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

D.P. Schissel[1], Gheni Abla[1], S. Flanagan[1], M. Greenwald[2], X. Lee[1], A. Romosan[3], A. Shoshani[3], J. Stillerman[2], J. Wright[2]

[1]General Atomics, P.O. Box 85608, San Diego, CA, USA
[2]Massachusetts Institute of Technology, Cambridge, MA, USA
[3]Lawrence Berkeley National Laboratory, Berkeley, CA, USA
email: schissel@fusion.gat.com, Phone: (858) 455-3387, Fax: (858) 455-4156

For scientific research, it is not the mere existence of experimental or simulation data that is important, but our ability to make use of it. This paper follows previous work [1] and 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. Our system is intended to 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 dependencies between the data elements; and allow human or automatic annotation for any data element.

Combining research on integrated metadata, provenance, and ontology (MPO) information with research on user interfaces, including graphical navigation has allowed the construction of early prototype tools for the fusion science domain. While using Fusion Energy Sciences as a test bed, our conceptual framework and data model is quite general and does not contain specific references to the fusion domain. Although the equations solved by simulations are different for each diverse field of science, the basic flow of information, the need to document workflow and provenance, and the need to allow traceability of results is common to all. Similar common needs exist for experimental data.

Results of three main research elements will be presented along with a prototype system. The first investigates the primitives and language for annotation required to automatically document provenance data. The second element investigates the best approaches and technologies for integrating metadata, provenance, and workflow documentation. This includes methods for representing graphs, ontologies and taxonomy. The third element examines user interfaces, including graphical navigation. The research is investigating the best methods for displaying, navigating, and interacting with the MPO system. A critical element is to provide users with the tools to interactively explore the MPO relationships.

[1] M. Greenwald, et al., "A Metadata Catologue for Organization and Systemization of Fusion Simulation Data," Fusion Eng. & Design **87** (2012) 2205-2208.

TOPICS: Advanced Computing and Massive Data Analysis

Preference: Oral
Journal Publication: Yes

*This work was supported by the US Department of Energy under DE-SC0008697.