

## NSF Workshop on the Challenges of Scientific Workflows

<http://www.isi.edu/nsf-workflows06>

May 1-2, 2006

Arlington, VA

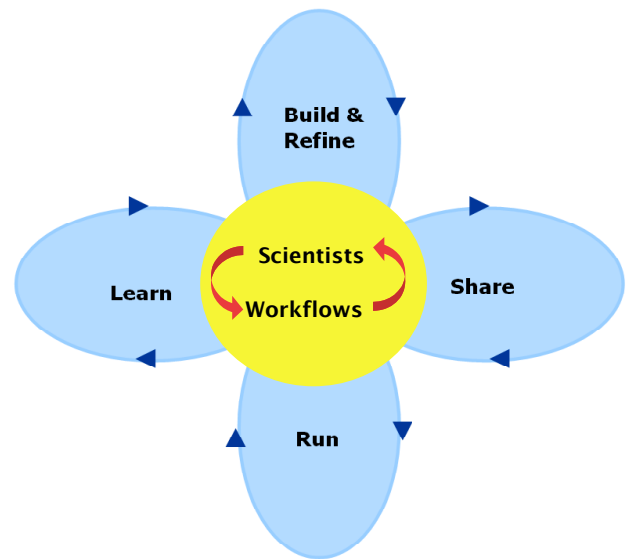
Sponsored by the National Science Foundation

**Ewa Deelman and Yolanda Gil (co-chairs)**

USC Information Sciences Institute  
University of Southern California

**Maria Zemankova (Program Director)**

Division of Information and Intelligent Systems  
National Science Foundation



### Executive Summary From the Final Workshop Report

Significant scientific advances today are achieved through complex distributed scientific computations. These computations, often represented as workflows of executable jobs and their associated data flow, may be composed of thousands of steps that integrate diverse models and data sources. Assembling and managing such workflows presents many challenges, and the limits of current technology are continuously pushed by increasingly ambitious scientific inquiry.

The recent Workshop on Challenges of Scientific Workflows was held at the National Science Foundation on May 1-2, 2006. The meeting brought together domain scientists, computer scientists, and social scientists, to discuss the requirements of future scientific applications and the challenges that they present to current workflow technologies.

Domain scientists consider workflows as a crucial and underrepresented ingredient in Cyberinfrastructure. Given the exponential growth in compute, sensors, data storage, network, and other performance elements, why is the growth of science not exponential? In part this may be due to the increasing complexity of managing the computations and data. Workflow systems reduce this complexity with assistance or automation to create, execute, and manage distributed computations and resulting data. Domain scientists also expressed concern that because most such computations are manually created and managed, the repeatability of the scientific process becomes nearly unattainable. Repeatability is a cornerstone of the scientific method, and as the complexity of scientific computations grows it will become impossible to replicate their results unless there is an economical and precise way to characterize and reproduce computations. Workflow systems can support repeatability very efficiently, since they track in detail the provenance of every data product in terms of the computations that generated it. Workflow systems can easily reuse the underlying computation to reproduce the results with alternative data. Workflow systems can also support the creation of variants of the workflow to support scientific exploration processes by facilitating and tracking their exploration trails. Therefore, workflow systems are an enabling factor to accelerate scientific progress.

Current workflow systems are able to manage quite complex computations that include thousands of components, using dozens of data repositories, and harnessing resources in dozens of sites. However, these applications are structurally simple compared with new emerging requirements from scientists to handle

streaming data, accommodate interactive steering, support event-driven analysis, and enable their creation through collaborative design processes involving many scientists across disciplines.

Computer scientists consider workflows as an enabler to automate and manage complex distributed computations. Workflows provide a formal and declarative representation of complex scientific processes that can then be managed efficiently through their lifecycle from assembly, to execution, to sharing. Much has been invested in scientific data repositories, instruments, and resource sharing, but the areas of workflow representation, execution management, and sharing are largely unexplored. While workflow systems can address some these issues in limited ways, current techniques are unlikely to adequately address the challenges of future scientific workflows in terms of their complexity, scope, heterogeneity, interactivity, collaborative nature, and execution management. Addressing these challenges will span diverse areas of computer science research, including distributed computing, artificial intelligence, software engineering, programming languages, semantic web, and collaborative software. Additional expertise will be needed from other areas of science, such as cognitive science, human computer interfaces, and operations research.

The following recommendations were made by the workshop participants:

- Basic research in computer science is needed to create a science of workflows.
- Explicit workflow representations that capture scientific analysis processes at all levels should become the norm when complex distributed scientific computations are carried out.
- Workflow representations should be integrated with other forms of scientific record.
- Cross-disciplinary projects need to be supported and encouraged involving relevant areas of computer science as well as domain sciences with distinct requirements and challenges.
- Long-term, stable (5 years and greater) collaborations and programs will be required.
- A roadmap to advance the research agenda of scientific workflows while building on existing Cyberinfrastructure should be elaborated.
- Coordination of workflow projects and interoperation frameworks for workflow tools will be needed among existing and new projects.
- Follow-up, cross-cutting workshops and meetings should be encouraged.

In summary, **workflows should become first-class entities in Cyberinfrastructure architecture.** For domain scientists, they are important because workflows document and manage the increasingly complex processes involved in exploration and discovery through computations. For computer scientists, workflows provide a formal and declarative representation of complex distributed computations that must be managed efficiently through their lifecycle from assembly, to execution, to sharing.

**Workshop Attendees:** Mark Ackerman, University of Michigan • Ilkay Altintas, San Diego Supercomputing Center • Roger Barga, Microsoft • Francisco Curbera, IBM • Mark Ellisman, University of California San Diego • Constantinos Evangelinos, MIT • Thomas Fahringer, University of Innsbruck • Juliana Freire, University of Utah • Ian Foster, University of Chicago & Argonne National Laboratory • Geoffrey Fox, Indiana University • Dennis Gannon, Indiana University • Carole Goble, University of Manchester • Alexander Gray, Georgia Institute of Technology • Jeffrey Grethe, University of California San Diego • Jim Hendler, University of Maryland • Carl Kesselman, USC Information Sciences Institute • Craig Knoblock, USC Information Sciences Institute • Chuck Koelbel, Rice University • Miron Livny, University of Wisconsin • Luc Moreau, University of Southampton • Jim Myers, National Center for Supercomputing Applications (NCSA) • Karen Myers, SRI International • Walt Scacchi, University of California Irvine • Ashish Sharma, Ohio State University • Amit Sheth, University of Georgia Athens • Alex Szalay, John Hopkins University • Gregor Von Laszewski, Argonne National Laboratory