Future Generation Computer Systems 29 (2013) 612–623

Contents lists available at SciVerse ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Enhancing the core scientific metadata model to incorporate derived data☆

Erica Yang [a,b,*], Brian Matthews [b], Michael Wilson [b]

[a] Bodleian Libraries, University of Oxford, Oxford OX2 0EW, UK

[b] STFC e-Science, Rutherford Appleton Laboratory, HSIC, Didcot, OX11 0QX, UK

ARTICLE INFO

Article history:
Received 28 February 2011
Received in revised form
1 June 2011
Accepted 5 August 2011
Available online 19 August 2011

Keywords:
Data management
Information management
Derived data management
Data analysis
Data provenance
Large scale facilities
Neutron sources
Scientific process

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 the tools it supports are considered.

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) [1,2] 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 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] are also adopting

☆ A preliminary version of the same title was published at 2011 eScience conference as a full paper.

* Corresponding author. Tel.: +44 1865 277609.
E-mail addresses: erica.yang@bodleian.ox.ac.uk, erica.yang@stfc.ac.uk (E. Yang), Brian.Matthews@stfc.ac.uk (B. Matthews), Michael.Wilson@stfc.ac.uk (M. Wilson).

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

0167-739X/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.future.2011.08.003

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 [4] and can now automatically run many computational workflows [5]. 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-analysis 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.

2. Core Scientific MetaData model

The Core Scientific MetaData model (CSMD) [1] 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 the Science and Technology Facilities Council (STFC) in the UK, 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. It was soon recognised that this metadata model had to have a more structured approach than offered by generic models such as the Dublin Core [7] to reflect the structured relationships between datasets.

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 depend on the equipment, the facility, or the program that the data is produced from. The Network Common Data Format (netCDF) [8] and Hierarchical Data Format (HDF) [9] are well defined formats used by many laboratories, while NeXus [10], derived from HDF5, is a common data format targeted 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 on the CCLRC Scientific Metadata

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

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