Scaling Up Scientific Data Management Infrastructure

Erica Yang (a) Manjula Patel (b) Brian Matthews (a)

a. STFC Rutherford Appleton Laboratory

b. University of Bath & UKOLN & DCC

Abstract

This abstract briefly highlights the data management challenges brought by the advent of modern computational methods and rapidly growing range of high through scientific equipments in the domain of structural sciences. Our requirement gathering exercises have revealed a significant gap between state of the art technologies and current data management practice. There are significant variations in data management requirements between individual researchers and facility service providers. Highly isolated technological solutions have been adopted by different stakeholders, making it hard for researchers to manage their experimental data which can be generated, collected, and analysed over a period of time at places across different collaborations.

We also describe our approach to address this problem by presenting a loosely coupled architectural framework for managing scientific data lifecycles. We expect that the support for overlapping investigations and datasets will open up a whole range of possibilities to cross-examine datasets from different angles over time and space, ultimately, enabling existing isolated data management solutions to scale up to embrace the excitements brought by open research era.

1. Introduction

Structural Science incorporates a number of disciplines including Chemistry, Physics, Materials, Earth, Life, Medical, Engineering, and Technology. Within these disciplinary communities scientific research is conducted at a range of differing scales involving the use of small laboratory equipment to institutional installations to large scale facilities such as the synchrotron facilities at CERN, the DIAMOND Light Source (DLS) and ISIS, based at the Science and Technology Facilities Council (STFC). With improvements in technology there is an increasing demand to make available raw, processed and derived data for validation and reanalysis purposes, necessitating data management of these types of data as well as the final results data.

It is however apparent, that many research scientists capture, manage, discuss and disseminate their data in relative isolation with highly fragmented data infrastructures and poorly integrated software applications. On the other hand, large centralised facilities have a responsibility to provide a data management infrastructure for their users and have spent considerable effort designing and implementing such systems. The outcome is that each large-scale facility has its own, often insular approach to data management resulting in vast ‘data silos’.

Consequently, there is currently a functional chasm between researchers working in their home institution and at centralised facilities such as DLS and ISIS. Researchers need to move data across institutional and domain boundaries in a seamless and integrated manner. Within the I2S2 (Infrastructure for Integration in Structural Sciences) project [1] we are seeking to “bridge” this chasm and develop a robust data infrastructure to enable these seamless flows to take place routinely and thereby increase efficiency. The project aspires to enable integrated research data management that will allow researchers to simply and efficiently manage their data across institutional and administrative boundaries.

2. Use Case Studies

Two complementary use cases are developed to explore the perspectives of ``data scale and complexity'' and ``research discipline'' throughout the research lifecycle.

Research Spanning Organisational Boundaries The first use case is in the domain of Chemistry. It is designed to understand the data management issues across organisational and administrative boundaries - from a lone university researcher (or a research group) in Cambridge, to a medium-size national research support service - National Crystallography Service (NCS) in Southampton, and to a large-scale international research facility – DLS synchrotron at STFC. The basic scenario here is a lone university scientist researching a type of crystal structure in the domain of Chemistry. Experiment facilities with progressively increased scale, power and capabilities, are needed to reach the level of accuracy and precision required at different stages of research. The service nature of facility operators implies an obligation to retain experiment data, maintain administrative and safety data, and transfer data to end-researchers.

Lone Researcher (or small research group) Scenario The second use case involves a closer exploration that studies in detail the processes involved in Earth Science to acquire raw experiment data at the ISIS facility and

generate derived and resultant data. This use case has two purposes: 1) to understand the differences and commonalities of the data management issues between research disciplines so as to ensure that the information model is sufficiently general to characterise the data models used by different research disciplines; and 2) to understand the nature of research data (e.g. size, volume, complexity) and that of research activities in structural sciences to ensure that the information model is powerful enough to embrace the diverse range of research data produced in these domains. Data management needs here largely comprise of a researcher (or another team member) being able to return to and validate results within the (local) research lifecycle.

3. Findings and Discussions

The outputs from our requirement capturing exercises have revealed significant challenges in bridging the gap between state of the art technologies and current data management practice.

Significant Variations in Requirements There is considerable variation in requirements between individual research scientists and (national and international) service facilities. Their view towards data management is often ad-hoc and short-term. Data are stored at personal computers, departmental or university storage systems with little (sometimes random or nil) data structure underlying research data. Because of the main interest of lone researchers is publications, the data lifecycle they are concerned with has a much shorter lifespan than that being considered in the facilities. Data management often ends when the resultant data is published. In contrast, facility operators have the duty and obligations to ensure the long term accessibility and usability of the data they have produced. They take a much forward looking view in ensuring the long term prospect of data. Hence, digital object curation and preservation issues are in their agenda, but not that of lone researchers.

Isolated Technological Solutions At present individual researcher, group, department, institution, facilities all working within their own technological frameworks. Tightly coupled integrations among these frameworks under one general technological umbrella are feasible but not desirable because of the existence of administration and separation between organisations. We expect this situation will remain to be the case in the foreseeable future. For example, university researchers use very diverse range of technologies in their daily data handling and management tasks, ranging from file systems or storage, web systems, in-house tools, to commercial technologies. On the other hand, facility service providers offer a much structured settings to support research data capturing, analysis and handling. For example, both NCS and STFC have defined their own data model to underpin the data infrastructure they provide to researchers, although there are significant semantic differences between the models.

4. Our Approach

A major conclusion from the studies is that there is merit in adopting an integrated approach which caters for all scales of science to facilitate a) efficient exchange and reuse of data across disciplinary boundaries; b) aggregation and/or cross-searching of related datasets; and c) data mining to identify patterns or trends across research datasets beyond the lifecycle of individual research datasets. We have taken up a *loosely coupled* integration approach by adopting a common information model (also called metadata model) underpinning the activities involved in the research lifecycle.

Figure 1 depicts our approach - a loosely coupled architectural framework for scientific data management lifecycles spanning across organisations. The key element in the framework is the information model (an extension of the Core Scientific MetaData (CSMD) model originally developed at STFC) which is a general model for the representation of scientific study metadata developed within the I2S2 project to represent the research data generated from scientific facilities (e.g. DLS, ISIS, and NCS), university research laboratories, and researchers' individual research portfolios. The production data management infrastructure at RAL has adopted an earlier version of CSMD [2].

The I2S2 CSMD is an extended version of that in [3] with newly added capabilities to capture: processes (e.g. program), data (analysis) provenance, secondary analysis, overlapping dataset(s) between investigations and between studies, and nested scientific studies. Capturing processes is an important milestone in enabling data provenance. The very definition of information model focusing on information, but not processes, means that it imposes no constraint (type, sequence) on the processes or process models related to the information or data it describes. Therefore, unlike a workflow model which emphasises processes and the order of processes, CSMD does not impose such restrictions. Hence, it is suited to incorporate any processes related to datasets.

[1] JISC Infrastructure for Integration in Structural Sciences (I2S2) Project, http://www.ukoln.ac.uk/projects/I2S2/
[2] Sufi, S., Matthews, B.: CCLRC Scientific Metadata Model:Version 2. DLTechnical Reports, DL-TR-2004-001, 2004. http://epubs.cclrc.ac.uk/work-details?w=30324, (2004)
[3] Matthews, B, Sufi, S, Flannery, D, Lerusse, L, Griffin, T, Gleaves, M, Kleese, K: Using a Core Scientific Metadata Model in Large-Scale Facilities, 5th International Digital Curation Conference, Dec. 2009, Oxford.

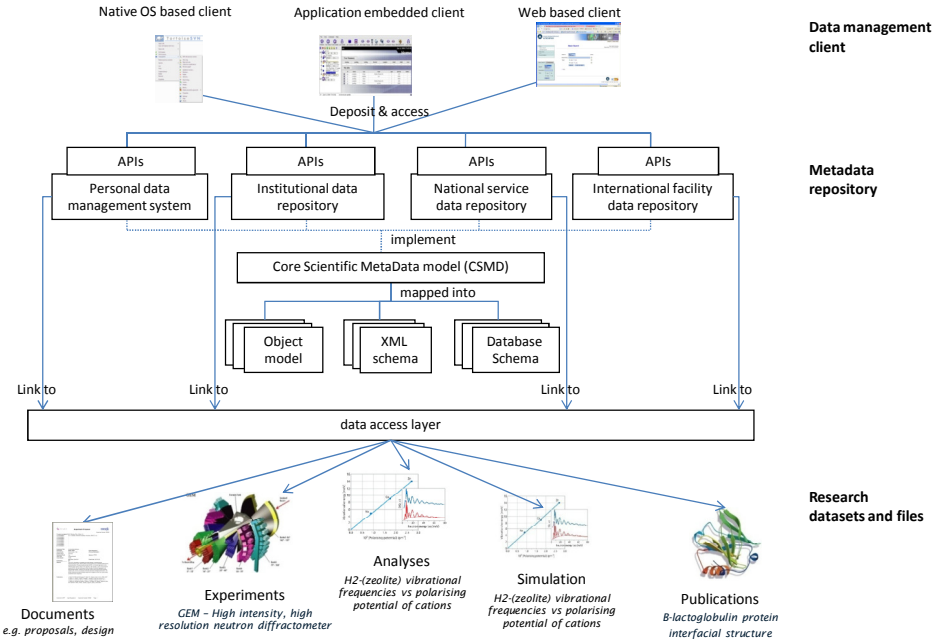

Figure 1 A Loosely Coupled Architectural Framework for Managing Scientific Data Lifecycles