Accelerating Scientific Discovery through Computational Workflows

2012 NSF Workshop on Discovery Informatics

Yolanda Gil
Information Sciences Institute and Department of Computer Science
University of Southern California

January 23, 2012

Workflows offer explicit representations of computational methods, and have long been recognized as a crucial element of scientific discourse [Gil et al 07]. A workflow can be seen as a software instrument that offers a new lens on data. Treating workflows as science currency could change the practice, accelerate the pace, and broaden participation in scientific endeavors:

Scientific discovery would advance at a faster pace if computational methods were published in an operational form so they could be readily inspectable and reusable. Several systems support the publication of workflows as an integral part of a scientific publication (e.g., [Falcon 07]). Extending publications with workflows makes computational experiments inspectable and reproducible [Bourne 10; Lehrer 10; Naik 11]. Workflow sharing sites are available to deposit and reuse workflows [De Roure et al 09]. Workflows can be shared through open standards [Moreau et al 11; Gil et al 10; Garijo and Gil 11]. However, the investment in creating workflows is not always undertaken in the scientific community. *What are the technical, social, and practical barriers for broader adoption of workflow publication as routine practice?*

Scientific discovery would be greatly facilitated by workflow systems that facilitate the exploration of the experiment design space. Computational experiments can often be realized by many alternative algorithms and algorithm combinations. In addition, these algorithms can have many parameter settings such as thresholds that affect significantly the results obtained. Scientists can in practice only explore a subset of that design space, and not in systematic or principled ways. Workflow elaboration algorithms can search through the space of possible instantiations of a high-level specification of an experiment [Gil et al 12]. Workflow discovery algorithms can find and suggest possible methods for given experimental goals [Bergmann and Gil 11]. *What kinds of automated or interactive exploration tools can assist scientists to improve the process of designing computational experiments?*

Publishing metadata about process provenance will significantly improve data sharing and reuse. Many data repositories have basic metadata about datasets but lack process provenance that describes how the data was obtained. Raw data (obtained from instruments and sensors) is rarely shared, rather scientists tend to share data after undergoing some quality control or preparation processes. Those processes are rarely

published when the data is published, which limits the utility of the data. *How do we change data sharing practices so that metadata routinely includes process provenance to facilitate data interpretability and reuse?*

Workflows could facilitate new kinds of cross-disciplinary discoveries. Workflows capture data analysis expertise and would enable easy exchange of that expertise to new scientists in other disciplines. Workflows could also document how data is processed, so that published data is appropriately documented for others to use. *How can workflows be published, discovered, and applied across scientific disciplines?*

References

Bergmann, R., and Gil, Y. "Retrieval of Semantic Workflows with Knowledge Intensive Similarity Metrics." Proceedings of the Nineteenth International Conference on Case Based Reasoning (ICCBR), Greenwich, London, 2011. *Available from http://www.isi.edu/~gil/papers/bergmann-gil-iccbr11.pdf*

Bourne, P. "What Do I Want from the Publisher of the Future?" PLoS Computational Biology, 2010. *Available from http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1000787*

De Roure, D.; Goble, C.; and Stevens, R. "The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows." Future Generation Computer Systems 25, 2009. *Available from doi:10.1016/j.future.2008.06.010*

Falcon, S. "Caching code chunks in dynamic documents: The Weaver package." Computational Statistics, (24)2, 2007. *Available from http://www.springerlink.com/content/55411257n1473414/*

Garijo, D.; and Y. Gil. "A New Approach for Publishing Workflows: Abstractions, Standards, and Linked Data." Proceedings of the Sixth Workshop on Workflows in Support of Large-Scale Science (WORKS'11), held in conjunction with SC 2011, Seattle, Washington. *Available from http://www.bibbase.org/cache/www.isi.edu__7Egil_publications.bib/garijo-gil-works11.html*

Gil, Y.; Deelman, E.; Ellisman, M. H.; Fahringer, T.; Fox, G.; Gannon, D.; Goble, C. A.; Livny, M.; Moreau, L.; and Myers, J. "Examining the Challenges of Scientific Workflows." IEEE Computer, 40(12), 2007. *Preprint available from http://www.bibbase.org/cache/www.isi.edu__7Egil_publications.bib/computer-NSFworkflows07.html*

Gil, Y. "From Data to Knowledge to Discoveries: Artificial Intelligence and Scientific Workflows." Scientific Programming, 17(3), 2009. Available from http://www.isi.edu/~gil/papers/gil-sp08.pdf

Gil, Y.; Cheney, J.; Groth, P.; Hartig, O.; Miles, S.; Moreau, L.; and daSilva, P. P. "Final Report of the W3C Provenance Incubator Group." Report from the W3C Provenance Incubator Group, first release: November 30, 2010. *Available from* *http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/*

Gil, Y.; Gonzalez-Calero, P. A.; Kim, J.; Moody, J.; and Ratnakar, V. "A Semantic Framework for Automatic Generation of Computational Workflows Using Distributed Data and Component Catalogs." To appear in the Journal of Experimental and Theoretical Artificial Intelligence, 2012. *Preprint available from* *http://www.bibbase.org/cache/www.isi.edu__7Egil_publications.bib/gil-etal-jetai10.html*

Lehrer, J. "The Truth Wears Off: Is There Something Wrong with the Scientific Method?" The New Yorker, December 13, 2010. *Available from* *http://www.newyorker.com/reporting/2010/12/13/101213fa_fact_lehrer*

Moreau, L.; Clifford, B.; Freire, J.; Futrelle, J.; Gil, Y.; Groth, P.; Kwasnikowska, N.; Miles, S.; Missier, P.; Myers, J.; Plale, B.; Simmhan, Y.; Stephan, E.; and denBussche, J. V. "The Open Provenance Model Core Specification (v1.1)." To appear in Future Generation Computer Systems, 2011. *Preprint available from* *http://www.bibbase.org/cache/www.isi.edu__7Egil_publications.bib/moreau-etal-fgcs11.html*

Naik, Gautam. "Scientists' Elusive Goal: Reproducing Study Results". The Washington Post, December 2, 2011. *Available from* *http://online.wsj.com/article/SB10001424052970203764804577059841672541590.html*