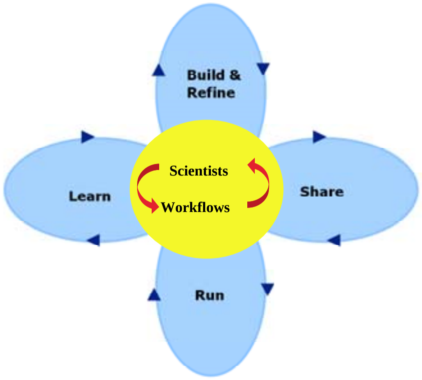

Workshop on the Challenges of Scientific Workflows

May 1-2, 2006

Arlington, VA

Sponsored by the National Science Foundation

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

Ewa Deelman and Yolanda Gil, co-chairs

Information Sciences Institute

University of Southern California

October 16, 2006

Workshop Organizers

Ewa Deelman, USC Information Sciences Institute, Marina Del Rey, CA (co-Chair)
Yolanda Gil, USC Information Sciences Institute, Marina Del Rey, CA (co-Chair)
Maria Zemankova, National Science Foundation, Washington, DC (Program Director)

Workshop Participants

Mark Ackerman, University of Michigan, Ann Arbor, MI
Ilkay Altintas, San Diego Supercomputing Center, San Diego, CA
Roger Barga, Microsoft, Redmond, WA
Francisco Curbera, IBM, Yorktown, NY
Mark Ellisman, University of California San Diego, San Diego, CA
Constantinos Evangelinos, MIT, Boston, MA
Thomas Fahringer, University of Innsbruck, Innsbruck, Austria
Juliana Freire, University of Utah, Salt Lake City, Utah
Ian Foster, University of Chicago & Argonne National Laboratory, Chicago, IL
Geoffrey Fox, Indiana University, Bloomington, IN
Dennis Gannon, Indiana University, Bloomington, IN
Carole Goble, University of Manchester, Manchester, UK
Alexander Gray, Georgia Institute of Technology, Atlanta, GA
Jeffrey Grethe, University of California San Diego, San Diego, CA
Jim Hendler, University of Maryland, College Park, MD
Carl Kesselman, USC Information Sciences Institute, Marina Del Rey, CA
Craig Knoblock, USC Information Sciences Institute, Marina Del Rey, CA
Chuck Koelbel, Rice University, Houston, TX
Miron Livny, University of Wisconsin, Madison, WI
Luc Moreau, University of Southampton, Southampton, UK
Jim Myers, National Center for Supercomputing Applications (NCSA), Urbana, IL
Karen Myers, SRI International, Menlo Park, CA
Walt Scacchi, University of California Irvine, Irvine, CA
Ashish Sharma, Ohio State University, Columbus, OH
Amit Sheth, University of Georgia, Athens, GA
Alex Szalay, John Hopkins University, Baltimore, MD
Gregor Von Laszewski, Argonne National Laboratory, Chicago, IL

Table of Contents

Executive Summary

Significant scientific advances are increasingly achieved through complex sets of computations and data analyses. These computations, often represented as workflows of executable jobs and associated data flows, may comprise thousands of steps. Each step may integrate diverse models and data sources, which may be developed by different groups. The applications and data may be also distributed in the execution environment. The assembly and management of such workflows present many challenges, and increasingly ambitious scientific inquiry is continuously pushing the limits of current technology. Today's workflow systems are able to manage quite complex computations that include thousands of components, use dozens of data repositories, and harness resources at 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.

To examine the nature of these challenges and to consider what steps should be taken to address them, a Workshop on the 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 requirements of future scientific applications and the challenges that they present to current workflow technologies.

This report summarizes the discussions and recommendations of the workshop. In this summary, we present two major findings and seven recommendations developed by workshop participants.

Domain scientists consider workflow as a crucial and underrepresented ingredient in Cyberinfrastructure. Domain scientists participating in the workshop pointed to the marked disparity between tremendous growth in the performance of computers, sensors, data storage, networks, and other system elements, and the decidedly slower growth in scientific insight. They asserted that this disparity is due, in part, to the increasing complexity of managing ever larger and more distributed computations and data. Domain scientists also expressed concern that because sequences of computational activities are typically performed manually, repeatability of scientific processes—the cornerstone of the scientific method—becomes nearly unattainable. These two concerns both point to a need for workflow systems that can assist with and/or automate the creation, execution, and management of sets of computations and the data that those computations produce. Workflow systems must also provide an efficient and precise way to characterize and reproduce computations. Such systems should be able to support repeatability, by tracking in detail the provenance of every data product in terms of the computations that generated it and input data used. They should allow for the convenient re-execution of a computational process with alternative data. Workflow systems can also support scientific exploration by facilitating and tracking the creation of alternative workflows. For all these reasons, workshop participants saw workflow systems as being an increasingly important foundation for scientific progress.

Computer scientists consider workflows as an enabler to automate and manage complex distributed computations. Computer scientists participating in the workshop argued that workflows can play a valuable role in scientific work by providing a formal and declarative representation of complex scientific processes that can then be managed efficiently through their lifecycle from assembly to execution and sharing. However, they agreed that important issues crucial to the more widespread application of workflow tools, such as workflow representation, execution management, and sharing, are largely unexplored. While existing workflow systems can address some of 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. Tackling these challenges will require contributions from 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 domain of science as well, such as cognitive science, human computer interfaces, operations research, and possibly others.

The following recommendations were made by the workshop participants:

- **Support basic research in computer science to create a science of workflows**. Although existing systems are addressing important issues such as workflow creation, planning, and execution, more comprehensive research is needed to provide easy-to-use workflow construction tools, develop sophisticated automation tools, provide robust workflow execution, manage complex dynamic workflows, etc. There are many open research issues to be resolved in computer science proper that will enable significant progress in the research agenda of scientific workflows.
- **Make explicit workflow representations that capture scientific analysis processes at all levels the norm when performing complex distributed scientific computations.** We need workflow representations at different levels of abstraction, so that we can represent workflows at different levels of refinement, from abstract application-level definition down to operational, system-specific description. These workflow representations can become a starting point for defining common representations that can be interpreted by a variety of workflow systems.
- **Integrate workflow representations with other forms of scientific record.** Data created through workflows should include representations of those workflows as metadata. Articles in scientific publications should include not only textual descriptions of the processes utilized, but also formal descriptions specified as workflows. Laboratory notebooks and invention records should be annotated with workflows and the rationale for their design and final configuration.
- **Support and encourage cross-disciplinary projects involving relevant areas of computer science as well as domain sciences with distinct requirements and challenges.** Cross-disciplinary projects between computer scientists and application scientists are needed to ensure that research efforts are directed towards areas where they can have a significant impact. Other disciplines, such as social sciences and cognitive science should also be engaged to meet the stated challenges.
- **Provide long-term, stable (five or more years) collaborations and programs.** Based on the experiences of the NSF ITR program and the UK e-Science program, the greatest successes were obtained in collaborations that were funded for five years or more, so that collaborations had time to mature and obtain significant results.
- **Define a roadmap to advance the research agenda of scientific workflows while building on existing cyberinfrastructure.** Significant investment in cyberinfrastructure, has resulted in production quality services for data management, high-end and large-scale computation, resource sharing, and distributed computing. It will be important to articulate anticipated requirements to support scientific workflows in the coming years, and develop a roadmap for how the current infrastructure can evolve to accommodate the challenging research agenda that lies ahead. A follow-up workshop on this topic in the near future would be highly beneficial.
- **Coordinate between existing and new projects on workflow systems and interoperation frameworks for workflow tools**. Many current projects have evolved in isolation, working with non-intersecting scientific communities. Capturing best practices will enable a better understanding of the existing capabilities. It will be beneficial to consider the development of a common framework, so that various workflow tools can be integrated and interchanged with others. Scientists will then be able to concentrate on the science rather than have to worry about the particulars of different workflow systems.
- **Hold follow-up, cross-cutting workshops and meetings**. More workshops are needed, to bring together scientists from various domains. Encourage discussion between sub-disciplines of computer science, to and bring in human factors and collaboration considerations to workflow management systems.

Introduction

The vision for a Cyberinfrastructure for science put forward by the National Science Board[1] and the National Science Foundation[2] is beginning to be realized. A variety of distributed resources are managed and shared by the broad scientific community in the form of hardware platforms and associated software infrastructure[3]. These efforts are laying a foundation for large-scale, high-performance scientific applications that routinely analyze Terabytes of data in many scientific domains. This cyberinfrastructure is enabling a significant paradigm shift in science, whereby the end-to-end scientific discovery process encompasses data collection from shared instruments and hypothesis formation through computationally-intensive analysis to the publication and dissemination of the resulting data products. Throughout this process, data may undergo many transformations at a variety of distributed locations and may be analyzed and processed by a number of collaborators. Today, such **complex distributed computational processes are enabling transformative research in many scientific communities**. Those communities include astronomy, physics, ecology, earthquake science, and many others[4]. This is a global trend in science and information technology research[5].

In most scientific disciplines, these complex computational processes are often generated, managed, and described with manual and ad-hoc approaches. Such approaches are problematic in practice, for several reasons. First, manually creating and managing these processes is time-consuming and error prone—indeed, this may be impractical as data size and complexity increases. Second, it is difficult to capture all steps in the data analysis process and their characteristics, including both high-level scientific properties and low-level execution environment features, and without this information, it is difficult to re-generate the results. Reproducibility of results is a cornerstone of the scientific method, and without it, large-scale computational science may soon be questioned in terms of its scientific soundness and integrity.

Workflows have recently emerged as a paradigm for representing and managing complex distributed scientific computations and therefore accelerate the pace of scientific progress. Scientific workflows capture the individual data transformations and analysis steps as well as the mechanisms to carry them out in a distributed environment. Each step in the workflow specifies a process or computation to be executed (e.g., a software program to be executed, a web service to be invoked). The steps are linked according to the data flow and dependencies among them. The representation of these computational workflows contain many details required to carry out each

[1] "National Science Board 2020 Vision for the NSF". National Science Board Report NSB 05-142. December 2005. Available at: http://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsb05142.

"Science and Engineering Infrastructure Report for the 21st Century: The Role of the National Science Foundation". National Science Board Report NSB 02-190. February 2003. Available at: http://www.nsf.gov/nsb/documents/2002/nsb02190/nsb02190.pdf.

[2] "Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure". January 2003. Available at: http://www.nsf.gov/publications/pub_summ.jsp?ods_key=cise051203.

"NSF's Cyberinfrastructure Vision for 21st Century Discovery". NSF Cyberinfrastructure Council. National Science Foundation Report, Version 7.1, July 20, 2006. Available at: http://www.nsf.gov/dir/index.jsp?org=OCI.

[3] http://www.teragrid.org, http://www.osg.org, and http://www.nsf-middleware.org.

[4] See for example the National Virtual Observatory (http://www.us-vo.org), the Grid Physics Network (http://www.griphyn.org), the Science Environment for Ecological Knowledge (http://seek.ecoinformatics.org), and the Southern California Earthquake Center (http://www.scec.org).

[5] UK e-Science Programme, http://www.rcuk.ac.uk/escience.

analysis step, including the use of specific execution and storage resources in distributed environments, Workflow systems can exploit these explicit representations of the complex computational processes to manage their lifecycle and to automate their execution. Workflows can capture complex analysis processes at various levels of abstraction, and also provide the provenance information necessary for scientific reproducibility, result publication, and result sharing among collaborators. By providing formalism and by supporting automation, workflows have the potential to accelerate and transform the scientific analysis process. Existing workflow systems have been demonstrated in a variety of scientific applications, were workflows composed of thousands of components processed large-distributed data sets on high-end computing resources. Some workflow systems have been deployed for routine use in scientific collaboratories. Appendix A provides more background and references on existing work on scientific workflows. Much research is underway to address issues of creation, reuse, provenance tracking, performance optimization, and reliability. However, to fully realize the promise of workflow technologies, many additional requirements and challenges must be met. 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, a more comprehensive treatment of workflows is needed to meet long-term requirements of scientific applications.

This report summarizes the discussions and recommendations of the recent Workshop on the Challenges of Scientific Workflows. The goal of this workshop was to examine workflow challenges from a variety of perspectives. To this end, the workshop brought together computer science researchers and practitioners working on various aspects of workflow management, as well as of domain scientists that use workflows for day-to-day data analysis and simulation. Application scientists were asked to describe requirements, desired new analyses, and computations that are not possible with today’s technologies. Computer science researchers were asked to identify challenges in their specific areas of expertise. The workshop discussions focused on four main topics:

> **TOPIC 1: Applications and requirements**: What are the requirements of future applications? What new capabilities are needed to support emerging applications?
> **TOPIC 2: Data and workflow descriptions**: How can workflow descriptions be improved to support usability and scalability? How to describe data produced as part of workflows? What provenance information must be tracked to support scalable data and workflow discovery?
> **TOPIC 3: Dynamic workflows and user steering**: What are the challenges in supporting dynamic workflows that evolve over time as data become available? What techniques can support incremental and dynamic workflow evolution due to user steering?
> **TOPIC 4: System-level management**: What are the challenges in supporting large-scale workflows scalably and robustly? What changes are needed in existing software infrastructure? What new research needs to be done to develop better workflow management systems?

The following sections summarize discussions and recommendations related to each of these four topics. The final section provides conclusions and recommendations of the workshop attendees. A short overview and pointers to the state of the art in the area of scientific workflows is provided in the Appendix, including citations and references to relevant work.

Discussion Topic 1: Application Requirements

A key motivating question posed by domain scientists participating in the workshop was:

> ***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?***

There was a broad consensus in the group that in the scientific community there is a perceived importance of workflows in accelerating the pace of scientific discoveries. Today, complex scientific analyses increasingly require tremendous amounts of human effort and manual coordination. Data is growing exponentially, but the number of scientists is roughly constant. Thus researchers need exponentially more effective tools to aid in their work, if they are not to be inundated in data and associated tasks. Workflow environments that support and improve the scientific process at all levels are crucial if we are to sustain the current rapid growth rate in data and processing.

The ability to combine 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. These data, and the experimental apparatus or simulation systems that produce them, typically do not belong to individuals but rather to collaborations. Within these collaborations, various individuals are responsible for different aspects of data acquisition, processing, and analysis, and in which publications are often generated by entire projects. Such environments demand tools that can orchestrate the steps of scientific discovery and bridge between the differing expertises of the members of the collaboration.

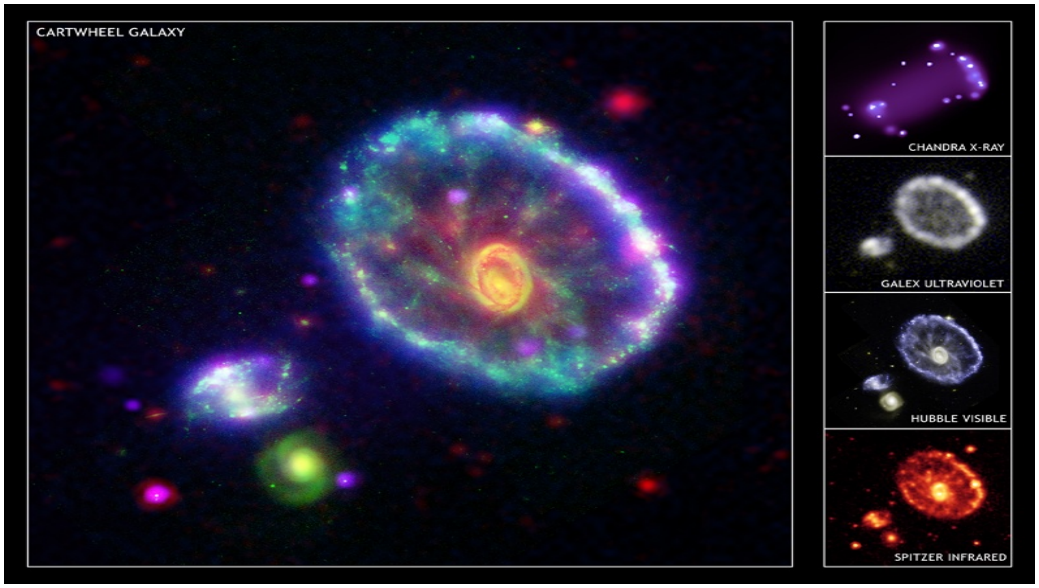

Figure 1: An example of transformative research through multi-perspective computations: a composite image of the Cartwheel galaxy integrating different light spectra from a variety of research collaborations (from http://antwrp.gsfc.nasa.gov/apod/ap060118.html).

An example of today's complex scientific applications is the analysis of astronomy data across multiple spectral bands. For example, Figure 1 shows the Cartwheel galaxy, created as the result of a collision of two galaxies. The composite image in the left panel was produced by a computation pipeline that mosaics images from different light spectra. The input data for this computation were obtained by different researchers in different collaborations; the ability to combine those data sources allows us to produce new results, such as that shown here, that can yield new insights. This example underscores the possibility of scientific discovery based on combining and processing information from a variety of sources. Many disciplines are benefiting from the use of workflow management systems to automate such computational activities. Examples of such disciplines include astronomy, biology, chemistry, environmental science, engineering, geosciences, medicine, physics, and social sciences.

An important application requirement identified by workshop participants is **reproducibility of scientific analyses and processes**. This requirement is at the core of the scientific method, in that it enables scientists to evaluate the validity of each other's hypothesis and provides the basis for establishing known truths. Reproducibility requires rich *provenance* information, so that researchers can repeat techniques and analysis methods to obtain scientifically similar results. Today, reproducibility is virtually impossible for complex scientific applications. First, because so many scientists are involved, the provenance records are highly fragmented, and in practice they are reflected in a variety of elements including emails, Wiki entries, database queries, journal references, codes (including compiler options), and others. All this information, often stored in a variety of locations and in a variety of forms, needs to be appropriately indexed and made available for referencing. Without tracking and integrating these crucial bits of information together with the analysis results, reproducibility can be largely impractical, and more likely impossible, for many important discoveries involving complex computations.

In order 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. Provenance must be associated and stored with the new data products and contain enough details to enable reproducibility. Another important requirement is for 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. Existing data repositories must be complemented with provenance and metadata repositories that enable the discovery of the workflows and application components that were used to create the data. An important concern for scientists in these highly collaborative endeavors is credit assignment and recognition of individual contributions. Novel approaches to track and annotate originators and consumer of workflows, models, analyses, hypotheses, and other important ingredients of cross-disciplinary collaborations need to be incorporated into the general infrastructure. Achieving these various goals in a manner that is accurate, efficient, and compatible with existing information systems is likely to be extremely challenging.

The environments provided should also be flexible in terms of **supporting both common analyses performed by many as well as unique individual analyses**. Routine analyses based on common cases should be easy to set up and execute. At the same time, individual scientists should be able to steer the system to conduct unique analyses and to create novel workflows with previously unseen combinations and configurations of models.

Many scientific applications require **interactive and dynamic workflow environments** in which users (or sensors, or other systems) can see preliminary results in a timely manner and reconfigure the remainder of a workflow based on results obtained. Providing this capability is especially challenging in the case of applications that require real-time interaction and response.

From an operational perspective, there is a need to provide **solutions that are secure, reliable, and scalable**. Scientists need to be able to trust that their input and output data are secure and free from inappropriate data access or malicious manipulation. Trust and reputation systems for data providers must be incorporated into current infrastructure. Tools need to be scalable in order to support large and complex analyses, TeraByte and greater size data sets, and large scientific communities.

An important concern is how to **address the inevitable heterogeneities and inconsistencies that arise when information comes from different sources and communities**. Mechanisms for curating, validating, translating, and integrating data are needed in order for scientific information to be shared in meaningful and truly integrative ways.

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. Interaction modalities that hide unnecessary complexities and speak the scientist's language will be crucial to success. Guidance to users will be useful to encourage the best scientific practices.

The principal conclusions of this discussion group are:

- **Scientific workflows have the potential of significantly accelerating the rate of scientific progress**. Workflows can serve as the "recipes" for cyberinfrastructure computations that automate scientific analysis processes. A significant investment in workflow systems may exponentially accelerate scientific progress and allow scientists to cope with the exponential growth of compute, sensors, storage, and network performance.
- **Scientific workflows are critical to reproducibility, which is the main ingredient of the scientific method.** Reproducibility is becoming increasingly impossible given the distributed and complex nature of many scientific computations. Scientific workflows can capture the information needed to enable reproducibility and repeatability of results in the form of provenance of data products detailing the computations and data sources involved.
- **Workflow management systems should integrate with many other sources of information** routinely used by scientists, such as publications, laboratory workbooks, discussion forums (Wikis, emails, etc). This integration is needed to capture in an appropriate level of detail the trail of decisions and reasoning behind specific experimental settings and to enable the sharing and the interpretation of results.
- **A close and continued collaboration between domain scientists and computer scientists will be required** in order to develop and articulate the requirements specific to scientific workflows, and to develop an appropriate and timely research agenda to address the many challenges.
- **Sharing workflows is an essential element of education, and acceleration of knowledge dissemination.** Thus we need to provide workflow libraries, descriptions and structures that enable sharing among a number of individuals and groups.

Discussion Topic 2: Data and Workflow Descriptions

A key issue addressed by this discussion group was:

> ***Given the broad practice and many benefits of sharing instruments, data, computing, networking, and many other science products and resources, why are scientific computations and processes not widely captured and shared as well?***

Scientists have always relied on technology to share information about experiments, from pen and paper to digital cameras, email, the Web, and computer software. Workflow description and execution capabilities offer a new way of sharing and managing information: one in which full processes can be captured electronically and shared for future reference and reuse. This new way of sharing information—agreeing on semantics of processes themselves and the infrastructure to support their execution—continues the historic push for making representations explicit and actionable, and reducing the barriers to coordination. **Scientists should be encouraged to bring workflow representations to their practices and share the descriptions of their scientific analyses and computations in**

ways that are as formal and as explicit as possible. However, there are no commonly accepted and sufficiently rich representations in the scientific community. Thus, more research in this area is needed.

Workflow representations need to **accommodate scientific process descriptions at multiple levels**. For instance, domain scientists may want a sophisticated graphical interface for composing relatively high-level scientific or mathematical steps, whereas computer scientists may be more concerned with the use of a workflow language, and with the detailed specifications of data movement and job execution steps. To link between these views and to provide needed capabilities, workflow representations must include rich descriptions that span abstraction levels, and must include models of how to map between them.. Further, to support the end-to-end description of multidisciplinary, community-scale research, **definitions of workflow and provenance must be broad enough to describe workflows-of-workflows that are linked through reference data, models backed by validation workflows, the scientific literature, and manual processes in general.** Other important dimensions of abstraction are experiment-critical vs. non-experiment-critical representations, where the former refers to scientific issues and the latter is more concerned with operational matters. A workflow system should support both sets of concerns.

Rich information about analysis processes needs to be incorporated in workflow representations to **support workflow discovery, creation, merging, and execution**. These activities will become a natural way to conduct experiments and share scientific methodology within and across scientific communities.

Workflow representations need to **support, wherever possible, automation of the workflow creation and management processes**. This capability will require 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 is possible to incorporate automated reasoning and planning capabilities that could automatically add data conversion and transformation steps. Similarly, with rich descriptions of the execution requirements of each workflow component, automated resource selection and dynamic optimizations would be possible.

A challenge for the computer science community is to be able to **manipulate the multiple levels of workflow abstraction simultaneously and to manipulate them individually**. For instance, several distinguishable levels of process abstraction were considered useful in the breakout group: scientific, engineering, and instance. Another classification distinguished among data description, functional behavior specification, non-functional aspects, and execution/run-time aspects. A capability of "workflow abstraction" would allow scientists to identify what level(s) of description are useful to share in their workflows, and package such a description as a self-contained sharable object, which can then be refined and instantiated by other scientists. Refinement and abstraction capabilities are needed for all first-class entities that have to be manipulated by workflow systems: workflow scripts (regarded as specifications of future execution), provenance logs (descriptions of process and data history), data, and metadata. There is relevant work in related fields of computer science, such as refinement calculi, model-driven architectures, and semantic modeling, but these techniques have not been applied widely to scientific workflows, which are potentially large scale, may involve multiple technologies, and have to operate on heterogeneous systems. We also note that sophistication of descriptions needed is dependent on the workflow capabilities needed. For example, a workflow that adapts dynamically to changes in environment or data values requires formal and comprehensive descriptions so that a machine can make a decision on adaptation. Even for a human to make choices related to making changes to a workflow would require access to a broad variety of descriptions.

Another important research issue is **whether scientific workflows can or even should build on existing workflow technologies**, or whether they require fundamentally new approaches. Workflows have been used for decades to represent and manage business processes. There are emerging standards for workflow representations as well as associated software (some of commercial quality) to manage 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 are not obviously distinguishable, since both may share common important characteristics. Indeed, in the literature, we find examples of workflows in both domains that are data intensive, highly parallel, etc.

On the other hand, scientific research requires flexible design and exploration capabilities that appear to depart significantly from the more prescriptive use of workflow in business; **workflows in science are a means to support detailed scientific discourse as well as a way to enable repeatable processes**. Another distinctive issue of scientific workflows is the variety and heterogeneity of data within a single workflow. For example, scientific workflow may involve numeric and experimental data in proprietary formats (such as those used for raw data produced by the scientific instruments involved in a process), followed by processed data resulting in description related to scientific element (e.g., molecule or biochemistry descriptions), leading to textual, semi-structured, 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. An important concern is to balance the desire for sharing workflow information against the dangers of premature standardization efforts that may constrain future requirements and capabilities. In this respect, it will be crucial to encourage computer scientists and domain scientists to collaborate closely in developing more workflow-based applications and to discuss representation requirements for future workflows.

In the discussion, it was recognized that most scientific activity consists of exploration of variants and experimentation with alternative settings, which would involve **modifying workflows to understand their effects and how to explain 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. As part of managing change, version control becomes important. The challenge of evolving workflows is further compounded by the need to validate data products and to disseminate and share experiments and data within the scientific community. Hence, traceability and sharing are key requirements of scientific workflows.

While acknowledging that the sharing of representations is important to the scientific process, the group recognized that **multiple collaboration and sharing practices must be accommodated**. In some cases, it is 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, it is a description of a specific previous execution (provenance) that is desirable, with or without providing execution details.

The group identified several recommendations:

- Encourage the practice in the scientific community of **using explicit and formal representations that cover an increasing number of aspects of the scientific analysis process**. These representations should enable workflow management systems to manage not only single workflows but also collections of workflow variants, entire workflow libraries, and synthesis of workflows by drawing from several libraries. These representations need to provide valuable metadata to annotate workflow data products.
- **Workflow is a broad enough concept that multiple language variants and tools will be needed.** There will be no one workflow language or workflow system, as there is no one programming language or operating system. A spectrum of factors influence the adoption of a particular workflow language or tool: workflow language expressivity; workflow language abstractions; user skills; user training; open access to workflow tools; data and provenance management; and the existence of a user community willing to share workflows and good practice in workflow design. The real world is dynamic and heterogeneous. All research activities in the workflow area should design for change and design for flexibility and diversity and aim to interoperate. The scope and potential implications of any standards should be carefully assessed.
- **Continue to investigate the relevance of existing business process model representations and standards**, as well as commercial quality software, to understand the tradeoffs between building on existing technologies and developing a new infrastructure that will accommodate anticipated needs.

- Develop research in **workflow representations that will support refinements and abstractions** of multiple workflow specifications involved in workflow systems in order to support the overall scientific process.

Discussion Topic 3: Dynamic Workflows and User Steering

The participants in this discussion were tasked with examining issues related to the dynamic nature of the scientific analysis and focused on the following question:

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

Given that the experimental context of the user is in flux (as the scientific discovery process evolves) and the distributed infrastructure that the workflows operate over is in flux (as networks, platforms and other resources come and go), 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 and 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 so that reproducibility of significant results is possible.

Scientific practice will routinely give rise to workflows that are dynamic where the decisions they make about which steps to take next are based on the latest available information. A workflow may need to be dynamically designed in the sense of looking at the results of the initial steps before a decision can be made about how to carry out later analysis steps. For example, by examining the results of some initial pre-processing of an image subsequent steps may be needed to look at specific areas identified by that pre-processing. A dynamic workflow could also be one where the basic structure or semantics or the workflow changes because of some external event. For example, in severe storm prediction, data analysis agents may examine radar data searching for specific patterns. Depending upon the specific pattern of events, different branches of a storm prediction workflow may be enacted which may require that significant computational resource be made available on-demand. Should the storm intensify or should resource availability change, the workflow must adapt. Some experimental regimes may draw on workflows that are heuristic or that employ untried activities, and thus these workflows may breakdown or fail during their execution, thereby necessitating fault diagnosis and repair. Another scenario which includes dynamic workflows is where two workflows could affect each other, for example by sharing results. They can be 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 engage in experimental methods or procedures that take long times to complete and require human intervention and dynamic steering throughout the process. For example, the study of deep-space phenomena in astrophysical studies may require the use and coordination of multiple observation devices operating in different spaces, capturing data at different frequencies or modalities, and the resulting data will need to be cleaned and aligned for proper interpretation. Any step in such scientific inquiry may be subject to both the exigencies of sensor operation, weather or spatial occlusions during scheduled observation periods, and other delays, not to mention reactive adjustments to later stage observations arising from preliminary discoveries in earlier observational steps.

The **management of dynamic workflows is complex due to their evolution and lifecycle**. Figure 2 shows our view of what we termed the "FLOWer" lifecycle for the design and synthesis of distributed scientific workflows. There is no beginning or end to this process – scientists can start at any point and flow through the figure in any direction. They might build or assemble a workflow, refine one that has previously been published to a shared repository, 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 are developing. They might settle on the workflow and run it many times, learning from the results produced, or maybe they run it just once, because that is 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 centre, interacting with the workflows and interpreting the outcomes.

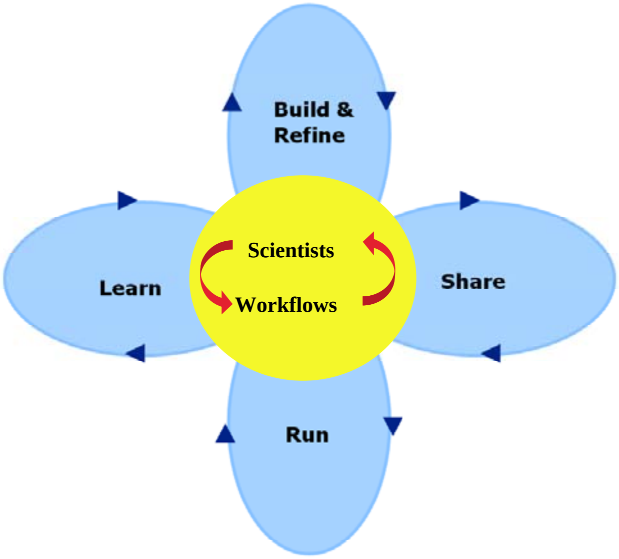

Figure 2: Life cycle for dynamic scientific workflows

Supporting scientists in complex exploratory processes involving dynamic workflows is an important challenge. A human-centered decision support system that accommodates the information needs of a scientist tracking and understanding such complex processes will need to be designed. Appropriate user interfaces that enable scientists to browse/traverse, query, re-capitulate, and understand this information will be needed. 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. Easily assembling workflows, finding services and adapting previous workflows is key.

An interesting direction for future research explores the question of how to **improve, redesign, or optimize workflows through data mining of workflow lifecycle histories** to learn successful (and unsuccessful) workflow patterns and designs and assist users to follow (or avoid) them. One kind of pattern can be extracted from successful execution trails. This information can be used 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. Another kind of pattern could be extracted from unsuccessful trails. These can, for example, help identify incompatible parameter settings, unreliable servers or services, gross inefficiencies in resource usage, etc. Workflow patterns can subsequently be analyzed, re-enacted (reproduced), and validated in order to facilitate their reuse, continuous improvement, and redeployment into new locations or settings.

The main recommendations from this discussion group are:

- **Workflow systems will need to integrate a variety of models in order for scientists to understand dynamic workflows.** These include models of collaboration, models of change, models of events, models of workflow execution, models of provenance, models of users, models of actions on change responding to events, and models of scientific experimentation and exploration. However, more than this, we need to understand how to integrate and inter-relate these models. They are not independent. Take the analogy of Integrative Systems Biology that includes models of the genome, protein-protein interaction, the cell, signalling pathways in the cell, metabolic pathway, the organism, disease, organisms in their environment etc. To support the dynamic scientific process, collaborations with cognitive scientists, human computer interface experts and others will be beneficial.

- **Dynamic workflows will require new computation paradigms that treat the workflow as a first-class object.** These paradigms should provide mechanisms to represent and manage the system's state in an evolving environment, control the process and avoid race conditions, integrate event-based and task-based computations, etc. For example, declarative workflow languages with pre-post step guards may be needed when external events cause a discontinuity in the workflow. It will be important to define what is a workflow at the different phases in the lifecycle.

- The role of **workflows in dynamic computational settings must be studied in collaborations between computer scientists and domain scientists**. Without real and extensive resource allocation, technical developments will focus on interesting technical problems and not user solutions. Long-term collaborations of five or more years are vital to the success of these investigations. There must also be a coordinated effort to share best practice and leverage solutions.

- **Dynamic workflows will need to support the exploratory nature of the scientific experimentation process**. As processes become more complex, scientists will benefit from appropriate user interfaces and interaction modalities that provide assistance in managing the process and understanding its results.

Discussion Topic 4: System-level Workflow Management

A key issue addressed by this discussion group was:

> ***Given the continuous evolution of infrastructure and associated technology, how can reproducibility of computational analyses be ensured over a long period of time?***

A key challenge in scientific workflows is **ensuring engineering reproducibility to enable the re-execution 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), etc., so that the results can be replicated bit-by-bit. The former capability is needed when researchers want to validate each other's hypotheses, whereas the latter is beneficial when unusual results or errors are found and their source needs to be traced and understood. The information needed to support both types of reproducibility is challenging to capture. When supporting scientific reproducibility, a high-level, yet meaningful, description of the process needs to be provided. Engineering reproducibility also necessitates low-level information such as what compiler flags were used to compile a particular code and the details of the execution environment and computer architecture.

An important challenge will be to **provide a stable view on the system in spite of continuous changes in technology and platforms at the system level**. The underlying execution system must be designed so that it provides a stable environment for the software layers managing the high-level scientific process. It must be possible to re-execute 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. In order to provide consistent and efficient access to resources, resource management must consider both physical resources (e.g., computers, networks, data servers) and logical resources (e.g., data repositories, programs, application components, workflows). Both should be exposed through uniform interfaces. By enhancing resource descriptions with semantic annotations, the provisioning, provenance, configuration and deployment of new resources can be organized more easily and possibly even automated. Extending current information services with meaningful semantic description of resources should enable semi-automatic discovery, brokering, and negotiation. Human interaction should be minimized through dynamic configuration and lifecycle management of resources. Some efforts have been made to provide semi-automatic discovery and brokering of physical resources and management of software components that may become part of scientific workflow environments. However, there is still much opportunity for improvements, since most existing systems require manual or semi-manual deployment of software components and force application builders to hardcode software component locations on specific resources into their workflows. Additionally, currently available information services are not well adapted to store complete description of software components, forcing the application builder to use only (name, location)-style information about available services and resources. As a consequence these applications are sensitive to dynamic changes in the resource infrastructure, and often fail during execution due to avoidable failures.

Workflow end users frequently want to be able to specify quality of service requirements. These requirements then should be guaranteed—or at least maintained on a best effort basis—by the underlying runtime environment. However, current systems are mostly restricted to best effort optimizations for time-based criteria such as reducing overall execution time or maximizing bandwidth. Several problems must be addressed to overcome current limitations. First, quality of service parameters need to be extended beyond time-based criteria to cover other important aspects of workflow behavior such as responsiveness, fault tolerance, security, and costs. This effort will require collaborative work on the definition of quality of service parameters that can be widely accepted among scientists, so as to provide a basis for interoperable workflow environments or services. Current optimization and planning approaches may have to be radically changed to cope with multi-criteria optimization or planning. Many systems exist for single and some for bi-criteria optimization, but hardly any systems tackle multi-criteria optimization problems. There is no ready-to-use methodology that can deal with this problem in an efficient and effective way; thus, there are many opportunities for research. In developing runtime environment support for quality of service, reservation mechanisms will be an important tool. Both immediate and advance reservations can make the dynamic behavior of infrastructures more predictable, an important prerequisite to guarantee quality of service such as responsiveness and dependability. Moreover, advance reservation can also simplify the scheduling of workflow tasks to resources. However, reservations also introduce challenges relating to policy (who gets to make reservations), fragility (in contrast to a best effort resource, reservable resources may suddenly become unavailable due to a reservation), and efficiency of resource utilization. In providing reservation mechanisms, we should address not only physical resources but also logical resources such as Web services, licenses, and executables. It should be the task of resource management systems to guarantee reservation of physical resources on which logical resources are executed or processed.

Challenging **issues of scale arise in workflow execution**, and these issues will increasingly require advances over the current state of the art. These issues occur in multiple dimensions. First, we see individual workflows becoming increasingly large in many discipilines, as (for example) the quantities of data operated on become larger. As workflows scale from 1,000 to 10,000 and perhaps 1,000,000 tasks or more, new techniques may be needed 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 may be submitting many workflows at once. If these workflows compete for resources or otherwise interact, then appropriate supporting mechanisms are needed in the runtime environment to arbitrating among competing demands. A third scaling dimension concerns the number of resources involved. Ultimately, we can imagine tasks running on millions of data and computing resources (indeed, some systems such as SETI@home already do 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 may be involved in defining the workflow, contributing relevant data, managing is execution, and interpreting results.

New infrastructure services to support workflow management must be provided. Some of these services are analogous to existing data management and information services: for example, workflow repositories and workflow registries. Other more novel services will be concerned with workflows as active processes, and the management of their execution state.

An important issue to address is the **perceived tension between research challenges of scientific workflows and the constraints imposed by existing production-quality infrastructure**. Shared infrastructures such as the TeraGrid and NMI[6] 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 have not been designed to address specific requirements of scientific workflows. Rather, they aim to meet the needs of a broader research community. It is unlikely that commitments can be made at this point by selecting particular architectures or implementations at the workflow layers of shared cyberinfrastructure. Alternative architectures must be explored to understand design tradeoffs in different contexts: for example, workflows designed and tested on a person's desktop that are then 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, these architectures could be designed to be interoperable and compatible, where feasible, with some overall end-to-end, multi-level framework. Follow-on discussions and workshops to understand and address these issues will be extremely beneficial.

The group had the following recommendations:

- Follow up **workshops to articulate how open research areas in scientific workflows can be accommodated by pre-existing cyberinfrastructure** architectures and roadmaps.
- It will be important to **provide a stable system layer on which to develop scientific workflows, while also accommodating new technologies in platforms, networks, and other underlying resources**. Separation of execution and implementation concerns from scientific-relevant aspects of workflows will be crucial to enable reproducibility over many-year timeframes.
- Research should be supported on how to **deliver quality of service guarantees to users.** Such concerns will become increasingly important, since competition for resources will continue to grow as workflow management systems become more widespread and thus facilitate the specification of complex computations. Quality of service requirements should encompass reliability, security, accessibility, and response time.
- Research is required to **address multiple issues of scale**, including workflow size, number of workflows, number of resources used by workflows, and number of participants in collaborations based around workflows.

Concluding Remarks and Recommendations

Workflows provide a formal specification of the scientific analysis process from the data collection, through analysis to the data publication. Workflows can be viewed as recipes for cyberinfrastructure computations,

[6] www.teragrid.org, www.nsf-middleware.org.

providing a representation to describe the end-to-end processes involved in carrying out heterogeneous interdependent distributed computations.

Once this process is captured in declarative workflow structures, workflow management tools could **accelerate the rate of scientific progress** by supporting scientists in creating, merging, executing, and re-using these processes. By assisting scientists in reusing well-known and common practices for analyses, complex computations will become a daily commodity for scientific discovery. By coaching scientists to conduct experiments in neighboring disciplines, cross-disciplinary scientific analyses will become commonplace.

Scientists view **workflows as key enablers for reproducibility of experiments involving large-scope computations.** Reproducibility is engrained in the scientific method, and there is a concern that without this ability there will be a rejection of cyberinfrastructure as a legitimate means to conduct scientific experiments. To enable reproducibility, workflow management systems are needed to capture the end-to-end process at all levels of abstraction, from the science domain level down to the system level. This information is generally termed as **provenance** and is key to reproducibility. 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.

Summary of Recommendations

The following recommendations were made by the workshop participants:

- **Support basic research in computer science to create a science of workflows**. Although existing systems are addressing important issues such as workflow creation, planning, and execution, more comprehensive research is needed to provide easy-to-use workflow construction tools, develop sophisticated automation tools, provide robust workflow execution, manage complex dynamic workflows, etc. There are many open research issues to be resolved in computer science proper that will enable significant progress in the research agenda of scientific workflows.
- **Make explicit workflow representations that capture scientific analysis processes at all levels the norm when performing complex distributed scientific computations.** We need workflow representations at different levels of abstraction, so that we can represent workflows at different levels of refinement, from abstract application-level definition down to operational, system-specific description. These workflow representations can become a starting point for defining common representations that can be interpreted by a variety of workflow systems.
- **Integrate workflow representations with other forms of scientific record.** Data created through workflows should include representations of those workflows as metadata. Articles in scientific publications should include not only textual descriptions of the processes utilized, but also formal descriptions specified as workflows. Laboratory notebooks and invention records should be annotated with workflows and the rationale for their design and final configuration.
- **Support and encourage cross-disciplinary projects involving relevant areas of computer science as well as domain sciences with distinct requirements and challenges.** Cross-disciplinary projects between computer scientists and application scientists are needed to ensure that research efforts are directed towards areas where they can have a significant impact. Other disciplines, such as social sciences and cognitive science should also be engaged to meet the stated challenges.

- **Provide long-term, stable (five or more years) collaborations and programs.** Based on the experiences of the NSF ITR program and the UK e-Science program, the greatest successes were obtained in collaborations that were funded for five years or more, so that collaborations had time to mature and obtain significant results.
- **Define a roadmap to advance the research agenda of scientific workflows while building on existing cyberinfrastructure.** Significant investment in cyberinfrastructure, has resulted in production quality services for data management, high-end and large-scale computation, resource sharing, and distributed computing. It will be important to articulate anticipated requirements to support scientific workflows in the coming years, and develop a roadmap for how the current infrastructure can evolve to accommodate the challenging research agenda that lies ahead. A follow-up workshop on this topic in the near future would be highly beneficial.
- **Coordinate between existing and new projects on workflow systems and interoperation frameworks for workflow tools**. Many current projects have evolved in isolation, working with non-intersecting scientific communities. Capturing best practices will enable a better understanding of the existing capabilities. It will be beneficial to consider the development of a common framework, so that various workflow tools can be integrated and interchanged with others. Scientists will then be able to concentrate on the science rather than have to worry about the particulars of different workflow systems.
- **Hold follow-up, cross-cutting workshops and meetings**. More workshops are needed, to bring together scientists from various domains. Encourage discussion between sub-disciplines of computer science, to and bring in human factors and collaboration considerations to workflow management systems.

In summary, **workflows should become first-class entities in the cyberinfrastructure architecture.** For domain scientists, they are 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 lifecycle from assembly, to execution, to sharing.

Acknowledgements

This material is based upon discussion at the workshop supported by the National Science Foundation under the Grant IIS-0629361.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation

Appendix A: References and General Background

In the last two decades, workflows and process models have been used as a paradigm to organize complex task structures. These workflows are often executed by a variety of parties and involve requirements and constraints that impose complex relationships among the tasks, often cutting across organizational boundaries. Commercial workflow editors and project management tools abound, and several standards for workflow languages have been developed in recent years.

Workflows provide a useful paradigm for conducting large-scale analyses in distributed scientific collaborations. The structure of a workflow specifies what analysis routines need to be executed, as well as details about the locations that contain relevant data to be used for the analysis. These workflows often need to be executed in distributed environments, where data sources may be available in different physical locations and the steps may have execution requirements calling for high-end computing and memory resources at remote locations. In contrast with business workflows and other process models, scientific workflows to date have focused on reflecting computational activity rather than human or organizational activity.

Scientific workflows present new challenges over business workflows and other kinds of process models. They typically use large data sets and computationally intensive tasks and require high-end and distributed computing technology. They are also often iteratively and interactively designed, since that is the nature of the scientific exploration and analysis process they reflect. But they also have simplified requirements in terms of their data flow structure and execution.

This appendix provides references to the literature on scientific workflows, including thematic compilations of recent research in the form of journals and books as well as recent NSF workshops on relevant topics. We also include a pointer to the statements the workshop participants were asked to write in advance of the workshop. Many workshops are convened in conjunction with conferences in specific research areas, although none of these workshops has the breadth of expertise of the NSF workshop that produced this report. The Global Grid Forum (now Open Grid Forum) has a Workflow Management Research Group that meets regularly. Other conferences that have held recent workshops on scientific workflows and distributed computational workflows include the IEEE International Symposium on High Performance Distributed Computing, the International Conference on Computational Science, the IEEE e-Science Conference, the IEEE Conference on Data Engineering, and the ACM Conference on Computer Supported Cooperative Work.

Workshop Participant Statements (provided prior to the workshop):

See the workshop web site http://www.isi.edu/nsf-workflows06 under "Participant Statements".

Special Issues of Journals:

Concurrency and Computation: Practice and Experience, Special Issue on Workflow in Grid Systems. Volume 18, Issue 10, August 2006.

SIGMOD Record, Special Issue on Scientific Workflows, Volume 34, Number 3, September 2005. Includes the overview article: "A Survey of Data Provenance in e-Science", by Y.L. Simmhan, B. Plale and D. Gannon.

Journal of Grid Computing, Special Issue on Scientific Workflows, Volume 3, Number 3-4, September 2005. Includes the overview article: "A Taxonomy of Workflow Management Systems for Grid Computing", by J. Yu and R. Buyya.

Scientific Programming Journal, Special Issue on Scientific Workflows, to appear December 2006.

Books:

Workflows for e-Science, Taylor, I.J.; Deelman, E.; Gannon, D.B.; Shields, M. (Eds.), Dec. 2006, *to appear.*

Related NSF workshops:

NSF Workshop Series on "Dynamic Data-Driven Applications Systems (DDDAS)", Workshop series information and reports available at http://www.nsf.gov/cise/cns/dddas.

NSF Workshop on "Distributed Information, Computation, and Process Management for Scientific and Engineering Environments (DICPM)", May 15-16, 1998, Herndon, Virginia. (report: http://deslab.mit.edu/DesignLab/dicpm/).

NSF Workshop on "Workflow and Process Automation in Information Systems: State-of-the-Art and Future Directions", May 8-10, 1996, Athens, Georgia. (report: http://lsdis.cs.uga.edu/library/download/Report.pdf.)

Other Related Publications:

J. Annis, Y, Zhao, J. Voeckler, M. Wilde, S. Kent, I. Foster, "Applying Chimera Virtual Data Concepts to Cluster Finding in the Sloan Sky Survey," *SC'2002*, 2002.

C. Argyris and D. Schon, *Organizational Learning: A theory of action perspective*, Addison-Wesley, Reading MA, 1978.

L. Bavoil, et al. "VisTrails: Enabling Interactive Multiple-View Visualizations", In *Proceedings of IEEE Visualization*, pages 135–142, 2005.

J Blythe, S Jain, E Deelman, Y Gil, K Vahi, A Mandal, K Kennedy, "Task Scheduling Strategies for Workflow-based Applications in Grids" CCGrid 2005, Cardiff, UK

J. Blythe, E. Deelman, Y. Gil: "Automatically Composed Workflows for Grid Environments," IEEE Intelligent Systems 19(4): 16-23 (2004)

G.A., Bolcer and R.N. Taylor, Advanced Workflow Management Technologies, *Software Process--Improvement and Practice*, 4, 125-171, 1998.

T.R. Browing, E. Fricke, and H. Negele, Key Concepts in Modeling Product Development Processes, *Systems Engineering*, 9(2), 104-128, 2006.

S. P. Callahan at al. "Using Provenance to Streamline Data Exploration through Visualization", SCI Institute Technical Report, No. UUSCI-2006-016, University of Utah, 2006.

E. Deelman et al. "Managing Large-Scale Workflow Execution from Resource Provisioning to Provenance tracking: The CyberShake Example", *e-Science 2006*, Amsterdam, December 4-6, 2006

E. Deelman et al, "Pegasus: a Framework for Mapping Complex Scientific Workflows onto Distributed Systems", *Scientific Programming Journal*, Vol 13(3), 2005, Pages 219-237

E. Deelman et al. "Workflow Management in GriPhyN," in *Grid Resource Management*, J. Nabrzyski, J. Schopf, and J. Weglarz editors, Kluwer, 2003.

E. Deelman et al. "Mapping Abstract Complex Workflows onto Grid Environments," *Journal of Grid Computing*, Vol.1, no. 1, 2003, pp. 25-39.

A. Fernandez, B. Garzaldeen, I. Grutzner and J. Munch, Guided Support for Collaborative Modeling, Enactment and Simulation of Software Development Processes, *Software Process--Improvement and Practice*, 9, 95-106, 2004.

I. Foster, "Service-Oriented Science," *Science*, 308:814-817, 2005.

I. Foster, J. Voeckler, M. Wilde, Y. Zhao, "Chimera: A Virtual Data System for Representing, Querying, and Automating Data Derivation," *14th Conference on Scientific and Statistical Database Management*, 2002.

J. Freire et al. "Managing Rapidly-Evolving Scientific Workflows.", To appear in Proceedings of the International Provenance and Annotation Workshop (IPAW'06), volume 4145 of Lecture Notes in Computer Science.

Y. Gil, E. Deelman, J. Blythe, C. Kesselman, and H. Tangmurarunkit "Artificial Intelligence and Grids: Workflow Planning and Beyond,". *IEEE Intelligent Systems*, January 2004

J.-R. Gruser, L. Raschid, V. Zadorozhny, T. Zhan: Learning Response Time for WebSources Using Query Feedback and Application in Query Optimization. *VLDB J.* 9(1): 18-37 (2000)

Y. Han, A. Sheth and C. Bussler, "A taxonomy of adaptive workflow management", *ACM CSCW 98 Workshop Proceedings*, Towards Adaptive Workflow Systems, Seattle, 1998. http://citeseer.ist.psu.edu/context/848353/0

C. Jensen and W. Scacchi, Experiences in Discovering, Modeling, and Reenacting Open Source Software Development Processes, in Mingshu Li, Barry Boehm, and Leon J. Osterweil (eds.), *Unifying the Software Process Spectrum*, 442-469, Springer-Verlag, 2006.

M. Klein, C. Dellarocas, and A. Bernstein, Introduction to the Special Issue on Adaptive Workflow Systems, *Computer Supported Cooperative Work*, 9, 265-267, 2000.

M. Klein and C. Petti, A Handbook-Based Methodology for Redesigning Business Processes, *Knowledge and Process Management*, 13(2), 108-119, 2006.

B. Latour, *Science in Action*, Cambridge, MA, Harvard University Press, 1987.

B. Latour and S. Woolgar, Laboratory Life: The Social Construction of Scientific Facts, London, Sage, 1979.

L. Ljung, *System Identification - Theory For the User*, 2nd ed, PTR Prentice Hall, Upper Saddle River, NJ, 1999.

P. Maechling et al, "Simplifying construction of complex workflows for non-expert users of the Southern California Earthquake Center Community Modeling Environment" SIGMOD Record 34(3): 24-30 (2005).

P. Mi and W. Scacchi, Articulation: An Integrated Approach to the Diagnosis, Replanning, and Rescheduling of Software Process Failures, *Proc. 8th. Knowledge-Based Software Engineering Conference*, Chicago, IL, IEEE Computer Society, 77-85, September 1993.

K. L. Myers and S. F. Smith, Issues in the Integration of Planning and Scheduling for Enterprise Control, In *Proceedings of the DARPA-JFACC Symposium on Advances in Enterprise Control*, 1999.

J. Noll and W. Scacchi, Supporting Software Development in Virtual Enterprises, J. Digital Information, 1(4), 19999.

J. Noll and W. Scacchi, Specifying Process-Oriented Hypertext for Organizational Computing, *J. Network and Computer Applications*, 24(1), 39-61, 2001.

P. Oreizy, M. Gorlick, R. N. Taylor, D. Heimbigner, G. Johnson, N. Medvidovic, A. Quilici, D. Rosenblum, and A. Wolf. An Architecture-Based Approach to Self-Adaptive Software, *IEEE Intelligent Systems*, 14(3), 54-62. May/June 1999.

W. Scacchi and P. Mi, Process Life Cycle Engineering: A Knowledge-Based Approach and Environment, *Intelligent Systems in Accounting, Finance, and Management*, 6(1), 83-107, 1997.

W. Scacchi, Experience with Software Process Simulation and Modeling, *J. Systems and Software*, 46(2/3), 183-192, 1999.

W. Scacchi, Understanding Software Process Redesign using Modeling, Analysis and Simulation, *Software Process--Improvement and Practice*, 5(2/3), 183-195, 2000.

G. Singh et al. "Optimizing Grid-Based Workflow Execution", *Journal of Grid Computing*, Volume 3(3-4), December 2005, Pages 201-219.

D. Sulakhe, A. Rodriguez, M. D'Souza, M. Wilde, V. Nefedova, I. Foster, N. Maltsev, "GNARE: An Environment for Grid-Based High-Throughput Genome Analysis," *Journal of Clinical Monitoring and Computing*, 2005.

J. Van Horn, J. Dobson, J. Woodward, M. Wilde, Y. Zhao, J. Voeckler, I. Foster, "Grid-Based Computing and the Future of Neuroscience Computation," *Methods in Mind*, MIT Press, 2006.

K. Verma and A. Sheth, "Autonomic Web Processes", *Proceedings of the International Conference on Services-oriented Computing*, 2005.

C. Wroe, C. Goble, A. Goderis, P. Lord, S. Miles, J. Papay, P. Alper, L. Moreau, Recycling workflows and services through discovery and reuse, *Concurrency and Computation: Practice and Engineering*, Published Online in advance: 11 May 2006, http://www3.interscience.wiley.com/cgi-bin/abstract/112609729/ABSTRACT

Y. Zhao, J. Dobson, I. Foster, L. Moreau, M. Wilde, "A Notation and System for Expressing and Executing Cleanly Typed Workflows on Messy Scientific Data," *SIGMOD Record*, 34(3):37-43, 2005.