From Data Sharing to Reproducible Science via Workflows and Provenance

Bertram Ludaescher[1] Paolo Missier[2] Karthik Ram[3]

[1]Dept. of Computer Science & Genome Center, University of California, Davis
[2]Dept. of Computing Sciences, Newcastle University, United Kingdom
[3]Dept. of Environmental Science, Policy & Management, University of California, Berkeley

A major objective of the open science paradigm is to make all outputs of the scientific process---data, software, analysis, results, and publications---freely available to the public and researchers. The inclusion of *provenance* information, in addition to the underlying data, not only benefits a wider community but also promotes greater transparency in scientific research, helps establish confidence in the trustworthiness of the data, and facilitates reproducible research.

Historically, methods sections in research publications have served as descriptions of the "scientific workflow" used and were meant to provide the necessary provenance to allow readers to reproduce results [DG10]. However, in many cases such representations of the methods only partially capture the details of the process that were used to produce the results. Records of the methods can take a broad variety of forms, ranging from handwritten lab notes on the procedures that were followed, to executable process specifications that use a formal (scientific) workflow language.

In order to scale processing and analysis to the ever increasing amounts of scientific data, e-science experiments are being automated as much as possible, using loosely linked scripts in languages such as Python, Perl or R, or using dedicated *scientific workflow systems*, e.g., see [Kepler,Taverna,Vistrails] and many others [DBE+07,DGS+08]. In addition to documenting and linking all steps from data acquisition to discovery, workflows provide a way for such steps to be easily shared, replicated, and extended by other researchers.

Even when workflows are machine processable, these specifications cannot always be used to repeat the experiment after a certain period of time, a phenomenon known as "workflow decay" [DBM+11]. For each workflow run, the specific parameter settings, versions of input and intermediate data, and other relevant provenance information may need to be captured, if a scientist wants to harness the full potential of provenance information. Thus, in addition to capturing different versions of workflow specifications, an even finer level provenance capture is needed, i.e., the recording of runtime data provenance. Scientific workflow scripts need to be instrumented to capture this information at runtime, while more and more scientific workflow systems offer this functionality as a built-in. The resulting detailed provenance traces allow scientists to interpret, validate, and debug workflow runs, resulting in data products whose quality is easier to assess, and which ultimately can be more trusted.

For example, when the research at the East Anglia climate research unit was called into question, a significant outcome of the resulting investigation was the establishment of a workflow server (http://westerly.badc.rl.ac.uk:8080/alws/about.html) which now formally

documents all associated workflows:

> *The overall aim of the Advanced Climate Research Infrastructure for Data (ACRID) project is to implement a linked-data approach for sharing some example climate datasets, and in doing so develop the necessary architecture, infrastructure and tools that might be implemented more widely within the climate science community. The ACRID project will try to demonstrate how the publication of datasets developed by the climate science community might be improved, so that*
>
> 1. *the provenance of the published data can be more clearly recorded (e.g. data sources and versions, software versions, and processing options);*
> 2. *published data can be recreated more straightforwardly from source data even a number of years after publication;*
> 3. *data can be cited in a way that links more directly to the precise version of data that was used and, by using the linked-data approach, make relationships between different datasets clearly visible.*

A challenge to the vision of reproducible science through workflows and provenance is the diversity in systems, models, and languages in use for workflows and provenance. This has led to community efforts towards a "standard" model of provenance [Mor+08], leading in particular to the specification of an Open Provenance Model (OPM) [MCF+11]. In addition, since January 2011, the W3C has started an effort to develop the *PROV* provenance data model within the Provenance Working Group of the W3C. The model is described in an evolving document [MM11]. Although it is designed to promote the exchange of provenance for Web data, it allows provenance assertions to be made regarding general data derivation relationships that are mediated by data consuming and producing processes. Due to the generality of its scope, a model of process structure (workflows) has deliberately been left out of the scope of the new PROV model. However, PROV can be extended to accommodate a description of the process description, using the extension mechanisms that are built into the specification.

The Data Observation Network for Earth [DataONE] has established a *Provenance and Scientific Workflows* Working Group, whose task it is to develop a provenance model that unifies the different provenance models employed by workflow systems and script-based approaches (e.g. the R data analysis system). This working group extends the OPM and the W3C PROV model to take into account (a) workflow-specific elements and observables, and (b) data elements, in particular data citation models to facilitate data interoperation and reuse via a shared model of provenance [MLD+11]. We believe that, although a unified workflow language is not on the horizon (and may not even be feasible or desirable), a unified provenance model is absolutely critical to facilitate open, reproducible science and data sharing through workflows and provenance.

References

[BG+11] Bauer, B., Gukelberger, J., Surer, B., & Troyer, M. Publishing Provenance-rich Scientific Papers. Procs. TAPP'11 (Theory and Practice of Provenance). Heraklyion, Crete, Greece, 2011.

[BMR+08] Bowers, S., McPhillips, T., Riddle, S., Anand, M., & B. Kepler/pPOD: Scientific workflow and provenance support for assembling the tree of life. In J. Freire, D. Koop, & L. Moreau (Eds.), Procs. Provenance and Annotation of Data and Processes (IPAW) 2008 (pp. 70-77). Springer.

[CFS+06] Callahan, S. P., Freire, J., Santos, E., Scheidegger, C. E., Silva, C. T., & Vo, H. T. VisTrails: visualization meets data management. Proc. SIGMOD 2006 (pp. 745-747).

[DataONE] Data Observation Network for Earth, http://dataone.org

[DBE+07] Susan B. Davidson, Sarah Cohen Boulakia, Anat Eyal, Bertram Ludäscher, Timothy M. McPhillips, Shawn Bowers, Manish Kumar Anand, Juliana Freire: Provenance in Scientific Workflow Systems. IEEE Data Eng. Bull. 30(4): 44-50, 2007.

[DBM+11] Roure, D. D., Belhajjame, K., Missier, P., & Al., E. (2011). Towards the preservation of scientific workflows. 8th International Conference on Preservation of Digital Objects (iPRES 2011). Singapore.

[DG10] Dave De Roure, D. and Goble, C. Anchors in Shifting Sand: the Primacy of Method in the Web of Data. In: Web Science Conference 2010, 26-27 April, 2010, Raleigh, NC, USA.

[DGS+08] Ewa Deelman, Dennis Gannon, Mathew Shields, Ian Taylor, Workflows and e-Science: An overview of workflow system features and capabilities, Future Generation Computer Systems, July 2008.

[Kepler] Kepler Project. http://www.kepler-project.org

[LAB05] Ludäscher, B., Altintas, I., & Berkley, C. Scientific Workflow Management and the Kepler System. Concurrency and Computation: Practice and Experience, 18, 1039-1065, 2005.

[MCF+11] Luc Moreau, Ben Clifford, Juliana Freire, Joe Futrelle, Yolanda Gil, Paul Groth, Natalia Kwasnikowska, Simon Miles, Paolo Missier, Jim Myers, Beth Plale, Yogesh Simmhan, Eric Stephan, Jan Van den Bussche, The Open Provenance Model core specification (v1.1), Future Generation Computer Systems, Volume 27, Issue 6, p. 743-756, June 2011.

[MLA+08] Luc Moreau, et al. Special Issue: The First Provenance Challenge. Concurrency and Computation: Practice and Experience 20(5): 409-418 (2008)

[MLD+11] P. Missier, B. Ludaescher, S. Dey, M. Wang, T. McPhillips, S. Bowers, M. Agun, Golden-Trail: Retrieving the Data History that Matters from a Comprehensive Provenance Repository, 7th International Digital Curation Conference (IDCC), Bristol, UK, 2011.

[MM11] Luc Moreau, Paolo Missier. The PROV Data Model and Abstract Syntax Notation. First working public draft, World Wide Web consortium, October 2011.

[MWF+07] Miles, S., Wong, S. C., Fang, W., Groth, P. T., Zauner, K.-P., & Moreau, L. Provenance-based validation of e-science experiments. J. Web Sem., 5, 28-38, 2007.

[N10] Nekrutenko, A. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biology, 11(8), 2010, R86.

[TMD+07] Turi, D., Missier, P., Roure, D. D., Goble, C., & Oinn, T., Taverna Workflows: Syntax and Semantics. Proc. of the 3rd e-Science Conference. Bangalore, India, 2007.

[Taverna] Taverna. http://www.taverna.org.uk/

[Vistrails] Vistrails http://www.vistrails.org/