Enhancing the Core Scientific Metadata Model to Incorporate Derived Data

Erica Yang, Brian Matthews, Michael Wilson
STFC e-Science
Rutherford Appleton Laboratory
HSIC, Didcot, Oxon, OX11 0QX, UK
Email: {Erica.Yang, Brian.Matthews, Michael.Wilson}@stfc.ac.uk

***Abstract*—The Core Scientific MetaData model (CSMD) is used by 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. 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. An analysis of the raw 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. 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.**

***Index Terms*—large scale facilities; neutron sources, scientific process, data management, data sharing, data linking.**

I. 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 need access to data to manage the use of their facilities; and other scientists 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.

The Core Scientific MetaData model (CSMD) [13], [8] 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 [3] which is used by several large scientific facilities, in particular, the ISIS neutron source[1] the Diamond Light Source (DLS)[2], and the Institut Laue-Langevin (ILL)[3]. Data cataloguing systems support access to scientific data, but the present CSMD only addresses the raw data produced by the facility and 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. At present these intermediary derived data sets must be stored locally by the scientists, and are not archived for other purposes. Thus the support for the intended users is partial.

Bioscientists have used workflow tools to capture and automate the flow of analyses and the production of derived data for many years [9] and can now automatically run many computational workflows [16]. 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 provenance [6], 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.

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

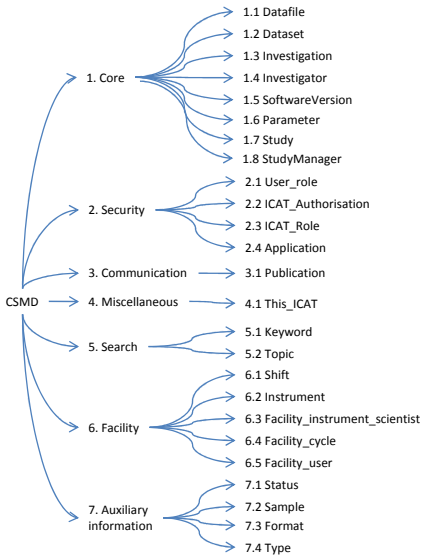

Fig. 1. A classification of the concepts in CSMD

II. Core Scientific MetaData model

The Core Scientific MetaData model (CSMD) [13] 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.

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.

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) [10] and Hierarchical Data Format (HDF) [4] are well defined formats used by many laboratories, while NeXus [7], 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 any data generation activity. 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 of a sample and a related simulation of the process, can 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, i.e. ICAT 3.3[4], is based on the CCLRC Scientific Metadata Model v2 [13] 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 six categories (see Figure 1):

a) **Core:** these concepts are core to scientific data management. Capturing the data outputs involve four data objects: datafile, dataset, investigation, and study. A datafile is 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. Investigator and StudyManager are people associated with an investigation and a study, respectively. Process[5] is an activity that produces or consumes data objects. Parameter is the context of the data production process.

b) **Search:** classifiers which facilitate the search and discovery of core concepts.

c) **Communication:** entities linking between research objects so that the provenance of a research publication can be traced back to the data holdings.

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

[5]In CSMD 2.0 and ICAT 3.3, the concept Process is called SoftwareVersion.

d) **Security:** entities which enforce access policies on the data holdings.

e) **Miscellaneous:** entities which identify the specific instance of ICAT metadata catalogue.

f) **Facility:** specific concepts related to facilities. They are introduced to capture the contextual information (e.g. which facility, instrument, shift, the data is collected, how it is collected, the instrument settings) associated with the (raw) data collection process.

g) **Auxilliary Information:** the information associated with data holdings. It is currently being used to store information related to *raw* data files, such as sample, parameters (e.g. temperature, humidity), file format. But it should be possible to extend or adapt them 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.

III. 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.

A. 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) [2] 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[6] provided by the GEM instrument[7] and the other small angle neutron scattering[8] offered by the Sandals instrument[9]. 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 [2].

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

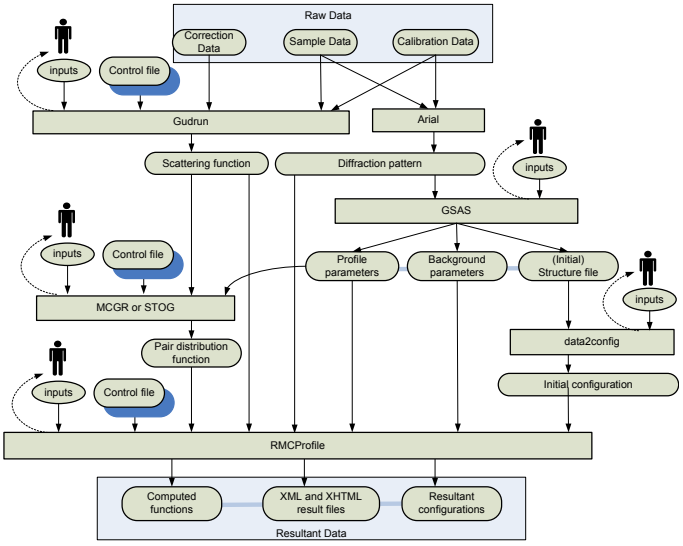

Fig. 2. The RMC data analysis flow diagram

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.

In practice, it can take months from the point that a scientist obtains the raw data to the point where 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 [12] has been applied to all the studies and the abstractions do generalise across instruments, techniques, programmes and disciplines.

B. 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 [15]. The RMC method is probabilistic, which means that a) it can only deliver 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 hand-written 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.

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`[10] are a set of scattering functions, one for each bank of detectors. For `Arial`[11], 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.

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 `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 byeway. 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) 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` [14] programme which implements 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.

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

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

C. 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 or better programs) 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.

IV. An Enhanced CSMD

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

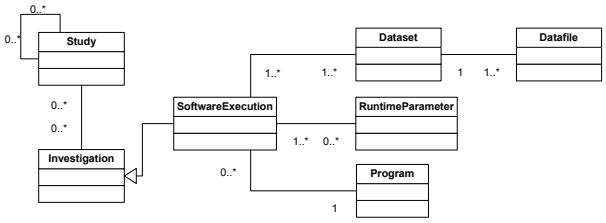

Fig. 3. An Extended CSMD UML Model for Supporting Derived Data

- the *data objects* involved;
- the *programs* that produce or consume data objects;
- the *ordering* of the programs; and
- the *parameters* to the programs.

Figure 3 is an UML object model depicting the extensions and modifications to the core of the existing CSMD model to support derived data. The actors are not included in the diagram because security is not the prime concern in this paper. Due to the space restriction, we cannot present the full details (attributes and methods) of each object in this paper, which are accessible via the sourceforge website of the implementation[12]. Specifically, the extensions are introduced to the model underpinning ICAT 3.3 along following directions:

- adding a SoftwareExecution type to investigation;
- linking program to a software execution;
- linking software executions with datasets;
- associating parameters with a software execution; and
- introducing nested study.

We shall now describe the rationales behind these extensions.

A. Adding a SoftwareExecution investigation type

As discussed in Section II, an investigation models a data handling activity, which, in the current model, means three types of investigations: measurements, experiments, and simulations [8]. None relates to the data handling activities in an analysis process.

A new type of investigation, *SoftwareExection* is introduced to model the *runtime* processing units in the process. This extension provides an end-to-end support for data management covering the experimental data gathered from instruments, to intermediate data generated in the process, to the resultant data finally appeared in papers.

B. Linking program to SoftwareExecution

A software execution represents an execution of a computer program for a part of the analysis in the process. It is a runtime notion meaning that it is not only associated with a static 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 to the program.

[12]http://icatlite.sourceforge.net/

C. Linking software executions to datasets

1) Input and output datasets: Two types of datasets are introduced to denote the inputs to and outputs from an execution of a program. *They are associated with an execution not the program*. This is an important aspect of the analysis we would like to capture reflecting the the open ended nature of scientific research.

2) Associating multiple software executions to an input dataset: 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.

D. Associating parameters with 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.

E. Study and nested study

Study is a notion for grouping related investigations. It is the means by which SoftwareExecutions are related to each other and SoftwareExecutions are related to other types of investigations. For example, in the analysis process, a study is used to group investigations in a particular order, which can be *sequential, parallel, and adjunctive*. The ordering depicts explicitly the relationship between the investigations reflecting the sequence of the data handling activities involved in a scientific endeavour.

Through a study, the investigations 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.

It is not uncommon that iterations of analyses are performed before a satisfied set of results can be obtained. Several of such "chains" can be formed when conducting an analysis process. A nested study is a notion for grouping related studies (or chains). Such relationship can be *adjunctive* in that the output from one study is used as the input to another. The studies can be *parameter sweeps* in that two studies use the same set of programs and input data files but with different runtime parameters. They can also be *functionally equivalent* when two studies use the same set of inputs (data files, parameters) but with a set of functionally equivalent programs.

V. ICatLite: A Pilot Implementation

A pilot implementation of the extended CSMD model, named **ICATlite**, has been developed and is available through

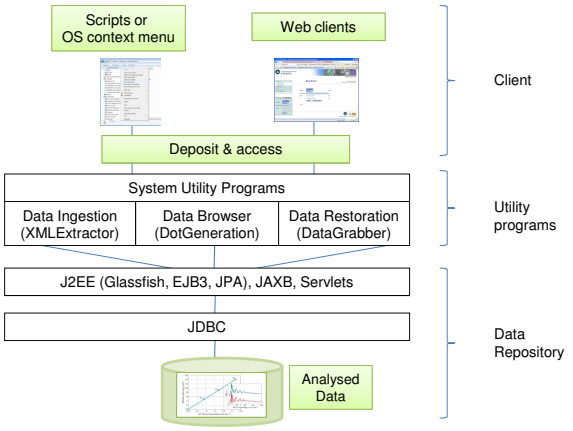

Fig. 4. ICATlite System Architecture

sourceforge website. It is a lightweight version of ICAT because it implements the core of the extended CSMD model to demonstrate the feasibility of capturing and cataloguing derived data. We describe its design and development focussing on the current capabilities of the implementation.

A. System Architecture

Figure 4 illustrates the system architecture for ICATlite. It consists of three layers: client, utility programs, and a repository. It supports two types of clients: command line scripts and or native OS context menu. The client tool, including the client-side of the utility programs, needs to be installed on users' computer. The client interacts with the server-side utility programs which connects to the persistent data repository through Java entity beans and other classes hosted by Glassfish, a J2EE container provided by Sun/Oracle. JAXB is used to parse the XML ingestion file and generate Java entity beans from the XML.

Three capabilities of data organization are supported, they are: data ingestion, browsing, and restoration. The targeted audience of this implementation is individual scientists who need a data management tool to assist their own research. Future releases will investigate how well the model accommodates issues of 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 UML model presented in the previous section is mapped into two data models: a XML schema and a database schema. Both are available through the sourceforge website. The former is used to guide the ingestion of data files and programs into an ICATlite repository whilst the latter is the structure underpinning the repository. We use the Gudrun program in the RMC workflow to explain their role in managing derived data.

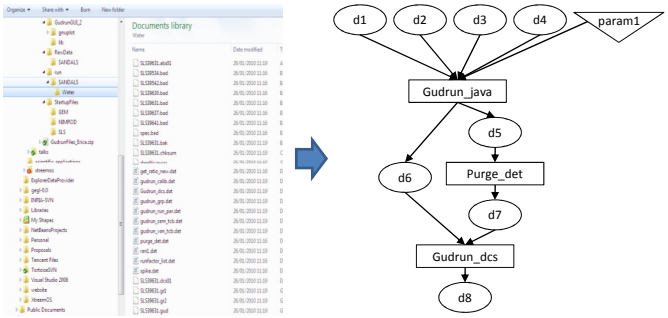

Fig. 5. Derived Data Management: An Example

B. Derived Data Management

Figure 5 illustrates the "before" and "after" scenarios of using ICATlite tools to manage derived data. The left hand side is a number of hierarchical file folders where scientists store the programs, run 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 directories, each storing the data resulted from one execution.

Managing the 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 linkages between the executions. Again, until the final results are available, all the linkages (which often mean many directories, and sub-directories) have to be kept.
- Different scientists have their own way of keeping the parameters (e.g. storing in the working directory, on a paper notebook). Without the parameters, it is hard to understand the outputs from the programs or continue other researchers' analyses. Even with the raw data, it is difficult for other people to reproduce derived data.

1) Data ingestion: On the right hand side of the diagram is a *structured representation of the execution* of the programs involved in Gudrun. The structure represents how the executions of different programs inside Gudrun are linked together. ICATlite tools store the structure as well as the contents inside the structure into an ICATlite repository underpinned by the J2EE technologies depicted in Figure 4. This process is called ICATlite data ingestion which is guided by an ICATlite XML schema compliant XML file. The file captures:

- the inputs, including data files and parameters (or parameter files), to and the outputs (e.g. data files, plots) from the programs;
- which files are produced or consumed in the same context (e.g. belonging to the same SoftwareExecution);

- the programs in the process; and
- the execution order of the programs.

The tools provide two further capabilities: browsing and restoration of the archived executions.

2) Data Browsing: An ICATlite tool, named DotGeneration, provides data browsing capability. It takes an ICATlite data ingestion XML file, transforms it into a Graphviz[13] dot file, and generates a flow diagram as depicted on the right hand side of Figure 5.

Datasets, depicted as `d1` to `d8` in the diagram, are used to capture the relationship between data files produced or consumed by one execution. In Figure 5, among all the input data files to Gudrun_java, four datasets are formulated, they represent four groups/types of data: raw data, sample and vanadium metadata, 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.

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,
  - 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 ICATlite data management tool. Structured data at various levels (dataset, investigation, and study) can then be restored using the ICATlite DataGrabber tool from the repository.

VI. 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 crystalography 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[14].

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 suggestions for change to the CSMD.

The changes proposed to the CSMD capture the source of the data, and the transformation process that is has gone through, but the implementation does not provide a comprehensive provenance management system. [6] 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 [5].

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, although beyond the scope of the current development.

The scientific process described above was undertaken as publically funded university research for which the main security concerns are to embargo release of data until after the scientists undertaking the experiments have publshed their results and then to make them as publically 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

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

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

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 [11]. 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]. An RO 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, tracable. 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.

REFERENCES

[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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

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

[13] 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.

[14] 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/.

[15] 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.

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