PUBLISHING COMPUTATIONAL WORKFLOWS OF SCIENTIFIC ARTICLES: THE CASE OF THE TUBERCULOSIS DRUGOME

PROJECT DESCRIPTION

Phil Bourne, bourne@sdsc.edu
Yolanda Gil, gil@isi.edu

April 12, 2011 (Last update: April 28, 2011)

The goal of this work is to extend articles with scientific workflows to 1) represent computations carried out to obtain the published results, essentially capturing explicitly data analysis pipelines, and 2) represent an abstraction of those computations that represents the semantics of the data analysis method in an execution-independent manner. This would make scientific results more reproducible because articles would have not just a textual description of the computational process described in the article but also a workflow that, as a computational artifact, could be analyzed and re-run automatically.

The idea of enhancing scientific publications with explicit workflows has been articulated before [Bourne 10]. Some systems exist that allow articles to be augmented with scripts or workflows, such as Weaver and GenePattern [Leisch 02; Mesirov 10]. However none of these systems can reproduce workflows created by others. This is a major impediment to reproducibility and sharing, since different scientists and labs choose to use different software infrastructure for their computational processes. What is needed is a framework that allows the publication of workflows in a system-independent manner that enables scientists to reuse them across systems.

Workflows represent scientific contributions, and should be treated as first-class citizens in science [Gil et al 08]. Unless they are published explicitly and completely, the methods and procedures are not completely revealed. Scientific articles describe computational methods informally, therefore requiring a significant effort from others to reproduce and to reuse, to the point of being referred to as "forensic" research [Baggerly and Coombes 09]. Retractions do occur more often than is desirable [The Scientist 10]. Studies have shown that reproducibility is not achievable from the article itself, even when datasets are published [Bell et al 09; Ioannidis et al 09]. Publishers themselves are asking the community for an end to "black box" science [Nature 06].

But the impact of this issue is well beyond scientific research circles. Clinical trials based on erroneous results pose significant threats to patients [Hutson 10]. The public has neutral to low trust on scientists for important topics such as flu pandemics, depression drugs, and autism causes [Scientific American 10]. The validity of scientific research methods has been put in question [Lehrer 10].

A workflow has value in and on itself as a "digital instrument" that enables scientists to analyze data through the lens of the method that the workflow represents. A technological challenge is how to make such instruments reusable across labs and institutions, since each has a diverse software and hardware infrastructure. What is needed is a mechanism to publish workflows that would give scientists access to such digital instruments at very low cost.

We propose to develop a framework to extend scientific articles with an explicit representation of the computational workflows used in the research in a manner that is both platform independent and easily reusable in different platforms. The work will include several major tasks:

* Representations to extend a scientific article with a high-fidelity provenance record of the computational workflows executed to obtained the results published in the article. The representation of workflow provenance will be based on the Open Provenance Model (http://openprovenance.org), a well-known model developed by the workflow community and that has been integrated into a variety of workflow systems [Moreau et al 11].

* Representations to extend a scientific article with a semantic representation of the executed workflow, which captures a conceptual and execution-independent view of the data analysis method. The semantic representation of the workflow will be based on the Wings workflow system (http://wings.isi.edu), the only workflow system with semantic representations of classes of steps and data as well as semantic constraints [Gil et al 11a; Gil et al 11b]. Both the provenance records and the semantic workflow representations will be consistent.

* A set of services to map the semantic workflow into the representations of other systems and run in other platforms used in other labs. The mappings to new platforms will be supported by a novel set of workflow services that can be instantiated by the community with additional workflow systems and software platforms.

The proposed framework would illustrate important benefits to both publishers and authors: 1) improve the review process, because workflows would allow reviewers to easily check the claims made in an article; 2) improve the content of the articles, because workflows would describe unequivocally the computational aspects of the methods used in the research; 3) improve the utility of published research as articles, because workflows would allow other researchers to re-execute the computational methods published in the article with minimal effort; and 4) extend the shelf life of an article by abstracting from implementation details that become obsolete faster than the research methods themselves.

We will start by developing a prototype of this framework to demonstrate that a workflow created in a lab using a workflow system (probably Wings), which exports

the provenance and abstract workflow as Linked Open Data, can be mapped to a different workflow system or infrastructure used in another lab (to be selected as the work progresses) and re-executed.

Ultimately, we would like to measure how individual researchers and laboratories benefit from this novel form of publication. Usability would improve, as lab members would interact with a proper workflow system rather than handling manually large low-level scripts. Productivity would improve, as lab members would take little time to incorporate into their work the workflows from other researchers. Exploration would increase, as scientists could easily mix and match components and change the workflow, perhaps serendipitously plugging in an untried component and discovering something unusual.

All the software will be available open source. All ontologies and models will be released under appropriate Creative Commons licenses. Documentation will be provided to allow the community to extend the framework to include other workflow systems and software infrastructure.

We will focus on an article that can have broad potential impact [Kinnings et al 11]. The original work did not use a workflow system, instead the computational steps were run separately and manually. The article describes a computational pipeline that accesses data from the Protein Data Base (PDB) and carries out a systematic analysis of the proteome of an organism against all FDA-approved drugs. The process uncovers protein receptors in an organism that could be targeted by drugs currently in use for other purposes. The result is a drug-target network (a "drugome") that includes all known approved drugs. Although the article focuses on Mycobacterium tuberculosis (TB), the method itself can be used for other pathogens or pathways and has the potential to be a key resource to develop new more comprehensive treatments for other diseases of interest.

The availability of the method in this particular drugome article as a workflow is in itself of significant value. First, the deeper significance of this article is that it represents a novel method for drug discovery that takes a comprehensive and systematic approach, moving away from current practice which is neither. Second, the workflow is composed of a number of computations that are common in PDB, such as finding drug binding sites in proteins, comparing ligand binding sites, molecular docking, and clustering of protein and drug binding profiles. Therefore, this work can have potential impact on the PDB community by allowing the publication and reuse of computational workflows within this large community. PDB is perhaps the most prominent and oldest database in the biomedical community (http://www.pdb.org). It has been available for 40 years and has 200,000 monthly users. Third, the article was selected as an exemplar in the BeyondThePDF workshop (http://sites.google.com/site/beyondthepdf) and can have an impact in the growing research community interested in structuring scientific articles (e.g., [Groth et al 10]).

References

Baggerly, K. A. and Coombes, K. R. "Deriving Chemosensitivity from Cell Lines: Forensic Bioinformatics and Reproducible Research in High-Throughput Biology." Annals of Applied Statistics, 3(4), 2009. *Available from* *http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.aoas/1267453942*

Bell AW, Deutsch EW, Au CE, Kearney RE, Beavis R, Sechi S, Nilsson T, Bergeron JJ, and the Human Proteome Organization (HUPO) Test Sample Working Group. "A HUPO test sample study reveals common problems in mass spectrometry–based proteomics." Nature Methods, 6(6), 2009. *Available from* *http://www.nature.com/nmeth/journal/v6/n6/full/nmeth.1333.html*

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*

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

Gil, Y.; Ratnakar, V.; Kim, J.; Gonzalez-Calero, P. A.; Groth, P.; Moody, J.; and Deelman, E. "Wings: Intelligent Workflow-Based Design of Computational Experiments." IEEE Intelligent Systems, 26(1), 2011. *Preprint available from* *http://www.bibbase.org/cache/www.isi.edu_7Egil_publications.bib/gil-etal-ieee-is-11.html*

Groth, P., Gibson, A., and Velterop, J. "The Anatomy of a Nanopublication." Information Services and Use, 30(1-2), 2010. *Available from* *http://iospress.metapress.com/content/ftkh21q50t521wm2/*

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.; 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, 2011. *Preprint available*

from http://www.bibbase.org/cache/www.isi.edu_7Egil_publications.bib/gil-etal-jetai10.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*

Hutson, S. "Data Handling Errors Spur Debate Over Clinical Trial," Nature Medicine, 16(6), 2010. *Available from http://www.nature.com/nm/journal/v16/n6/full/nm0610-618a.html*

Ioannidis JP, Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, Falchi M, Furlanello C, Game L, Jurman G, Mangion J, Mehta T, Nitzberg M, Page GP, Petretto E, van Noort V. "Repeatability of Published Microarray Gene Expression Analyses." Nature Genetics, 41(2), 2009. *Available from http://www.nature.com/ng/journal/v41/n2/full/ng.295.html*

Kinnings, S. L.; Xie, L.; Fung, K. H.; Jackson, R. M.; Xie, L.; and Bourne, P. E. "The *Mycobacterium tuberculosis* Drugome and Its Polypharmacological Implications." To appear in PLoS Computational Biology, 2011. *Preprint available from http://sites.google.com/site/beyondthepdf/file-cabinet/FinalPaper.pdf?attredirects=0&d=1*

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*

Leisch, F. "Sweave: Dynamic Generation of Statistical Reports Using Literate Data Analysis", Proceedings of Computational Statistics, 2002. *Preprint available from http://www.statistik.lmu.de/~leisch/Sweave/Sweave-compstat2002.pdf*

Mesirov, J. P. "Accessible Reproducible Research." Science, 327:415, 2010. *Available from http://www.sciencemag.org/cgi/rapidpdf/327/5964/415?ijkey=WzYHd6g6IBNeQ&keytype=ref&siteid=sci*

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*

Nature Editorial. "Illuminating the Black Box." Nature, 442(7098), 2006. *Available from http://www.nature.com/nature/journal/v442/n7098/full/442001a.html*

Scientific American. "In Science We Trust: Poll Results on How you Feel about Science" Scientific American, October 2010. *Available from http://www.scientificamerican.com/article.cfm?id=in-science-we-trust-poll*

The Scientist. "Top Retractions of 2010." The Scientist, December 16, 2010. *Available from http://www.the-scientist.com/news/display/57864/*