EDUCAUSE Review, vol. 43, no. 5 (September/October 2008)

E-Content

Curating Scientific Web Services and Workflows

Carole Goble and David De Roure

Carole Goble is a full professor in the School of Computer Science, University of Manchester, United Kingdom, and is Director of the ^{my}Grid e-Science consortium. David De Roure is a full professor in the School of Electronics and Computer Science, University of Southampton, United Kingdom.

Comments on this article can be sent to the authors at <carole.goble@manchester.ac.uk> and <dder@ecs.soton.ac.uk> and/or can be posted to the web via the link at the bottom of this page.

A bewildering array of digital resources is available to the modern researcher, ranging from libraries of articles and data collections to analytical tools and visualization applications, many publicly available. Take bioinformatics, for example. *Nucleic Acids Research* describes more than 1,000 databases[1] drawn from a "Bioinformatics Nation"[2] of different subdisciplines, research teams, and institutes. The same is true for chemistry, astronomy, earth sciences, and just about any information-rich scientific area. These digital resources are combined and their data aggregated and analyzed in the day-to-day work of skilled scientific investigators. But even though we are familiar with the need to curate our data for dissemination and for the long term, we must not neglect the curation and cataloguing of the *processes* that we use to search, integrate, and analyze that data.

Where researchers may once have used applications or software libraries, increasingly we see functionality provided instead through web-based resources. In nanotechnology, for example, nanoHUB (http://www.nanohub.org) provides more than 1,000 resources for research, education, and collaboration—including simulation tools accessed from the web browser. Hence the web has become a distributed computing platform supporting research on an everyday basis.

In this article we will highlight two specific kinds of processes—web services and workflows—that are enjoying increasing adoption. They have already become prevalent in the *in silico* research of life scientists, providing a revealing glimpse into the future. Web services offer a well-defined programming interface that software applications, written in various programming languages and running on various platforms, can use to process data over the Internet. An increasing number of resources are available via web services interfaces, turning these resources into services that can be combined into complex networked applications. In the life sciences, open source and commercial integration systems, data warehouses, and integration frameworks use web services behind the scenes.

Workflows are an alternative to using precooked applications, with embedded data pipelines and analysis scripts. Scientific workflow management systems—such as Taverna, Triana, Kepler, and Pipeline Pilot—provide a mechanism to automatically orchestrate the execution of services, coordinating processes (control flow) and managing the flow of data between them (data flow).[3] The workflows are explicit and precise descriptions of a scientific process—the instruction scripts that define the flow of data and the order of execution of the service steps. In turn, these workflows can become services within other workflows and applications.

Workflows are becoming rather fashionable. As more resources become available, exposed as web services, they provide an attractive means for rapid assembly of customized integrations. They link together and cross-reference data in different repositories, both public and private, which could be widely distributed. For example, workflows can assist in automatically text-mining the literature. From a developer's standpoint, they are an agile means of application delivery of a process. From a scientific programmer's standpoint, they are a means to automatically, repetitively, and systematically run a process while accurately tracking the provenance of results. From a scientist's standpoint, they are a reliable and transparent means for encoding a scientific method that supports reproducible science and the sharing and replicating of best-of-practice and know-how through reuse.

Curating Processes

Processes are the methods that form a core component of scientific discovery. Given a predicted rise in the number of openly available web services and workflows, it would seem necessary, and certainly prudent, to curate processes as effectively as we curate the data they consume and the publications they generate. The systematic curation of processes would enable programmers and scientists to survey available, well-characterized, and established methods, to avoid unnecessary reinvention, and to be better informed of best-practice techniques and how they are used correctly and appropriately. The lack of adequate and standard metadata describing individual services often prevents their discovery unless users already know that these services exist, know what they do, and know how to use them.

We should be able to

- find a process based on what it does (or was meant to do), what it consumes as inputs and produces as outputs, and find copies or similar services usable as alternates,
- understand how and when it works, how to operate and configure it correctly with some examples and defaults, and how to predict its performance properties,
- know the conditions for use: permissions, licenses, platforms, and costs (financially or in-storage or computational resource),
- judge the benefits of adoption based on its reputation (its popularity and known use cases), its provenance (its source and history), and its validation by peers,
- estimate the risk of adoption based on its reliability and stability, and

- get assistance for its incorporation into applications and workflows.

Both web services and workflows need accurate and flexible metadata that is understandable both by people and by software applications. However, the comprehensive cataloguing needed to serve the broader research community, beyond project-specific efforts, is lacking. Web services and workflows are scattered across the web. They are most likely to be located by word-of-mouth or by Google searches, which find them through textual references. Groups or individuals gather them on websites or portals. Broader initiatives such as seekda (http://www.seekda.com) gather together a very wide range of web services but are insufficiently curated. Yet curation is crucial:

- Web services tend to be poorly described, often with documentation that is insufficient or inappropriate; they lack basic information, such as semantic descriptions to tell a prospective user what parameters mean, how they should be used, or what their data formats should be. Workflows are defined using system-specific files (although BPEL is a standardized language, it is not widely adopted in science) without standards for their documentation.
- Reports of operational behavior and current standing have no central place to be gathered, processed, and disseminated.
- The change in availability, interfaces, permissions, performance, and other basic properties of public processes needs constant vigilance.

Although some curators are domain experts who understand web services and workflows, we see two other key approaches. One is community curation: the trend is to follow in the footsteps of popular Web 2.0 social computing sites and encourage community curation through user feedback, blogging, e-tracking, recommendations, and folksonomy-based tagging. Community curation requires built-in incentive models, such as credit and attribution, for people to contribute. Second, operational and usage metadata is ripe for automation, generated from monitoring services, application diagnostics, customer reports, and social network analysis. *Workflow analytics* is the term used for processing workflow collections to identify, for example, service co-use patterns and service popularity.

Two Approaches

We are tackling these challenges in two efforts to systematically catalogue processes for the benefit of specific scientific communities:

- *myExperiment* (http://www.myexperiment.org) is a social networking and scientific workflow repository that emphasises community participation for workflow developers to upload their workflows for the common good of the community.[4] Although first developed for the Taverna workflow community, it is intended as a resource for many workflow systems, recently incorporating Triana workflows. In particular, myExperiment has addressed the issues of sharing and credit, with an underlying model to support author attribution, versioning, and cross-workflow reuse analysis.

- *BioCatalogue* (http://biocatalogue.org), an offshoot of myExperiment, incorporates the experiences of the Taverna service registry. The aim is to improve process reuse, reliability, and validation by encouraging self-curation and community curation alongside automated-curation and expert-curation pipelines managed by the European Bioinformatics Institute, a major service provider with a reputation in scientific curation. By generating content that can be indexed by third-party information providers, such as Google, BioCatalogue should make the content easy to find. By presenting programmable APIs, this effort should make the catalogue easy to mashup and to incorporate into third-party applications. As a consequence, BioCatalogue aims to provide the missing part of the equation that yields science from data integration.

Conclusion

We have an increasing understanding of the practices of data curation, but we should not neglect the curation and cataloguing of the *processes* that we use to work with the data. A well-curated resource would potentially enable reuse by including knowledge of and about processes, and would hence avoid wasteful reinvention, increase reliability by pooling operational histories and reputations, and improve validation by promoting best-practice, verified procedures and popular processes. However, an absence of curated processes leads to ignorance of availability and creates obstacles to adoption. Active curation of these resources with accurate and flexible descriptions to check their availability, reliability, and general quality of service is required. Community curation and automation provide a powerful approach to addressing these challenges.

Notes

1. Michael Y. Galperin, "The Molecular Biology Database Collection: 2008 Update," November 19, 2007 (online), *Nucleic Acids Research,* vol. 36 (2008), <http://nar.oxfordjournals.org/cgi/reprint/36/suppl_1/D2>.
2. Lincoln Stein, "Creating a Bioinformatics Nation," *Nature,* vol. 417 (May 9, 2002), pp. 119–20, <http://www.nature.com/nature/journal/v417/n6885/full/417119a.html>.
3. Yolanda Gil, Ewa Deelman, Mark Ellisman, Thomas Fahringer, Geoffrey Fox, Dennis Gannon, Carole Goble, Miron Livny, Luc Moreau, and Jim Myers, "Examining the Challenges of Scientific Workflows," *Computer,* vol. 40, no. 12 (December 2007), pp. 24–32.
4. David De Roure, Carole Goble, and Robert Stevens, "Designing the myExperiment Virtual Research Environment for the Social Sharing of Workflows," *E-SCIENCE '07: Proceedings of the Third IEEE International Conference on e-Science and Grid Computi*ng*, Bangalore, India, December 10–13, 2007* (Washington, D.C.: IEEE Computer Society, 2007), pp. 603–10.