Computational Provenance

Record keeping has always been an essential component of science and engineering, but it has become even more so recently. As computers get faster, we perform increasingly complex computations—and as storage gets cheaper, we accumulate larger volumes of data. The complete process, from data acquisition through analysis, is inherently exploratory: users experiment with different simulation models, parameters, and data mining and visualization techniques. But when they find an interesting result, it can be hard to remember which of the many trial-and-error paths produced a particular result without a detailed record.

For complex computations that manipulate a lot of data, the traditional laboratory notebook or other manual approaches to maintaining this information just aren't feasible. Today, users must expend substantial effort managing and recording data to answer the most basic questions: Who created this data product and when? Who modified it and when? What process created the data product? Did the same raw data produce two different products? Not only is this process time-consuming, it's also error-prone.

1521-9615/08/$25.00 © 2008 IEEE
Copublished by the IEEE CS and the AIP

Cláudio T. Silva
University of Utah
Joel E. Tohline
Louisiana State University

Systematic mechanisms for capturing this information are at the heart of a new field of research called *computational provenance*. Most dictionaries define provenance as an object's source or origin—a record of an item's ultimate derivation and passage through various owners. Ultimately, provenance helps determine an object's value, accuracy, and authorship. But in addition to enabling result reproducibility,[1] provenance for computational tasks and the data they manipulate and derive has other benefits as well. In particular, it helps users interpret and understand results—in some cases, it can be more important than the actual results themselves.

The articles in this special issue show many uses of provenance that go beyond result reproducibility. The first article—"Provenance for Computational Tasks: A Survey," by Juliana Freire and her colleagues—targets potential users of provenance technology who aren't quite experts on the topic. It covers the key issues involved in capturing, storing, and querying computational task provenance and describes some existing systems.

The next two articles discuss ongoing research projects and their use of provenance information. In "Provenance in High-Energy Physics Workflows," Andrew Dolgert and his colleagues describe their efforts as part of a large-scale international collaboration of physicists analyzing data from CERN's Large Hadron Collider. Their research involves petabytes of data accessed by thousands of systems and collaborators, so they're aiming for a software-based infrastructure that will replace the traditional lab notebook. In "Provenance in Comparative Analysis: A Study in Cosmology," Erik Anderson and his colleagues focus on a collaborative visualization framework to help analyze data from the Cosmic Code Comparison Project, which aims to establish the robustness of results from a set of cosmological simulations. The large number of simulations, plots, graphs, and visualizations involved has rendered manual bookkeeping of result provenance nearly impossible. In particular, the authors describe how they customized the VisTrails system to get the type of high-end visualization required.

The last two articles discuss specific aspects of provenance management technology. In "Provenance: The Bridge between Experiments and Data," Simon Miles and his colleagues present an innovative use of a provenance project coupled with a workflow engine. Finally, in "Problem-Solving Methods for Understanding Process Executions," Jose Manual Gómez-Pérez and Oscar Corcho approach provenance from the perspective of providing users with meaningful interpretations of process executions. These interpretations aim to explain provenance in a way that's closer to how domain experts see it.

We hope this theme issue raises awareness and contributes to a better understanding of the issues surrounding computational provenance to the broader *CiSE* community. Research and technology being developed in this area have the potential to be transformative and improve how people do science in a variety of domains. By examining the sequence of steps an expert followed to produce a result, a user can gain insights into the chain of reasoning used, learn by example, and potentially reduce the time to insight. Combined with social networking, provenance could even serve as a catalyst to mass collaboration. Likewise, given the ease with which we can share digital information, the provenance infrastructure currently available could serve as strong motivation for authors to publish, along with their scientific articles, data, and codes, the actual process they used to solve a problem. Provenance is destined to gain a higher profile in coming years as the broad field of computational sciences matures and more strongly emphasizes the reproducibility of archival simulation results.

Acknowledgments

This work was funded by the US National Science Foundation, the US Department of Energy, and an IBM Faculty Award.

Reference

1. M. Schwab, N. Karrenbach, and J. Claerbout, "Making Scientific Computations Reproducible," *Computing in Science & Eng.*, vol. 2, no. 6, 2000, pp. 61–67.

Cláudio T. Silva *is an associate professor at the University of Utah. His full vita information appears on p. 21. Contact him at csilva@cs.utah.edu.*

Joel E. Tohline *is an alumni professor at Louisiana State University. He has a PhD in astronomy from the University of California, Santa Cruz. Contact him at tohline@lsu.edu.*

Editor's Note: *Although Cláudio T. Silva is listed as a guest editor and helped pull together the articles for this special issue, he wasn't involved in the peer-review process at all (co-guest editor Joel E. Tohline handled this duty).*