Guest Editors' Introduction to the Special Section on Scientific Workflows

Bertram Ludäscher
Dept. of Computer Science & Genome Center
University of California, Davis
1 Shields Ave, Davis, CA 95616, USA
ludaesch@ucdavis.edu

Carole Goble
School of Computer Science
The University of Manchester
Manchester, M13 9PL, UK
carole@cs.man.ac.uk

1. INTRODUCTION

Business-oriented workflows have been studied since the 70's under various names (office automation, workflow management, business process management) and by different communities, including the database community. Much basic and applied research has been conducted over the years, e.g. theoretical studies of workflow languages and models (based on Petri-nets or process calculi), their properties, transactional behavior, etc.

Recently, and largely unnoticed by the database community, *scientific workflows* have gained momentum due to their central role in e-Science and cyberinfrastructure applications, i.e., where scientists need to "glue" together data management, analysis, simulation, and visualization services over often voluminous and (structurally and semantically) complex, distributed scientific data and services. While sharing commonalities with their business workflow relatives, scientific workflows often pose different challenges. For example, scientific workflows are typically data-centric, dataflow-oriented "analysis pipelines" (as opposed to task-centric and control-flow oriented business workflows) and can be very computationally expensive (often requiring parallel and/or Grid computing capabilities).

Another characteristic is that scientific workflows are often more metadata and annotation-intensive, since repurposing of a scientific data product in another scientist's study requires detailed (and preferably machine-processable) context and data provenance information. Finally, scientists typically are rather individualistic and are more likely to create their own "knowledge discovery workflows", whereas in business, users are commonly restricted to using carefully designed and predetermined automation workflows in a constrained way.

Scientific workflow *systems* are related to (and can have features of) mathematical problem solving environments [1], LIMS (Laboratory Information Management Systems), dataflow visualization systems (AVS, IBM's OpenDX, SciRUN, etc.), and distributed (Grid) scheduling and execution environments. Users of scientific workflow systems range from bench scientists to computational scientists and of course include the new breed of "hybrid" e-scientists. Scientific workflows are useful to capture, document, archive, share, execute, and reproduce scientific data analysis pipelines from all disciplines (e.g., biology, medicine, ecology, chemistry, physics, geosciences, and astronomy). Clearly, different disciplines and subdisciplines can have different requirements and characteristics w.r.t. data volume, (structural and semantic) heterogeneity, computational complexity, etc.

Grid computing (now largely web service based) has stimulated workflow developments, from the orchestration of long running applications to the scheduling of job submissions to marshalled compute resources. Many scientific workflow systems can execute remote web services and local tools (e.g., via a command-line interface).

2. SPECIAL SECTION OVERVIEW

With this special section we aim at providing a glimpse of a number of research and development activities and technical challenges in scientific workflows. Due to space limitations, we can only provide a very limited snapshot of ongoing work. Nevertheless, we hope that this special section can serve as a first sample of the range of issues that define the current state-of-the-art in scientific workflows and that provide a starting point for further research and contributions by the database community.

The call for papers attracted 31 submissions, indicating the large interest in the topic. Based on the peer reviews by about thirty external reviewers, 9 papers were accepted. Several of the papers use case studies from the life sciences: two papers use applications in biomedical image analysis, and two others use bioinformatics and phylogenetics examples. The geosciences are also represented. These disciplines are characterized by large, distributed and heterogeneous data sets, which are subject to change and regular re-interpretation, and need to be combined and processed in differing and non-prescriptive ways by third party scientists.

Scientific Workflow Systems. There is a plethora of scientific workflow environments covering a range of scientific disciplines. YU and BUYYA's *"Taxonomy of Scientific Workflow Systems for Grid Computing"* sets the scene by briefly characterizing and classifying various approaches for building and executing workflows on the Grid. A comprehensive scientific workflow system has demanding execution requirements. They should be able to schedule workflow tasks (typically in a distributed/Grid environment), monitor and control execution, allow on-the-fly visualization and computational steering, facilitate "pause and rerun", gracefully manage failure, and support various static and dynamic analysis and optimization techniques.

Metadata for Workflow Reuse and Provenance. Scientific workflows are pivotal knowledge components in e-Science. The scientific protocol encapsulated by the workflow provides a context and history for its products that enables their interpretation. Data provenance is such a critical component of scientific workflows, that the *"Survey of Data Provenance in e-Science"* by SIMMHAN, PLALE and GANNON is a welcome summarization of the key research efforts and open challenges.

A workflow is itself is know-how about a scientific method that can be shared and reused, or act as a template for new versions of workflows. By reusing workflows we can spread best practice, avoid wasteful duplicated effort, and foster scientific collaboration. Along with shared data warehouses and service registries, we envision shared catalogues of workflows indexed by metadata, as do MEDERIOS ET AL in *"WOODSS and the Web: Annotating and Reusing Scientific Workflows"*. This presents challenges of how to describe and query workflows, understanding models of reuse, and presenting the workflows in terms of a user model rather than a delivery paradigm.

In *"Simplifying Construction of Complex Workflows for Non-Expert Users of the Southern California Earthquake Centre Community Modelling Environment"* MAECHLING ET AL pick up this theme by returning the scientist at the center of a real scientific workflow application. Workflow templates are shared and reused by scientists; metadata is used to intelligently guide scientists to build and refine their own workflows.

Workflow Support for Data Collections. As scientific data analysis is the main use of workflows, it is becoming apparent that large-scale data-intensive workflows will dominate e-Science. Two papers take the management of data collections in workflow as their theme, whilst two more take a more conventional database line, arguing that database technologies can support workflow environments.

When constructing workflows that operate on large and complex datasets, the ability to describe and introspect on the types of both datasets and workflow components is invaluable – for type checking and iteration over collections, for example. If the datasets were described using clearly defined and shared metadata, and stored in well-organized databases, then this would be straightforward. However, the real world is not like this. Datasets are commonly files, and metadata is encoded in directory and file names, employed in ad-hoc ways. The physical manifestation of the dataset is conflated with its logical structure.

In *"A Notation and System for Expressing and Executing Cleanly Typed Workflows on Messy Scientific Data"*, ZHAO ET AL present a typed workflow notation and system that allows workflows to be expressed in terms of abstract XML data types that are then executed over diverse physical representations, decoupling the physical and logical descriptions without forcing change in the datasets themselves.

MCPHILLIPS and BOWERS in *"An Approach for Pipelining Nested Collections in Scientific Workflows"* take up the theme of appropriate approaches for workflow execution over large-scale nested data collections. Their framework illustrates a new scientific workflow programming paradigm, emphasizing extensibility through collection-aware actors, concurrent operations, on the fly component customization and exception management.

Database Support for Workflow Execution. Several works focus on database support for scientific workflows. SHANNON ET AL pick up on the prevalence of XML for describing and representing datasets, that there should be *"XML Database Support for Distributed Execution of Data-intensive Scientific Workflows"*. They use the Mobius framework for on demand creation and federation of XML databases and DataCutter for streaming data between processes.

SHANKER ET AL go further by arguing in *"Integrating Databases and Workflow Systems"* that workflow execution and data management are so co-dependent that this calls for a workflow modelling language that tightly integrates workflow management systems and database management systems. Rather than a process dominated viewpoint, where data is a product of a workflow engine, they see workflow execution as a means of generating data products as an extension of SQL, putting the database at the center rather than the workflow execution machinery.

This resonates with the Virtual Data Language discussed by ZHAO ET AL. A more loosely coupled approach has been proposed by the OGSA-DQP project; they suggest that database queries can be workflow jobs and workflow components can be queries [2].

Workflow Scheduling. In *"Scheduling of Scientific Workflows in the ASKALON Grid Environment"*, WIECZOREK, PRODAN and FAHRINGER's paper focuses particularly on execution performance for scheduling in Grid environments, and represent a particular use of workflow, that is scheduling *job* submissions over compute resources, sometimes termed "workflow in the small". This is in contrast to workflows for orchestrating *applications*, termed "workflow in the large", such as those developed by myGrid's Taverna system [3] or Kepler [4].

Conclusion. It has been recognized by funding agencies and their respective programmes and initiatives (e.g., NIH Roadmap, NSF ITR, Cyberinfrastructure, DOE SciDAC, UK e-Science, various EU programmes, etc.) that scientific advances and discoveries are facilitated through novel IT infrastructure and tools. Scientific workflows provide the interface between scientists and this infrastructure. We think that the many and various types of technical challenges in scientific workflow modeling, design, optimization, verification etc. provide a rich playing field and great opportunity for database researchers.

Acknowledgements. We thank the SIGMOD-Record editor, Mario Nascimento, and all external reviewers of this special section for their support.

3. REFERENCES

[1] R. F. Boisvert and E. N. Houstis, editors. *Computational Science, Mathematics, and Software.* Purdue University Press, 1999.

[2] OGSA-DQP: Service Based Distributed Query Processor. http://www.ogsadai.org.uk/dqp/.

[3] myGrid. http://www.mygrid.org.uk/.

[4] Kepler. http://kepler-project.org/.