Scientific Workflows: Catalyzing the Grid ⇌ Semantic Web Reaction

Bertram Ludäscher

University of California, Davis

Scientific workflows allow scientists to automate repetitive data management, analysis, and visualization tasks, and to document the provenance of analysis results. Scientific workflows are composed of interlinked computational components (sometimes called *actors*), and the datasets that are consumed and produced by those components. Scientific workflow systems are problem-solving environments to design, reuse, share, execute, monitor, and archive scientific workflows. As such, they are the primary tool that end user scientists use when interacting with the emerging e-Science cyberinfrastucture. Scientific workflow systems can often benefit from both, Grid and Semantic Web capabilities. Thus, scientific workflows can bring together these otherwise loosely connected technologies and "catalyze the reaction" between them. For example, *compute-intensive* workflows (e.g. simulation models for quantum chemistry, protein folding, ecological niche modeling, weather forecasting, ocean currents, supernova explosions, etc.) require significant computational power available through Grid and/or dedicated cluster resources, and thus can benefit from Grid services for distributed and parallel execution and scheduling. Scientific workflows can also be *data-intensive* (consuming and producing large amounts of data, thus suggesting the use of Data-Grid middleware), and *metadata-intensive*, i.e., requiring and producing information on the provenance of datasets and the meaning of data, in particular, to facilitate reproducability of experiments (e.g. by capturing essential parameters of experiment protocols) and for reuse of workflows and analysis products. Metadata-intensive workflows often employ Semantic Web standards such as RDF and OWL to capture metadata in a machine-processable form. Taken together, resource management provided by Grid services and knowledge capture and management through Semantic Web technologies, provide essential capabilities of any general purpose, large-scale scientific workflow systems. In various ongoing projects we employ these technologies to enhance the Kepler scientific workflow system and make it more versatile for the scientist and more interoperable with other e-Science/cyberinfrastructure tools and services.

See Also:

http://kepler-project.org and
B. Ludäscher, et al. Scientific Workflow Management and the Kepler System. *Concurrency and Computation: Practice & Experience*, 2005, to appear.

Dagstuhl Seminar Proceedings 05271
Semantic Grid: The Convergence of Technologies
http://drops.dagstuhl.de/opus/volltexte/2006/407