GACP 20003321

Automated Metadata, Provenance Cataloging and Navigable Interfaces: Ensuring the Usefulness of Extreme-Scale Data

Multi-institutional Collaborative Project
Lead Principal Investigator:
David P. Schissel
General Atomics
3550 General Atomics Court, San Diego, CA 92121
Schissel@fusion.gat.com

Other Principal Investigators:
Dr. Martin Greenwald
MIT Plasma Science and Fusion Center
77 Massachusetts Avenue, NW17, Cambridge, MA 02139
g@psfc.mit.edu

Dr. Arie Shoshani
Lawrence Berkeley National Laboratory
1 Cyclotron Road, Berkeley, CA 94720
shoshani@lbl.gov

PROJECT SUMMARY/ABSTRACT

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 our ability to make use of it. The goal of this research is to create a data model, infrastructure, and a set of tools that support data tracking, cataloging, and integration across a broad scientific domain. Our system would document workflow and data provenance in the widest sense, enabling us to answer the questions "who, what, when, how and why" for each data element, provide information about the connections and dependences between the data elements, and allow human or automatic annotation for any data element. We aim to capture information from the creation, recording or importing of physical data through various levels of analysis, data preparation, data staging, HPCC code execution, storage, post processing, data exporting and publication. These tools would be demonstrated in large national and international fusion sciences collaborations – from which user experience would be collected and lessons-learned tabulated to provide feedback for improved design. We believe that rapid prototyping and testing by real users on real problems at scale is crucial. Solutions to "toy" problems do not provide the opportunity for real feedback. While using Fusion Energy Sciences as a test bed, our conceptual framework and data model will be quite general and not contain specific references to the fusion domain. We expect that what we develop 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. Our work will in effect create a modern "scientific notebook" for computational science. Similar common needs exist for experimental data and all fields of science struggle to integrate information from simulation and experiment and to extract knowledge from the confrontation between the two.