Fusion Engineering and Design 89 (2014) 745–749

Contents lists available at ScienceDirect

Fusion Engineering and Design

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

Automated metadata, provenance cataloging and navigable interfaces: Ensuring the usefulness of extreme-scale data

D.P. Schissel[a,*], G. Abla[a], S.M. Flanagan[a], M. Greenwald[b], X. Lee[a],
A. Romosan[c], A. Shoshani[c], J. Stillerman[b], J. Wright[b]

[a] General Atomics, P.O. Box 85608, San Diego, CA 92186-5608, USA
[b] Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[c] Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

ARTICLE INFO

Article history:
Received 20 May 2013
Received in revised form 17 January 2014
Accepted 21 January 2014
Available online 23 February 2014

Keywords:
Metadata
Workflow
Data management
Data provenance

ABSTRACT

For scientific research, it is not the mere existence of experimental or simulation data that is important, but the ability to make use of it. This paper presents the results of research to create a data model, infrastructure, and a set of tools that support data tracking, cataloging, and integration across a broad scientific domain. The system is intended to document workflow and data provenance in the widest sense. Combining research on integrated metadata, provenance, and ontology information with research on user interfaces has allowed the construction of early prototype. While using fusion science as a test bed, the system's framework and data model is quite general.

1. Introduction

Data, from large-scale experiments and extreme-scale computing, is expensive to produce and may be used for high-consequence applications. However, it is not the mere existence of data that is important, but the ability to make use of it. The practice today with scientific experiments and simulation data is to collect the data first, store it, analyze, and process later. However, much of the data rapidly becomes useless without orderly methods for capturing how the data was generated (what programs, process, and parameter setup), and what products are generated later as a result of analysis processes, and the lineage of these products. Much of the data captured today from experiments and simulations is left untouched (e.g. Ref. [1]), or only touched once. It is essential that data from previous experiments is preserved and all the information about their generation (often referred to as metadata) is captured. In the past this problem was manageable as long as the volume and rate of data generation is small. However, as the volume and rate of data generation grows this problem is becoming a major issue for getting the most value from raw data and data products.

Scientists use workflow scripts to orchestrate the automatic processing of data as it is generated, or for generating data products in subsequent analysis. In such cases, the workflow information is imbedded in the scripts, which are typically not well documented nor easily visible to collaborators.

Experience has shown that as the associated metadata is better organized and more complete, the more useful the underlying data becomes. Further, the set of tools that automatically create, discover, display or explore the semantic relationships of complex data from experiments and computer simulations does not currently exist, though concepts or paradigms from the semantic web may prove useful. Therefore, there is presently an unmet need to better document workflows that create, transform, or disseminate data and to capture (and later present) data provenance. Provenance refers to the lineage of data products. Generous provisioning of metadata, including data provenance and data relations, is critical to enhance data sharing, to allow data to retain its usefulness over extended periods of time, and to provide traceability of results.

The motivation of this research is to greatly increase the value of experimental and computational data across a wide range of scientific domains. Our approach is to reduce the barriers associated with collaboratively using data by creating a complete infrastructure that supports data provenance, cataloging, and ontology along with interactive tools for rapid searching and browsing.

Space limitations preclude an overview of similar work (e.g. Ref. [2]). However, this paper follows previous work [3] and presents early results of research to create a data model, infrastructure, and a set of tools that support data tracking, cataloging, and integration across a broad scientific domain. The system aims to document

[*] Corresponding author. Tel.: +1 858 455 3387.
E-mail address: schissel@fusion.gat.com (D.P. Schissel).

0920-3796/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.fusengdes.2014.01.053

workflow and data provenance in the widest sense, enabling scientists to answer the questions "who, what, when, how and why" for each data element; provide information about the connections and dependencies between the data elements; and allow human or automatic annotation for any data element. The goal is to capture information from the creation, recording/importing of physical data, through various levels of analysis, data preparation, High Performance Computing (HPC) code execution, storage, post processing, data exporting, and publication.

While using Fusion Energy Sciences as a test bed, the goal is to create a conceptual framework and data model that is quite general and does not contain references to the fusion domain. The expectation is that results of this research will be applicable to many if not most science areas. Although the equations solved by simulations are different for different fields of science, the basic flow of information, the need to document workflow and provenance, allowing traceability of results is common to all. Similar common needs exist for experimental data. All fields of science struggle to integrate information from simulation and experiment and to extract knowledge from the comparison of the two. Our goal is to aid this task as well.

2. Approach

The overall development approach is to perform rapid prototyping and testing by real users on real problems at scale. Solution to "toy" problems does not provide the opportunity for useful feedback. Therefore, agile software development is being utilized allowing iterative and incremental development and tools to be demonstrated in large national and international fusion sciences collaborations from which user experience is collected and lessons-learned tabulated to provide feedback for improved design.

The research has been divided into three main areas: (1) an application programming interface (API), (2) an integrated metadata, provenance, and ontology (MPO) data store, and (3) a suitable user interface (UI) for displaying, interacting, and navigating with the data.

2.1. Application programming interface

Today, most applications use scripting languages, such as Python, to express workflows. Although workflow systems have been developed over the last few years (e.g. Kepler, Swift), these require installation of the system and support for the language they are written in. Furthermore, they require learning a new system, how to express the workflow's structure, and how to invoke activity components. Our goal is to allow application users to continue with the tools they are accustomed to working with (i.e. scripting languages) augmented with capabilities for annotating the scripts. This allows the structure to be viewed externally, as well as allowing for commands that can automatically write actions to a database that can later be queried to extract provenance. Any extensible workflow engine will be able to take advantage of the API but the near-term focus is what scientists are using today, scripts.

The research problem associated with this task is the development of the annotation capabilities [4]. This includes two parts, the first being the annotation of the scripts at the beginning and end of the code that can be considered a single activity step of a workflow (Fig. 1). This annotation must also include information on what are the input/output data objects that resulted from previous steps. This will allow a definition of the workflows as data flows, since dependencies are based on availability of data objects. The second part is the annotations to write provenance data when the scripts are run. This includes information on input data objects for each workflow step, output of product data objects, and the activity involved. These will generate links between an instance of a workflow run and the corresponding activities in the workflow database. In addition, these annotations will include directives to allow users to provide additional information on each activity or an entire run to capture user comments on what was the purpose of a run and how parameters were chosen. An important capability that is planned is to provide the ability to capture an ensemble of runs where, for example, the same workflow is run multiple times with different parameters.

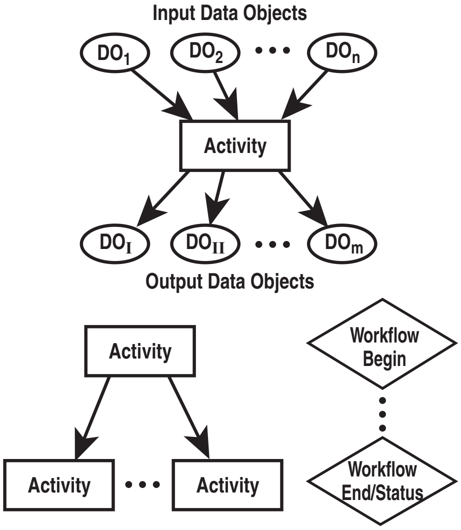

Fig. 1. Depicted are three different types of workflow structures that can be represented by the MPO system.

Since one of our goals is better integration between experimental and simulation data, tools will be needed for both. This goal is common to many application domains where simulations are validated by comparing the results with experimental data. As mentioned, this research uses as an initial realistic example existing tools from the fusion domain. For experiments, our work takes advantage of the workflow functionality that already exists within MDSplus. For simulations, several test cases will be created where the scripts used to prepare data and execute the simulation code can be appropriately instrumented. A realistic environment ensures that the challenge of development of a useful and realistic annotation language will be met. Such annotation is an innovative approach for the automation of capturing workflows embedded in scripts, and automating provenance collection when running scripts.

The API research uses Representational State Transfer technology (RESTful) interfaces and references to data object as Unique Identification Numbers (UID). The reasons for these choices are as follows. Using a RESTful interface is a choice of using a data centric API instead of a procedural one. It simplifies the implementation of provenance tracking by reducing the number of procedures that need to be instrumented. The workflow may be viewed as the evolution of the state or a sub-state of the database. The data or states are accessed through GET (select) procedures and changed through PUT (update) procedures. Including POST (insert) and DELETE methods, only four procedures need to be instrumented to track workflow and provenance. These procedures all act on a route with a UID that represents a data instance of a particular type of resource, or in the case of POST, create the UID of a resource. As a result, the complexity has been moved from the API into the data or state representation. The resources exposed via the REST architecture may be qualified with parameters in the URL in the case of GET, as "&param1=value&param2=value" attached to the resource