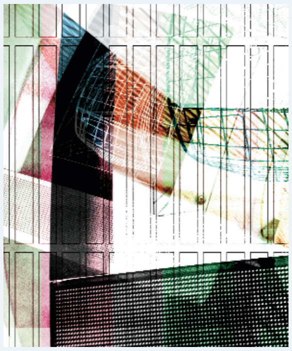
# Examining the Challenges of Scientific Workflows

**Workflows have emerged as a paradigm for representing and managing complex distributed computations and are used to accelerate the pace of scientific progress. A recent National Science Foundation workshop brought together domain, computer, and social scientists to discuss requirements of future scientific applications and the challenges they present to current workflow technologies.**



*Yolanda Gil and Ewa Deelman*
University of Southern California

*Mark Ellisman*
University of California, San Diego

*Thomas Fahringer*
University of Innsbruck

*Geoffrey Fox and Dennis Gannon*
*Indiana University*

*Carole Goble*
Manchester University

*Miron Livny*
University of Wisconsin-Madison

*Luc Moreau*
University of Southampton

*Jim Myers*
National Center for Supercomputing Applications

Significant scientific advances are increasingly achieved through complex sets of computations and data analyses. These computations may comprise thousands of steps, where each step might integrate diverse models and data sources that different groups develop. The applications and data might be also distributed in the execution environment. The assembly and management of such complex distributed computations present many challenges, and increasingly ambitious scientific inquiry is continuously pushing the limits of current technology.

Workflows have recently emerged as a paradigm for representing and managing complex distributed scientific computations, accelerating the pace of scientific progress.[1-6] Scientific workflows orchestrate the dataflow across the individual data transformations and analysis steps, as well as the mechanisms to execute them in a distributed environment.

Each step in a workflow specifies a process or computation to be executed (for instance, a software program or Web service). The workflow links the steps according to the data flow and dependencies among them. The representation of these computational workflows contains many details required to carry out each analysis step, including the use of specific execution and storage resources in distributed environments. Figure 1 shows an example of a high-level workflow developed within the context of an earthquake science application, CyberShake (www.scec.org), which generates shake maps of Southern California.[7]

Workflow systems exploit these explicit representations of complex computational processes at various levels of abstraction to manage their life cycle and automate their execution. In addition to automation, workflows can provide the information necessary for scientific reproducibility, result derivation, and result sharing among collaborators. By providing automation and enabling reproducibility, they can accelerate and transform the scientific-analysis process.

Workflow systems have demonstrated these capabilities in a variety of applications where workflows comprising thousands of components processed large,

Published by the IEEE Computer Society

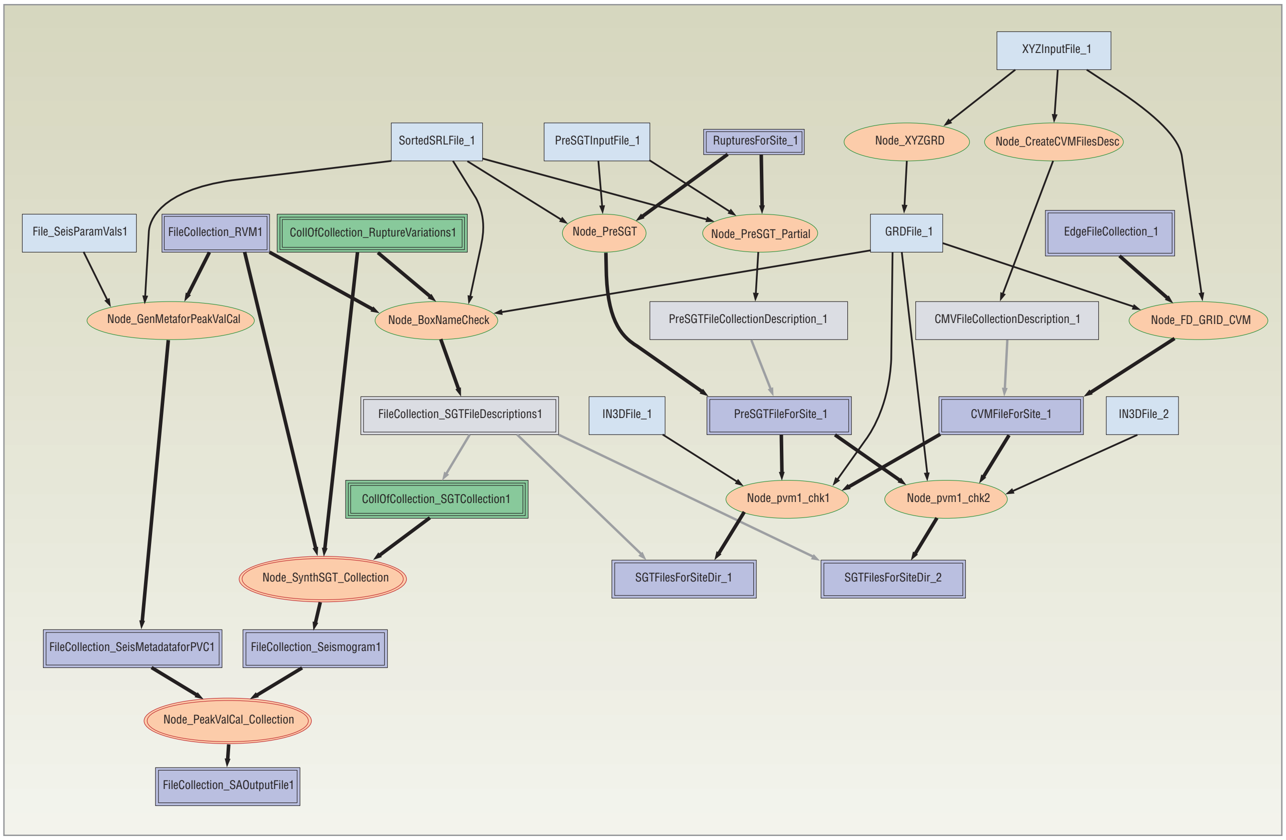


*Figure 1. A visual representation of a high-level workflow developed within the context of an earthquake science application. Double-lined nodes indicate computations that the system will parallelize automatically.*

distributed data sets on high-end computing resources. Some workflow systems are deployed for routine use in scientific *collaboratories*—virtual entities that allow scientists to collaborate with each other across organizations and physical locations. Figure 2 shows an image of the Orion Nebula that the Montage[8] application produced. Montage uses workflow technologies[9] to generate science-grade mosaics of the sky. Researchers recently used such mosaics to verify a bar in the M31 galaxy.[10]

Much research is under way to address issues of creation, reuse, provenance tracking, performance optimization, and reliability. However, to fully realize the promise of workflow technologies, we must meet many additional requirements and challenges. Scientific applications are driving workflow systems to examine issues such as supporting dynamic event-driven analyses, handling streaming data, accommodating interaction with users, intelligent assistance and collaborative support for workflow design, and enabling result sharing across collaborations.

As a result, we need a more comprehensive treatment of workflows to meet the long-term requirements of scientific applications. The National Science Foundation's 2006 Workshop on Challenges of Scientific Workflows brought together domain, computer, and social scientists to discuss requirements of future scientific applications and the challenges they present to current workflow technologies. As part of the workshop, we examined application requirements, workflow representations, dynamic workflows, and system-related challenges.

*Figure 2. The Montage application uses workflow technologies to generate science-grade mosaics of the sky.*

## APPLICATION REQUIREMENTS

Given the exponential growth in computing, sensors, data storage, network, and other performance elements, why is the growth of scientific data analysis and understanding not proportional?

### Collaborations

Combining distributed data, computation, models, and instruments at unprecedented scales can enable transformative research. The analysis of large amounts of widely distributed data is becoming commonplace. This data, and the experimental apparatus or simulation systems that produce it, typically belong to collaborations rather than individuals. Within these collaborations, various individuals are responsible for different aspects of data acquisition, processing, and analysis, and entire projects often generate publications. Such environments demand tools that can orchestrate the steps of scientific discovery and bridge the differing expertise of collaboration members.

**Many disciplines benefit from the use of workflow-management systems to automate computational activities.**

Many disciplines benefit from the use of workflow-management systems to automate such computational activities, including astronomy, biology, chemistry, environmental science, engineering, geosciences, medicine, physics, and social sciences.

The scientific community perceives that workflows are important in accelerating the pace of scientific discoveries. Today, complex scientific analyses increasingly require tremendous amounts of human effort and manual coordination. Thus, researchers need more effective tools to prevent being inundated by the ever-growing data and associated computational processing tasks.

### Reproducibility

The NSF workshop participants identified reproducibility of scientific analyses and processes as an important application requirement. Reproducibility is at the core of the scientific method, enabling scientists to evaluate the validity of each other's hypotheses and providing the basis for establishing known truths. Reproducibility requires rich provenance information so researchers can repeat techniques and analysis methods to obtain scientifically similar results.

Today, reproducibility for complex scientific applications is virtually impossible. Many scientists are involved, and the provenance records are highly fragmented, existing in e-mails, wiki entries, database queries, journal references, codes, and other sources for communication. All this information, often stored in various locations and forms, must be appropriately indexed and made available for referencing. Without tracking and integrating these crucial bits of information with the analysis results, reproducing important discoveries involving complex computations can be impractical or even impossible.

To support reproducibility, workflow-management systems must capture and generate provenance information as a critical part of the workflow-generated data. Workflow-management systems must also consume the provenance information associated with input data and associate that information with the resulting data products. Systems must associate and store provenance with the new data products and contain enough details to enable reproducibility.

Scientists also need interoperable, persistent repositories of data and analysis definitions, with linkage to open data and publications, as well as to the algorithms and applications used to transform the data. Workflow systems must complement existing data repositories with provenance and metadata repositories that enable the discovery of the workflows and application components used to create the data. Two important concerns for scientists in these highly collaborative endeavors are credit assignment and recognition of individual contributions.

### Flexible environments

Systems must be flexible in terms of supporting both common analyses that many scientists perform, as well as unique analyses. Researchers should find it easy to set up and execute routine analyses based on common cases. At the same time, individual scientists should be able to steer the system to conduct unique analyses and create novel workflows with previously unseen combinations and configurations of models.

From an operational perspective, there's a need to provide secure, reliable, and scalable solutions. Scientists must trust that their input and output data is secure and free from inappropriate data access or malicious manipulation. Current infrastructure must incorporate trust and reputation systems for data providers.

Finally, scientists need easy-to-use tools that provide intelligent assistance for such complex workflow capabilities. Automation of low-level operational aspects of workflows is a key requirement. Success will depend on interaction modalities that hide unnecessary complexities and speak the scientist's language.

## SHARED WORKFLOW DESCRIPTIONS

Given the broad practice and benefits of sharing instruments, data, computing, networking, and many other science products and resources, why don't researchers widely capture and share scientific computations and processes as well?

## Process sharing

Scientists have always relied on technology to share information about experiments, from pen and paper to digital cameras, e-mail, the Web, and computer software. Workflow description and execution capabilities offer a new way of sharing and managing information to electronically capture full processes and share them for future reference and reuse.

This new way of sharing information—agreeing on processes' semantics and the infrastructure to support their execution—continues the historic push for making representations explicit and actionable and reducing the barriers to coordination. We should encourage scientists to bring workflow representations to their practices and share the descriptions of their scientific analyses and computations in ways that are as formal and explicit as possible. However, no commonly accepted and sufficiently rich representations exist in the scientific community.

**Workflow representations must incorporate rich information about analysis processes to support discovery, creation, merging, and execution.**

## Representations

Workflow representations must accommodate scientific process descriptions at multiple levels. For instance, domain scientists might want a sophisticated graphical interface for composing relatively high-level scientific or mathematical steps, whereas the use of a workflow language and detailed specifications of data movement and job execution steps might concern computer scientists.

To link these views and provide needed capabilities, workflow representations must include rich descriptions that span abstraction levels and include models of how to map between them. Further, to support the end-to-end description of multidisciplinary, community-scale research, we need definitions of workflow and provenance that are broad enough to describe workflows-of-workflows that are linked through reference data, the scientific literature, and manual processes in general.

Other important and necessary dimensions of abstraction are experiment-critical versus non-experiment-critical representations, where the former refers to scientific issues and the latter is more concerned with operational matters.

## Abstractions

Workflow representations must incorporate rich information about analysis processes to support discovery, creation, merging, and execution. These activities will become a natural way to conduct experiments and share scientific methodology within and across scientific communities.

**Automation.** Wherever possible, workflow representations need to support automation of the workflow creation and management processes. This capability requires rich semantic representations of requirements and constraints on workflow models and components. With semantic descriptions of the data format and type requirements of a component, it's possible to incorporate automated reasoning and planning capabilities that could automatically add data conversion and transformation steps. Similarly, rich descriptions of the execution requirements of each workflow component would enable automated resource selection and dynamic optimizations.

**Levels of description.** Abstractions would let scientists identify what levels of description are useful to share in their workflows, and they could package such descriptions as a self-contained sharable object that other scientists could then refine and instantiate. We need refinement and abstraction capabilities for all first-class entities that workflow systems must manipulate: workflow scripts (regarded as specifications of future execution), provenance logs (descriptions of process and data history), data, and metadata. There's relevant work in related fields of computer science, such as refinement calculi, model-driven architectures, and semantic modeling, but researchers haven't applied these techniques widely to scientific workflows, which are potentially large scale, might involve multiple technologies, and must operate on heterogeneous systems.

The sophistication of required descriptions depends on the workflow capabilities needed. For example, a workflow that adapts dynamically to changes in environment or data values requires formal and comprehensive descriptions to enable automatic adaptation. Even for a human to make choices related to making changes to a workflow would require access to a broad variety of descriptions.

## Scientific versus business workflows

Understanding the differences between scientific workflows and practices and those used in business could yield useful insights. On the one hand, scientific and business workflows aren't obviously distinguishable, since both might share common important characteristics. Indeed, the literature contains examples of workflows in both domains that are data-intensive and highly parallel. On the other hand, scientific research requires flexible design and exploration capabilities that appear to depart significantly from the more prescriptive use of workflows in business. Workflows in science are a means to support detailed scientific discourse as well as a way to ensure repeatable processes.

Another distinctive issue of scientific workflows is the variety and heterogeneity of data within a single workflow. For example, a scientific workflow might involve numeric and experimental data in proprietary formats (such as those used for raw data that scientific instruments involved in a process produce), followed by processed data resulting in descriptions related to scientific elements, leading to textual, semistructured, and structured data, and formats used for visual representation.

To clarify the research issues in developing scientific workflow capabilities, the community needs to identify where there are real differences between scientific and business activities, beyond domain-specific matters. It's important to balance the desire for sharing workflow information against the dangers of premature standardization efforts that might constrain future requirements and capabilities.

### Workflow variants

Most scientific activity consists of exploration of variants and experimentation with alternative settings, which would involve modifying workflows to understand their effects and provide a means for explaining those effects. Hence, an important challenge in science is representation of workflow variants, which aims at understanding the impact that a change has on the resulting data products as an aid to scientific discourse.

**The management of dynamic workflows is complex due to their evolution and life cycles.**

While acknowledging that sharing representations is important to the scientific process, the workshop group recognized that workflows must accommodate multiple collaboration and sharing practices. In some cases, it's suitable to share workflows, but not data. In other cases, scientists want to share an abstract description of the scientific protocol without actually communicating details, parameters, and configurations, which are their private expertise. In other situations, a description of a specific previous execution (provenance) is desirable, with or without providing execution details.

## DYNAMIC WORKFLOWS

How can workflows support both the exploratory nature of science and the dynamic processes involved in scientific analysis?

### Changing context and infrastructure

Given that both the user's experimental context and the distributed infrastructure that the workflows operate over are in flux, the notion of static workflows is an odd one. The vision of supporting dynamic, adaptive, and user-steered workflows is to enable and accelerate distributed and collaborative scientific methodology via rapid reuse and exploration accompanied by continuous adaptation and improvement. Reproducibility becomes ever more elusive in this kind of setting. The challenge is to develop mechanisms to create, manage, and capture dynamic workflows to allow for reproducibility of significant results.

Scientific practice will routinely give rise to dynamic workflows that base decisions about subsequent steps on the latest available information. Researchers might need to dynamically design a workflow to look at the initial steps' results before making a decision on carrying out later analysis steps. For example, examining the results of an image's initial preprocessing might require subsequent steps to look at specific areas that preprocessing identified.

### External events

A dynamic workflow could also result from an external event changing the workflow's basic structure or semantics. For example, in severe-storm prediction, data-analysis computations might search for patterns in radar data. Depending upon the specific pattern of events, enacting different branches of a storm-prediction workflow might require significant computational resources on-demand.

In this case, the workflow must adapt to changes in storm intensity or resource availability. Some experimental regimens might draw on workflows that are heuristic or employ untried activities, thus these workflows might break down or fail during their execution, necessitating fault diagnosis and repair. Two workflows could also affect each other by sharing results, being classified as dynamic as they respond to events arising in each other's execution.

Finally, some scientific endeavors are large scale. They involve large teams of scientists and technicians, and they engage in experimental methods or procedures that take a long time to complete and require human intervention and dynamic steering throughout the process. For example, an astrophysical study of deep-space phenomena might require the use and coordination of multiple observation devices operating in different spaces, capturing data at different frequencies or modalities.

### Workflow life cycle

The management of dynamic workflows is complex due to their evolution and life cycles. As Figure 3 shows, there's no beginning or end to the life-cycle process of a workflow—scientists can start at any point and flow through the figure in any direction. They might build or assemble a workflow, refine one that a shared repository has previously published, run their design, evolve it, run it again, share fragments of it as they go along, find other fragments they need, run it a few more times, and learn from the protocol they're developing.

They might settle on the workflow and run it many times, learning from the results produced, or they might run it just once because that's all they need. While running, the workflows could adapt to external events and user steering. The results of the whole activity feed into the next phases of investigation. The user is ultimately at the center, interacting with the workflows and interpreting the outcomes.

Supporting scientists in complex exploratory processes involving dynamic workflows is an important challenge. Researchers will need to design a human-centered decision-support system that accommodates the information needs of a scientist tracking and understanding such complex processes. The workflow will need appropriate user interfaces that enable scientists to browse/traverse, query, recapitulate, and understand this information. Simplifying the exploratory process also requires novel and scalable means for scientists to manipulate the workflows, explore slices of the parameter space, and compare the results of different configurations.

### Learning workflow patterns

An interesting direction for future research explores the question of how to improve, redesign, or optimize workflows through data mining of workflow life-cycle histories to learn successful (and unsuccessful) workflow patterns and designs, and assist users in following (or avoiding) them. Researchers can extract one kind of pattern from successful execution trails and use the information to build recommendation systems. For example, if a model *M* is added, the system could suggest additional models that other people often use together with *M* in a workflow, or suggest values commonly used for the parameters in the model. Researchers could extract another kind of pattern from unsuccessful trails. These patterns can, for example, help identify incompatible parameter settings, unreliable servers or services, or gross inefficiencies in resource usage. Researchers can subsequently analyze, reenact (reproduce), and validate workflow patterns in order to facilitate their reuse, continuous improvement, and redeployment into new locations or settings.

## SYSTEM-LEVEL WORKFLOW MANAGEMENT

Given the continuous evolution of infrastructure and associated technology, how can we ensure reproducibility of computational analyses over a long period of time?

### Engineering reproducibility

A key challenge in scientific workflows is ensuring engineering reproducibility to enable the reexecution of analyses, and the replication of results. Scientific reproducibility implies that someone can follow the general methodology, relying on the same initial data, and obtain equivalent results. Engineering reproducibility requires more knowledge of the data manipulations, of the actual software and execution environment (hardware, specific libraries), to replicate the results bit-by-bit. Researchers need the former capability when they want to validate each other's hypotheses, whereas the latter is beneficial when they find unusual results or errors and need to trace and understand them.

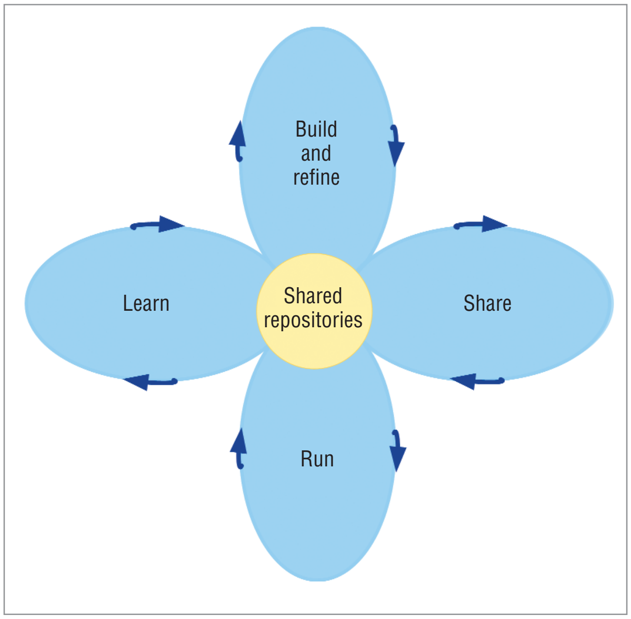


*Figure 3. A view of the workflow life cycle, where the processes of workflow generation, sharing, running, and learning are continuous.*

### System stability

Providing a stable view in spite of continuous technology and platform changes at the system level will be challenging. Researchers must design the underlying execution system to provide a stable environment for the software layers managing the high-level scientific process. It must be possible to reexecute workflows many years later and obtain the same results. This requirement poses challenges in terms of creating a stable layer of abstraction over a rapidly evolving infrastructure while providing the flexibility needed to address evolving requirements and applications and to support new capabilities.

To provide consistent and efficient access to resources, resource management must consider both physical resources (computers, networks, and data servers) and logical resources (data repositories, programs, application components, and workflows). Uniform interfaces should inform both. Enhancing resource descriptions with semantic annotations can enable easier, more organized, and possibly automated provisioning, provenance, configuration, and deployment of new resources. Extending current information services with meaningful semantic descriptions of resources should allow for semiautomatic discovery, brokering, and negotiation.

Dynamic configuration and life-cycle management of resources should minimize human interaction. Researchers have made some efforts to provide semiautomatic discovery and brokering of physical resources and management of software components that might become part of scientific-workflow environments. However, there's still much opportunity for improvements, since most existing systems require manual or semimanual deployment of software components and force application builders to hard-code software component locations on specific resources into their workflows.

## Quality of service

Workflow end users frequently need to specify quality-of-service (QoS) requirements. The underlying runtime environment should then guarantee, or at least maintain, these requirements on a best-effort basis. However, current systems are mostly restricted to best-effort optimizations for time-based criteria such as reducing overall execution time or maximizing bandwidth. Researchers must address several problems to overcome current limitations.

**Issues of scale will increasingly require advances over the state of the art, and they occur in multiple dimensions.**

We must extend QoS parameters beyond time-based criteria to cover other important aspects of workflow behavior such as responsiveness, fault tolerance, security, and costs. To provide a basis for interoperable workflow environments or services, this effort will require collaborative work on the definition of QoS parameters that scientists can widely accept.

Coping with multicriteria optimization or planning might require radically changing current optimization and planning approaches. Many systems exist for single or bi-criteria optimization, but few systems tackle multicriteria optimization problems. There's no ready-to-use methodology that can deal with this problem in an efficient and effective way; thus, many opportunities for research exist.

Reservation mechanisms will be an important tool in developing runtime environment support for QoS. Both immediate and advance reservations can make the dynamic behavior of infrastructures more predictable, an important prerequisite to guarantee QoS parameters such as responsiveness and dependability. Moreover, advance reservation can also simplify the scheduling of workflow tasks to resources.

## Scaling

Challenging issues of scale arise in workflow execution. These issues will increasingly require advances over the state of the art, and they occur in multiple dimensions.

First, in many disciplines, individual workflows are becoming large as the quantities of data operated on become larger. As workflows scale from 1,000 to 10,000 and perhaps 1 million or more tasks, researchers might need new techniques to represent sets of tasks, manage those tasks, dispatch tasks efficiently to resources, monitor task execution, detect and deal with failures, and so on.

A second important scaling dimension is the number of workflows. Particularly in large communities, many users might submit many workflows at once. If these workflows compete for resources or otherwise interact, the runtime environment needs appropriate supporting mechanisms to arbitrate among competing demands.

A third scaling dimension concerns the number of resources involved. Ultimately, we can imagine workflows running on millions of data and computing resources (indeed, some systems such as SETI@home already operate at that scale). A fourth scaling dimension concerns the number of participants. In a simple case, a single user prepares and submits a workflow. In a more complex case, many participants might help define the workflow, contributing relevant data, managing its execution, and interpreting results.

We need to provide new infrastructure services to support workflow management. Some of these services are analogous to existing data management and information services, such as workflow repositories and registries. Other novel services will be concerned with workflows as active processes and the management of their execution state.

## Infrastructure constraints

There's a perceived tension between workflow research challenges and the constraints that existing production-quality infrastructures impose. Shared infrastructures such as the Open Science Grid (www.opensciencegrid.org), the TeraGrid (www.teragrid.org), and NMI (www.nsf-middleware.org) provide widely used and well-tested capabilities to build on. These system-level infrastructure layers are designed to be production quality, but out of necessity haven't been designed to address workflows' specific requirements. Rather, they aim to meet a broader research community's needs. It's unlikely that we can make commitments by selecting particular architectures or implementations at the workflow layers of shared cyberinfrastructure. We must explore alternative architectures to understand design tradeoffs in different contexts. Examples include

- workflows designed and tested on a desktop and run with larger data in a cluster,
- workflows to handle streaming data,

- event-driven workflow-management engines, and
- architectures centered on interactivity.

At the same time, we could design these architectures to be interoperable and compatible, where feasible, with some overall end-to-end, multilevel framework. Follow-on discussions and workshops to understand and address these issues will be extremely beneficial.

## RECOMMENDATIONS

Workflows provide a formal specification of the scientific-analysis process from data collection, through analysis, to data publication. We can view workflows as recipes for cyberinfrastructure computations, providing a representation describing the end-to-end processes involved in carrying out heterogeneous interdependent distributed computations.

Once scientists capture this process in declarative workflow structures, they can use workflow-management tools to accelerate the rate of scientific progress by creating, merging, executing, and reusing these processes. By assisting scientists in reusing well-known and common practices for analyses, complex computations will become a daily commodity for use in scientific discovery. As scientists conduct experiments in neighboring disciplines, cross-disciplinary scientific analyses will become commonplace.

The NSF workshop participants made the following recommendations:

- Support basic research in computer science to create a science of workflows.
- Make explicit workflow representations that capture scientific analysis processes at all levels the norm when performing complex distributed scientific computations.
- Integrate workflow representations with other forms of scientific record.
- Support and encourage cross-disciplinary projects involving relevant areas of computer science as well as domain sciences with distinct requirements and challenges.
- Provide long-term, stable collaborations and programs.
- Define a road map to advance the research agenda of scientific workflows while building on existing cyberinfrastructure.
- Coordinate between existing and new projects on workflow systems and interoperation frameworks for workflow tools.
- Hold follow-up, cross-cutting workshops and meetings and encourage discussions between subdisciplines of computer science.

Scientists view workflows as key enablers for reproducibility of experiments involving large-scale computations. Reproducibility is ingrained in the scientific method, and there's concern that without this ability, scientists will reject cyberinfrastructure as a legitimate means for conducting experiments. Representing scientific processes with enough fidelity and flexibility will be a key challenge for the research community. Recognizing that science has an exploratory and evolutionary nature, workflows need to support dynamic and interactive behavior. Thus, workflow systems need to become more dynamic and amenable to steering by users and be more responsive to changes in the environment.

Workflows should become first-class entities in the cyberinfrastructure architecture. For domain scientists, they're important because workflows document and manage the increasingly complex processes involved in exploration and discovery through computation. For computer scientists, workflows provide a formal and declarative representation of complex distributed computations that must be managed efficiently through their life cycle from assembly, to execution, to sharing. ■

**Acknowledgments**

The NSF sponsored this workshop under grant #IIS-0629361. We thank Maria Zemankova, program manager of the Information and Intelligent Systems Division, for supporting the workshop and contributing to the discussions. We also thank the workshop attendees (http://vtcpc.isi.edu/wiki/index.php/Participants) for their contributions.

### References

1. E. Deelman and Y. Gil, eds. *Final Report of NSF Workshop on Challenges of Scientific Workflows*, Nat'l Science Foundation; http://vtcpc.isi.edu/wiki/images/b/b2/NSFWorkshopFlyer-final.pdf.
2. E.Deelman and I. Taylor, eds., *J. Grid Computing*, special issue on scientific workflows, vol. 3, no. 3-4, Sept. 2005.
3. E. Deelman, Z. Zhao, and A. Belloum, eds., *Scientific Programming J.*, special issue on workflows to support large-scale science, vol. 14, no. 3-4, 2006.
4. G. Fox and D. Gannon, eds., *Concurrency and Computation: Practice and Experience*, special issue on workflow in grid systems, vol. 18, no. 10, Aug. 2006.
5. B. Ludaescher and C. Goble, eds., *SIGMOD Record*, special issue on scientific workflows, vol. 34, no. 3, Sept. 2005; www.sigmod.org/record/issues/0509/index.html.
6. I.J. Taylor et al., eds., *Workflows for e-Science: Scientific Workflows for Grids*, Springer-Verlag, 2006.
7. Y. Gil et al., "Wings for Pegasus: Creating Large-Scale Scientific Applications Using Semantic Representations of Computational Workflows," *Proc. Conf. Innovative Applications*

*of Artificial Intelligence* (IAAI), 2007, pp. 1767-1774; http://dblp.uni-trier.de/rec/bibtex/conf/aaai/GilRDMK07.

8. G.B. Berriman et al., “Montage: A Grid-Enabled Engine for Delivering Custom Science-Grade Mosaics On Demand,” *Proc. SPIE,* vol. 5493, SPIE, 2004, pp. 221-234.
9. E. Deelman et al., “Pegasus: A Framework for Mapping Complex Scientific Workflows onto Distributed Systems,” *Scientific Programming J.*, vol. 13, 2005, pp. 219-237; http://vtcpc.isi.edu/wiki/images/f/f4/Pegasus.doc.
10. R.L. Beaton et al., “Unveiling the Boxy Bulge and Bar of the Andromeda Spiral Galaxy,” *Astrophysical J. Letters* (in submission).

***Yolanda Gil*** *is the  associate division director for research of the Intelligent Systems Division at the University of Southern California’s Information Sciences Institute (ISI) and a research associate professor in the Computer Science Department. She cochaired the NSF’s Workshop on Challenges of Scientific Workflows. Her research interests include intelligent interfaces for knowledge-rich problem solving. Gil received a PhD in computer science from Carnegie Mellon University. She is a member of the Association for the Advancement of Artificial Intelligence. Contact her at gil@isi.edu.*

***Ewa Deelman*** *is a project leader in the Advanced Systems Division at USC’s ISI and cochaired the NSF’s Workshop on Challenges of Scientific Workflows. Her main research interest is scientific workflow management in distributed environments. Deelman received a PhD in computer science from Rensselaer Polytechnic Institute. She is a member of the IEEE Computer Society. Contact her at deelman@isi.edu.*

***Mark Ellisman*** *is the director of the Center for Research in Biological Systems and National Center for Microscopy and Imaging Research and a professor of neurosciences and bioengineering at the University of California, San Diego. His research interests include the molecular and cellular basis of nervous system function as well as the use of advanced imaging and information technologies in brain research. Ellisman received a PhD in molecular, cellular, and developmental biology from the University of Colorado, Boulder. He is a founding fellow of the American Institute of Medical and Biological Engineering. Contact him at mellisman@ucsd.edu.*

***Thomas Fahringer*** *is a professor and head of the Distributed and Parallel Systems Group at the Institute of Computer Science at the University of Innsbruck, Austria. His research interests include distributed and parallel systems. Fahringer received a PhD in computer science from the Technical University of Vienna. He is a member of the IEEE and the ACM. Contact him at tf@dps.uibk.ac.at.*

***Geoffrey Fox*** *is a professor of physics and computer science at Indiana University and a distinguished scientist in its Community Grids Laboratory. His research interests include grids and parallel computing. Fox received a PhD in theoretical physics from Cambridge University. He is a member of the ACM and the IEEE Computer Society and a fellow of the American Physical Society. Contact him at gcf@indiana.edu.*

***Dennis Gannon*** *is a professor of computer science in the School of Informatics at Indiana University. His research interests include cyberinfrastructure, programming systems and tools, and distributed computing. Gannon received a PhD in mathematics from the University of California, Davis, and a PhD in computer science from the University of Illinois. Contact him at gannon@cs.indiana.edu.*

***Carole Goble*** *is a professor at the University of Manchester, and director of the $^{my}$GRID project. Her research interests are the Semantic Web, e-science, and grid communities. Goble received a BSc from Manchester University. She is a member of the IEEE, the ACM, and the British Computer Society. Contact her at carole.goble@manchester.ac.uk.*

***Miron Livny*** *is a computer science professor at the University of Wisconsin-Madison. His research interests include high-throughput computing, visual data exploration, and experiment-management environments. Livny received a PhD in computer science from the Weizmann Institute of Science, Israel. Contact him at miron@cs.wisc.edu.*

***Luc Moreau*** *is a professor of computer science at the University of Southampton. His research interests include large-scale open distributed systems and provenance. He received a PhD from the University of Liège, Belgium. He is a fellow of the British Computer Society and a member of the ACM. Contact him at L.Moreau@ecs.soton.ac.uk.*

***Jim Myers*** *leads the Cyberenvironments and Technologies Directorate at the National Center for Supercomputing Applications. His research interests include open source collaborative tools. Myers received a PhD in chemistry from the University of California, Berkeley. He is a member of the American Chemical Society, the American Physical Society, the ACM, and the IEEE. Contact him at jimmyers@ncsa.uiuc.edu.*