of software. Our model does *not* mandate what a piece of software might be. It can be a system of programs, for example `MATLAB`, which consist of many programs implementing various functionalities. Or, it can be one program implementing a specific function as part of a system of programs, for example, a Fast Fourier transformation function in `MATLAB`. The decision of such is left to the users of our software, the researchers, because only they know about what is the most suitable and useful representation of the execution of a task in their data analysis process.

As illustrated in Figure 4, the SoftwareExecution concept captures the scenario that a piece of software is executed many times, yielding results, each corresponding to a combination of the following components:

- the software used for the execution,
- the parameters (e.g. the settings of experiments or simulations where an input dataset is obtained, or the settings for an analysis methodology),
- the input datasets (e.g. readings captured from instruments or simulations, or from another SoftwareExecution), and
- the output datasets (e.g. the results of an analysis function).

We use an analogy to explain the rationale behind the modelling of this concept. A (research) problem can be analogous to a 'puzzle', a task the

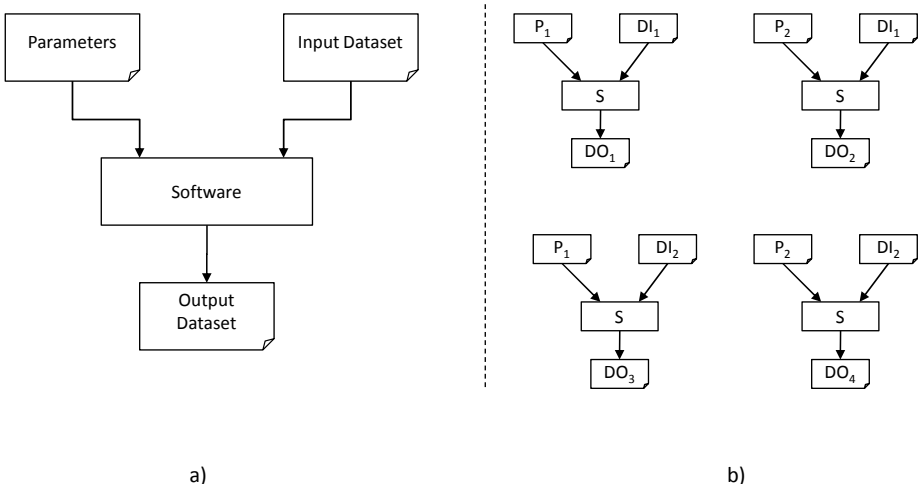

Figure 4: a) The SoftwareExecution Concept; b) Four SoftwareExecutions depicting the scenario of running the same piece of software four times, each taking a set of parameters and input datasets, and yielding different output datasets

building of a 'jigsaw' in order to solve one part of the 'puzzle', and the execution of a task a 'jigsaw' built with one combination of parameters, datasets, and program settings to solve that part of the 'puzzle'. Every researcher has an idea of what tasks are required to solve a problem and how to execute the tasks. Because research is often open-ended and iterative, the usefulness and impact of an (early-stage) analysis is often difficult to judge. Therefore, keeping track of all the potentially useful 'jigsaws' along a research trial is not only important but also valuable to researchers.

'Jigsaws' can be built by the same person or independently by different people in solving a part of a puzzle. Once the 'jigsaws' are built, they can be put together to a) either lead to different solutions to the same problem; or b) form new 'jigsaws' to solve a different problem, leading to unexpected new discoveries. For the former, judging which solution is the best is an issue beyond the scope of this paper. However, we believe that capturing the different 'paths' presented in derived data provenance trails can be a powerful approach for addressing research problems. Sections 4.5 present how we put these 'jigsaws' together to form a big picture of a solution to solve a 'puzzle'.

4.2. Linking Software to SoftwareExecution

SoftwareExecution is a runtime notion meaning that it is not only associated with a software program but also inputs (including data files and the parameters) that drive the program and the corresponding outputs resulted from running the program using those inputs. A software execution comprises of: one (*and only one*) program, one or more input datasets, one or more output datasets, and zero or more parameters (also parameter or configuration files) to the program. It is worth noting that a program in this context *could be* formed from a number of programs linked together. A typical example of such is the use of a script to pipe data through a number of different stages and each stage is controlled by a different program. But, because the intermediate inputs and outputs, streamed or kept in temporary files for intermediate steps, are not significant in the derived data provenance trial, all the programs involved in the script are abstracted into one program, i.e. the script, to be represented in the trial.

In contrast to the temporary nature of the intermediate inputs, outputs, and the programs, the inputs, outputs, and the program associated with a SoftwareExecution stage are persisted and catalogued systematically. We should emphasis that it is often also the case that researchers *themselves* decide what aspects of a derived data provenance trail need to be persisted and catalogued. Neither our model nor any software implementing our model should mandate that.

Capturing computer software related to data is a complex issue; see for example [22, 21] for a consideration of the complexity of characterising software. For example, even with the same software, there are often different versions in existence and they may or may not compatible with each other, and may behave differently in different execution environments. In this paper, we take a simplified view of software by focusing on the data aspect of the derived data provenance trails. In practical terms, this means that the inputs (parameters or datasets) used for one version of a program may not be workable with another one. We are aware of this issue and consider it a topic for further research.

4.3. Linking SoftwareExecutions to Datasets

Software executions are linked to datasets to capture their context in the provenance process as follows.

Input and Output Datasets

Two types of datasets are introduced to denote the inputs to and outputs from an execution of a program. Note that they are associated with an *execution of a program* not the program itself. This is an important aspect of the analysis we would like to capture reflecting the the open ended nature of scientific research.

Associating Multiple SoftwareExecutions to Input Datasets

In the current model, there is an one to many relationship between investigation and dataset. However, a program can run many times using different sets of parameters but with the same input dataset. Hence, the relationship between investigation and dataset is extended to be many to many so that it accommodates this scenario.

4.4. Associating Parameters with a SoftwareExecution

One program can be executed several times resulting in several (program) executions. All can correspond to the same input dataset(s) but with different output datasets and runtime parameters. A program can take zero or more parameters, but a parameter must be associated with at least one software execution. The linkage between RuntimeParameter and Program is through SoftwareExecution.

4.5. Re-introducing Study and Nested Study

The *Study* is a concept designed for grouping related investigations together to capture a common intentionality, such as a research programme, or the analysis of a particular compound. It is a part of the existing CSMD model, although not currently implemented in the current ICAT 3.3 as that it tailored to capture the generation of raw data from a facility instrument, thus each investigation is the unit of intentionality and investigation and study are in a one to one relationship, so study is seen as superfluous. However, this is not the case when we need to consider derived data, and we use it to capture the means by which SoftwareExecutions are related to each other. For example, in the analysis process, a study is used to group SoftwareExecutions together in a particular order. The ordering depicts explicitly the relationship between the investigations reflecting the sequence of the data handling activities involved in a scientific process.

Through a study, SoftwareExecutions can be chained together to form a connected sequence of analysis activities in the process. For example, using

the same set of programs, executions can be chained together to form an analysis flow reflecting the use of a set of input data files and parameters. A different chain can be formed reflecting the use of a different set of files and parameters. The exended study concept provides an end-to-end support for data management covering the experimental data gathered from instruments, to intermediate derived data generated during the analysis process, to the resultant data finally appeared in papers.

A nested study is a notion for nesting a study inside one or more *other* studies. When a study nests inside another study, the former is called a `childstudy` and the latter a `parentstudy`. When a study is a child of several studies, the parent studies share the same task in their analysis process.

4.6. Observation

Although the extended model we have presented has been developed for solving the data management problem in context of "big science" carried out in large-scale facilities, we believe that it is a discipline neutral approach that can be used to solve derived data management problems in other disciplines, and also in the small-scale context found in the university research laboratory.

5. iCat-Personal: A Pilot Implementation

A pilot implementation of the extended CSMD model has been developed for the purpose of supporting the capturing, cataloguing and storing of derived data for *individual researchers*, typically working in university research laboratories. Such researchers may have limited capability for systematic data management, and thus this approach offers a rigorous but feasible method to capture data generated in laboratory analysis and make it retrievable and reusable. Because it is designed to tackle data management problems of individuals, it is named **iCat-Personal**. It is available through the sourceforge website. We shall describe its design and development focusing on the current capabilities of the implementation.

5.1. System Architecture

Figure 5 illustrates the system architecture of iCat-Personal, an Java-based implementation for data ingestion and restoration, and a PHP-based web application for data browsing. It consists of three layers: clients, utility programs, and a repository. Two types of clients are supported: command line scripts for getting the data into the repository and restoring data from a

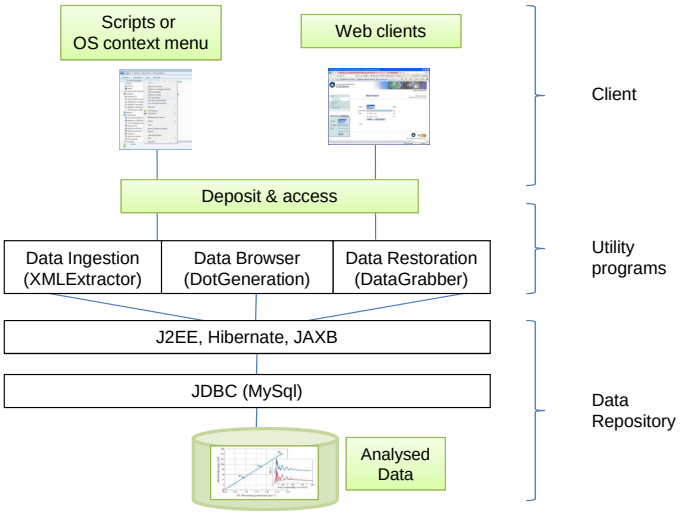

Figure 5: ICAT-Personal System Architecture

repository, and a web browser interface for browsing and navigating derived data stored in the repository. The utility program transforms the data sent from the script and ingest them into a persistent data repository through Java entity beans over a Hibernate-based persistence layer underpinning by a MySQL relational database.

Three key functions of data management are supported, they are: data ingestion, data browsing, and data restoration; all are underpinned via a data catalogue. The targeted audience of this implementation is individual scientists who need a data management tool to assist their research. Future research will investigate how well the model accommodates issues of remote data storage (instead of storing the data locally on the same computer as the source of the derived data), data reuse (e.g. secondary analysis and cross analyses study), and data sharing (e.g. derived data publication, linked data, and its relevance to automated experimentation). As a pilot implementation, data annotation, searching and discovery, although important, are not considered in the implementation.

The object model presented in the previous section is mapped into two data models: a XML schema and a database schema. The former is used by the client to guide the ingestion of derived data provenance into an iCAT-Personal data repository; whilst the latter is the database structure under-

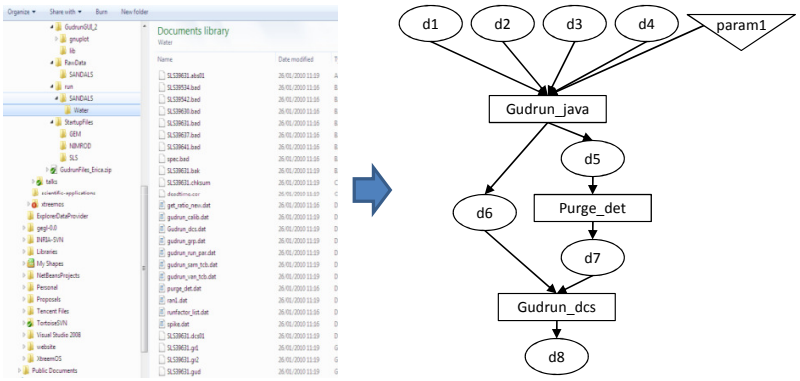

Figure 6: Derived Data Management: An Example

pinning the repository. We use the Gudrun program in the RMC workflow to explain their role in managing derived data.

5.2. Derived Data Management

Figure 6 illustrates the "before" and "after" scenarios of using ICAT-Personal tools to manage derived data. The left hand side is a number of hierarchical file folders where scientists store the programs, scripts, raw data files, instrument settings, and initial parameter inputs to programs, each in a separate directory. The last one is called a working directory where the parameters (stored in a configuration file), raw data input files, intermediate and final output files reside. Each *execution* of the programs corresponds to a *separate* working directory. As with the RMC analysis process, many scientific analyses involve several programs. Scientists often end up with many working directories, each storing the data resulted from one execution. Managing such working directories is challenging because:

- Most programs are run many times. Until the final results at the end of the analysis process are available, it is sometimes difficult to tell which executions are useful. So, all the potentially useful ones need to be kept.

- Scientists also need to keep track of the relationships between the executions. Again, until the final results are available, all the links (which often mean many directories, and sub-directories) have to be kept.

- Different scientists have their own way of keeping the parameters and settings of each execution. These may be stored in annotation files in the working directory, or in a paper lab notebook for example. Without the parameters, it is hard to understand the outputs from the programs or continue other researchers' analyses.

As a consequence, even with the raw data, it is often difficult for one scientist to understand, reproduce, and reuse the derived data produced by another scientist.

5.2.1. Data ingestion

On the right hand side of Figure 6 is a *structured representation of the executions* of the programs involved in Gudrun. The structure represents how the executions of different programs inside Gudrun are linked together. ICAT-Personal tools store the structure as well as the contents[16] (e.g. datafiles, datasets, programs, parameters) inside the structure into an ICAT-Personal repository underpinned by the J2EE technologies depicted in Figure 5. This process is called ICAT-Personal data ingestion. It is guided by an ICAT-Personal XML schema compliant XML file. Figure 7 illustrates a snippet of such an XML file, which captures:

- programs in the process (line 1),
- inputs, including data files and parameters (or parameter files), to and outputs (e.g. data files, plots) from the programs (lines 2 - 7),
- datasets, the logical groupings of the datafiles (lines 8 - 18),
- SoftwareExecutions and their linkage with datasets (lines 19 - 31), and
- the order of the software executions (lines 32 - 39)

A Java based `XMLExtractor` program, built upon the JAXB technology, is used to parse the XML file and generates Java entity beans from the XML. The beans are ingested into the database via a Hibernate-based persistence layer over MySQL.

[16]Initially, the contents are stored in the database. But soon, the size of the database increases rapidly as more derived data are pushed into the database. So, currently, the files are stored in the file system and the database only stores the links to datafiles.

```
<processes><process id="gudrun_java" type="java program">...</process>...</processes>
<parameters><parameter id="param1"> ... </parameter> ... </parameters>
<datafiles>
    <datafile id="df1"><name>Gudrun_dcs.txt</name> ... </datafile>
    ...
    <datafile id="df18"><name>SLS39542.RAW</name> ... </datafile>
</datafiles>
<datasets>
 <dataset id="d1">
         <datafileref idref="df1"/>
         <datafileref idref="df6"/>
         ...
     </dataset>
     <dataset id="d2">
         <datafileref idref="df2"/>
     </dataset>
     ...
 </datasets>
 <investigations>
     <investigation id="i1" type="softwareexecution">
         <processref idref="gudrun_java"/>
         <datasetref idref="d1" type="output"/>
         <datasetref idref="d2" type="output"/>
     </investigation>
     <investigation id="i2" type="softwareexecution">
         <datasetref idref="d2" type="input"/>
         <processref idref="purge_det"/>
         <datasetref idref="d3" type="output"/>
     </investigation>
     ...
 </investigations>
 <studies>
     <study id="s1">
         <name>Gudrun Data Reduction Study</name>
         <investigationref idref="i1" />
         <investigationref idref="i2" />
         <investigationref idref="i3" />
     </study>
 </studies>
```

Figure 7: Example XML file capturing a Gudrun execution process

5.2.2. Data Browsing

An ICAT-Personal tool, named DotGeneration, provides data browsing capability. It takes an ICAT-Personal data ingestion XML file, transforms it into a Graphviz[17] dot file, and generates a flow diagram as depicted on the right hand side of Figure 6. In the current database schema, the dot file and the corresponding snippet of the XML are also stored in the database. A PHP based web application has also been developed to display the relationship between programs, parameters, datafiles, datasets, SoftwareExecutions and studies[18].

Datasets, labelled as `d1` to `d8` in the Figure 6, are used to capture the relationship between data files produced or consumed by one execution. Among all the input data files to Gudrun_java, four datasets are used, they represent four groups/types of data: raw data, sample and vanadium metadata[19] , instrument data, and neutron/x-ray information, respectively. Other scientists may consider different types of relationships between the files by classifying them into three datasets: raw, correction, and calibration data. Such grouping is important because the relationships between the files are not self evident by examining them directly.

5.2.3. Data Restoration

As presented in the previous section, a SoftwareExecution is an encapsulation of the objects (the program, and the inputs and parameters to and outputs from the program) involved in running a software application. Three ordered SoftwareExecutions, corresponding to Gudrun_java, Purge_det, and Gudrun_dcs, respectively, are grouped into one study, which represents an *instance* of the data reduction process, involving

- all the programs, and
- all the raw and derived data, comprising of:
  - all the initial input data files,
  - environment and instrument settings,
  - parameters that used to drive the programs,

[17]http://www.graphviz.org

[18]See [23] for a screenshot of the web interface.

[19]Vanadium metadata is used as a calibration dataset in the Gudrun process.

– all the intermediate outputs, and

– finally to the reduced data files.

This process can be repeated many times leading to many studies (i.e. execution instances) of the process. Each corresponds to a combination of three SoftwareExecutions captured by the ICAT-Personal data management tool. Structured data at various levels (dataset, investigation, and study) can then be restored using the ICAT-Personal DataGrabber tool from the repository.

6. Discussion and Future Work

The data management approach to handling the analysis process would seem well matched to the infrastructure supporting structural science in facilities and potentially a wider scientific community. Storing and retrieving data from throughout the scientific process is a common problem across many disciplines that exploit computational methodologies and high throughput data handling techniques. The analysis presented here in detail only addresses a single study in earth sciences, while other studies in chemistry and crystallography have contributed to the analysis leading to the proposals for changes to the CSMD, and the approach described is also now being generalised into a common information model for structural science in the I2S2 project[20]. This common model combines the expressive power in describing the context and structure of data collections offered by the CSMD with the conceptual framework for modelling experimental process planning and enactment offered by the ORE-CHEM [3], and models a wide range of activities within the scientific life cycle.

It is nevertheless a concern whether the breadth of tasks analysed reflects the whole scope of the target system. At present the usage patterns of the facilities considered are reflected in the sample of tasks analysed, but that may change over time. Other facilities may need to be supported by the CSMD which will introduce further disciplines and different data transformation processes. In particular, if disciplines such as astronomy and earth observation data were to be included, the data collection and analysis processes from those disciplines might lead to further changes to the CSMD.

[20]http://www.ukoln.ac.uk/projects/I2S2/

The changes proposed to the CSMD capture the source of the data, and the transformation process that is has gone through, and reflects the Open Provenance Model [11], but the implementation does not provide a comprehensive provenance management system. [8] argues that a provenance management system can only be useful for a real world application if it allows querying of provenance information for resultant data items. It is unrealistic to expect a complete provenance management system which will use provenance data to automatically recreate resultant data items by executing the transformations that were used in its creation [7].

It would be possible to enhance the ICAT prototype to allow the propagation of the complete provenance of resultant data so that researchers can query it for the transformations used without having to successively unpack the datasets involved. In a simple example, if it becomes known that a particular version of a piece of software was unsafe for a parameter range, the provenance could be queried to provide all resulting data that was produced by using that software in its unsafe range. A more complex example would query for a combination of transformations within the provenance from different datasets in a study, e.g. programs X and Y were used consecutively in the transformation when their underlying models have been found to be incompatible and the resultant data could be unsafe. Such advances on the current implementation would clearly add to the safety of the scientific results derived from the transformations recorded in the provenance. However, such a query system would require the automated splitting of datasets to isolate the transformation data and the subsequent merging of that transformation data for each stage into a single provenance item describing the overall process. Such provenance records would then have to be open to be queried for co-occurring transformations or transformation parameter values Such a system is beyond the scope of the current development, although it could not be attempted without the work presented here to build upon, and they could be a topic for future work.

The tools considered in the analysis simply consume and produce files. Some more sophisticated, but very commonly used, tools such as Chimera [13] offer a Virtual Data Catalog as a relational database for provenance information where users register transformations, data objects and derivations. Such records could be incorporated into datasets in CSMD, but to use them profitably would require the access by the appropriate RDBMS to provide a query interface. This would add further complexity to any attempt to develop a complete provenance management system around the CSMD.

The scientific process described above was undertaken as publicly funded research for which the main security concerns are to embargo release of data until after the scientists undertaking the experiments have published their results and then to make them as publicly open as possible to gain maximum value from the investment. However, large facilities of the class considered in this paper are also used by commercial organisations, or academics funded by commercial organisations. In these cases there may be more exacting security concerns. The modifications proposed here to account for derived data address the Core part of the CSMD only. The second main module of the CSMD addresses security metadata. It is common in these circumstances for all derived data to be required to be handled as the original data received in which case a single data policy would apply to the whole CSMD record. However, security policies are becoming more sophisticated and it is possible for the derivation process to either reduce or, more likely, increase the security constraints on data as it moves through the scientific process and its value increases. When different policies apply to the derived data from the original data then the current single CSMD security node will not be enough, but would have to link policies to individual datasets. Alternatively, the current single security node could be maintained with the use of more sophisticated policies that refer to differently labelled data items explicitly [15]. As commercial use of large facilities becomes more common security issues will become increasingly important to resolve and standardise.

A recent proposal advocates encapsulating published data files in self-contained units of knowledge which they term *Research Objects* - semantically rich aggregations of resources, that possess some scientific intent or support some research objective [1, 2]. A research object bundles together essential information relating to experiments and investigations. This includes not only the data used, and methods employed to produce and analyse that data, but also the people involved in the investigation. The authors present a number of principles that they expect such objects and their associated services to follow: reusable, repurposeable, repeatable, reproducable, playable, traceable. These are indeed the properties which the CSMD records have in principle after the inclusion of the modifications proposed in this paper. The authors propose the use of rich ontologies to encode these properties as an essential requirement for their usability. The current CSMD lacks such semantically rich encoding, but this again would appear to be a clear direction for further development.

Finally, we should point out that the current work only captures several

fairly limited aspects of software and software executions. At this stage, our aim is to understand its relevance to data provenance. It is not our aim to realise the so-called "one-click" execution dimension of scientific process management. We feel that this is just the beginning to unveil the challenges of dealing with software and executions (e.g. hardware, OS, environment variables, support libraries) in the process, which embrace issues such as handling the relationship between a software execution and a software version, deciding what aspects of a software and executions are needed to be captured, and how to capture them.

Acknowledgments

This research was supported by the JISC's Managing Research Data Programme under the Infrastructure for Integration in Structural Sciences (I2S2) project. The authors would like to thank Prof Martin Dove from Earth Sciences at the University of Cambridge, Simon Coles from the UK National Crystallography Service, and Dr. Alan Soper from STFC ISIS facility for providing case studies of the scientific process on STFC facilities leading to the evidence for the proposed modifications to the CSMD.

[1] Bechhofer, S., De Roure, D., Gamble, M., Goble, C. and Buchan, I. (2010) Research Objects: Towards Exchange and Reuse of Digital Knowledge. In: The Future of the Web for Collaborative Science (FWCS 2010), April 2010, Raleigh, NC, USA.

[2] Sean Bechhofer, John Ainsworth, Jiten Bhagat, Iain Buchan, Philip Couch, Don Cruickshank, David De Roure, Mark Delderfield, Ian Dunlop, Matthew Gamble, Carole Goble, Danius Michaelides, Paolo Missier, Stuart Owen, David Newman, Shoaib Sufi, *Why Linked Data is Not Enough for Scientists*, eScience, IEEE International Conference on, pp. 300-307, 2010 IEEE Sixth International Conference on e-Science, 2010

[3] Mark Borkum, Carl Lagoze, Jeremy Frey, and Simon Coles. *A Semantic eScience Platform for Chemistry*, eScience, IEEE International Conference on, pp.316-323, 2010 IEEE Sixth International Conference on e-Science, 2010

[4] Simon J. Coles, Jeremy G. Frey, Michel B. Hursthouse, Mark E. Light, Andrew J. Milsted, Leslie A. Carr, David DeRoure, Christopher J. Gutteridge, Hugo R. Mills, Ken E. Meacham, Michael Surridge, Elizabeth

Lyon, Rachel Heery, Monica Duke, and Michael Day, *An E-Science Environment for Service Crystallography - from Submission to Dissemination*, J. Chem. Inf. Model., 2006, 46(3), pp.1006 - 1016.

[5] Damian Flannery, Brian Matthews, Tom Griffin, Juan Bicarregui, Michael Gleaves, Laurent Lerusse, Roger Downing, Alun Ashton, Shoaib Sufi, Glen Drinkwater, Kerstin Kleese, *ICAT: Integrating Data Infrastructure for Facilities Based Science*, e-science, pp.201-207, 2009, Fifth IEEE International Conference on e-Science, IEEE Computer Society.

[6] M Folk, A Cheng, K Yates (1999) HDF5: A file format and I/O library for high performance computing applications, Proceedings of Supercomputing'99, ACM SIGARCH and IEEE, (Portland, OR), Nov. 1999.

[7] Ian T. Foster. *The virtual data grid: a new model and architecture for data-intensive collaboration*. In SSDBM 2003: Proceedings of the 15th international conference on Scientific and statistical database management, Washington, DC, USA, 2003. IEEE Computer Society.

[8] Boris Glavic and Klaus R. Dittrich. *Data Provenance: A Categorization of Existing Approaches*. In Datenbanksysteme in Business, Technologie und Web (BTW 2007), pp. 227 - 241, 2007.

[9] P. Klosowski, M. Koennecke, J. Z. Tischler and R. Osborn, *NeXus: A common format for the exchange of neutron and synchroton data*, Physica B: Condensed Matter, Vol 241-243, Dec 1997, pp151-153, Proceedings of the International Conference on Neutron Scattering.

[10] Brian Matthews, Shoaib Sufi, Damian Flannery, Laurent Lerusse, Tom Griffin, Michael Gleaves, Kerstin Kleese. *Using a Core Scientific Metadata Model in Large-Scale Facilities*. The 5th International Digital Curation Conference, London, England, 2-4 December 2009.

[11] Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., Plale, B., Simmhan, Y., Stephan, E. and Van den Bussche, J. *The Open Provenance Model core specification (v1.1)*. Future Generation Computer Systems, in Press, 2010.

[12] Tom Oinn, Matthew Addis, Justin Ferris, Darren Marvin, Martin Senger, Mark Greenwood, Tim Carver, Kevin Glover, Matthew R. Pocock,

Anil Wipat and Peter Li, *Taverna: a tool for the composition and enactment of bioinformatics workflows*, Bioinformatics, Vol. 20(17) 2004, pp3045-3054.

[13] Eric F. Pettersen, Thomas D. Goddard, Conrad C. Huang, Gregory S. Couch, Daniel M. Greenblatt, Elaine C. Meng, Thomas E. Ferrin, *UCSF Chimera - A visualization system for exploratory research and analysis*, Journal of Computational Chemistry, 25(13), 1605 - 1612, 2004.

[14] R Rew, G Davis, *NetCDF: an interface for scientific data access IEEE Computer Graphics and Applications*, 10(4), 76-82, 1990.

[15] Enrico Scalavino, Vaibhav Gowadia, and Emil C. Lupu (2010) A Labelling System for Derived Data Control, in Sara Foresti and Sushil Jajodia (Eds.) Data and Applications Security and Privacy XXIV: Proceedings of DBSec 2010, the 24th Annual IFIP WG 11.3 Working Conference, LNCS, Springer-Verlag:Berlin.

[16] Andrew Shepherd (2001) Hierarchical task analysis, Taylor & Francis: London.

[17] Shoaib Sufi and Brian Matthews, *A Metadata Model for the Discovery and Exploitation of Scientific Studies*. In Domenico Talia, Angelos Bilas and Marios D. Dikaiakos (Eds.) Knowledge and Data Management in GRIDs, 2007, pp135-149, Springer: Berlin.

[18] MG Tucker, DA Keen, MT Dove, AL Goodwin and Q Hui. *RMCProfile: Reverse Monte Carlo for polycrystalline materials*. Journal of Physics: Condensed Matter 19, art no 335218 (16 pp), 2007, available at: `http://wwwisis2.isis.rl.ac.uk/rmc/`.

[19] Erica Yang. *Martin Dove's RMC Workflow Diagram*. Project Requirement Report (supplementary report) for the I2S2 project, July 2010. Available at: `https://www.jiscmail.ac.uk/cgi-bin/filearea.cgi?LMGT1=I2S2\&f=/Deliverables/RequirementsReport`.

[20] Yu, J. and Buyya, R. 2005. *A taxonomy of scientific workflow systems for grid computing*, SIGMOD Rec. 34, 3 (Sep. 2005), pp44-49.

[21] Brian Matthews, Arif Shaon, Juan Bicarregui, Catherine Jones, *A Framework for Software Preservation*, International Journal of Digital Curation, 5 (1), 2010.

[22] Brian Matthews, Arif Shaon, Juan Bicarregui, Catherine Jones, J Woodcock, Esther Conway, *Towards a Methodology for Software Preservation*, Proc. 6th International Conference on Preservation of Digital Objects (iPres 2009), San Francisco, USA, 05-06 Oct 2009.

[23] Erica Yang, Brian Matthews, Michael Wilson, *Accommodating Derived Data with an Enhanced Core Scientific Metadata Model*, Technical Report Series, Rutherford Appleton Laboratory, 28, Oct. 2010, available at: `http://epubs.stfc.ac.uk/bitstream/6032/RAL-TR-2010-028.pdf`.