An Unmet Need: Documenting Complex Scientific Workflows – End to End

Martin Greenwald (MIT)*, David Schissel (GA), and John Wright (MIT)
*Corresponding Author: g@psfc.mit.edu
Request Oral Presentation: No
Primary Panel: F, Secondary Panel: G

Summary: The goal is to increase the productivity and impact of fusion research carried out via simulation by better management of data and documentation of the scientific workflow.

Introduction: Data from simulations and experiments are the common currency that underlies scientific research. However, data are useful only to the extent that their meaning can be conveyed and preserved. Before the advent of electronic computing, the scientific notebook captured on paper all of the data along with descriptions of assumptions and processes that went into capturing or calculating the data and summarized the research results. The advent of modern computing has allowed us to generate and store vastly more data, but has fragmented or destroyed the records of how the data was produced or used. New tools are required to restore the previous record keeping capabilities in a new environment with vastly different technologies.

Use Cases: The unmet requirements for data management can be understood by considering several practical use cases that might face researchers.

a) A recently graduated PhD student left behind output from thousands of gyrokinetic simulations. Which of these were used in her thesis? Which might be useful in the future? What were the assumptions, inputs and parameters used in the interesting runs?

b) How did a researcher arrive at the data plotted in figure 6 of their 2014 Phys. Plasmas article? What set of simulations was used? Which experimental shots were used as input?

c) A calibration error was found in Thomson Scattering data taken during 2011. The data has now been recalculated, but the original data was likely used as input to important simulations. In which runs was the old data used? What publications used that data? Were they critical for the published conclusions? Was any of that data contributed to an ITPA database?

d) A researcher is designing a diagnostic to validate large-scale simulations. He wants to use a range of simulation data, processed through a synthetic diagnostic, to help choose the characteristics and view for the physical diagnostic. He can select simulations of interest by searching a metadata catalog, find out which are considered "good" by the person who ran them and couple their own workflow to the ones used to create the runs.

Building toward a solution: The first steps toward addressing this problem are underway. The MPO (Metadata, Provenance and Ontology) project is aimed at documenting scientific workflow and is carried out by a team of researchers and software engineers from GA, LBNL and MIT. **Metadata:** The MPO stores descriptive metadata that allows users to search and browse information about the data they've produced and the processes that were used to create that data. The MPO is built on structural metadata that represents the relations between objects contained in the system. **Provenance:** The MPO provides users with a well-structured mechanism to define the origins of any piece of data that they have stored. The users control the completeness and

granularity of that description. **Ontology:** In our context, ontology refers to a formal system of names that allows users to more efficiently and consistently define metadata

Project Objectives: The overall objective of the MPO project is to help preserve the meaning of stored data by documenting all of the steps that were taken to produce that data, that is the data provenance. The capture of process information and metadata along with links to the data also supports reproducibility – that is the ability to repeat any experiment or calculation. At the same time, the MPO supports more systematic management of analysis and simulation data that is often not as carefully handled or archived as experimental data.

Implementation Philosophy: To be useful and to be adopted by researchers, the system developed by the MPO project is designed to intrude as little as possible on users practices and tools. Adoption of the MPO does not require that users adopt a particular, monolithic tool to control their workflow. Rather it allows them to instrument their existing set of tools to the extent that they find useful, recording as much or as little information as they need. Once set up, the MPO is meant to work as automatically as possible (and so best suited for scripted rather than one-time use). The MPO is intended to support all types of scientific workflows including both experimental and computational or any mixture. In most fields, the broader research effort involves processing of raw data, with small or large codes that often provide inputs to larger simulations. Code output usually requires extensive processing as well.

Looking Forward: The MPO system has recently gone from beta software to a full production release with installations at MIT and GA. Initial beta users instrumented both experimental analysis and simulation workflows and provided valuable feedback that was folded into the production release. Additional usage in new workflows will help to further assess the usability, functionality, and performance of the MPO for automated workflow documentation. From a practical standpoint, further development of the MPO software will be limited since the project's funding expires at the end of FY15. But whether new work picks up where the MPO left off or uses the lessons learned and starts anew, it is clear that further development and support of a production system will be critical to the successful long-term usage of the large amount of simulation data that is anticipated to be produced. This need is especially true given the recent work to couple typically distinct workflows under one umbrella for faster turnaround (e.g. OMFIT); this will only accelerate the production of data.

Conclusion: Lacking in today's workflow ecosystem is end-to-end data management and provenance capture. The motivation to fill this gap is to greatly increase the value of experimental and computational fusion data. Therefore, this white paper puts forth the argument that a fundamental requirement of future simulation research is the creation of a comprehensive data management capability that spans the entire range of scientific activities including both experimental and simulation activities. There exists a good body of work to build on but resources will need to be dedicated for a community-wide production system.