The Impact of Workflow Tools on Data-centric Research

Carole Goble, University of Manchester

David De Roure, University of Southampton

We are in an era of data-centric scientific research, where hypotheses are not only tested through directed data collection and analysis but are also generated by combining and mining the pool of data already available [1-3]. The scientific data landscape we draw upon is expanding rapidly in both scale and diversity. Taking the life sciences as an example, high-throughput gene sequencing platforms are capable of generating terabytes of data in a single experiment, and data volumes are set to increase further with industrial-scale automation. Between 2001 and 2008 the number of databases reported in the Life Sciences' annual review jumped from 96 to 1070 [4]. Not only are the datasets growing in size and number, they are only partly coordinated and often incompatible [5], which means discovery and integration tasks are significant challenges. At the same time, we are drawing on a broader array of data sources: modern biology draws insights from combining different types of "omic" data (proteomic, metabolomic, transcriptomic, genomic) as well as data from other disciplines such as chemistry, clinical medicine, and public health; Systems Biology links multi-scale data with multi-scale mathematical models. These data encompass all types: from structured database records to published articles, raw numeric data, images, and descriptive interpretations that use controlled vocabularies.

Data generation on this scale must be matched by scalable processing methods. The preparation, management, and analysis of data are bottlenecks and also beyond the skill of many scientists. Workflows [6] provide: (i) a systematic and automated means of conducting analyses across diverse datasets and applications; (ii) a way of capturing this process so that results can be reproduced and the method can be reviewed, validated, repeated and adapted; (iii) a visual scripting interface so that computational scientists can create these pipelines without low-level programming concerns ; and (iv) an integration and access platform for the growing pool of independent resource providers so that computational scientists need not specialise in each one. The workflow is thus becoming a paradigm for enabling science on a large scale by managing data preparation and analysis pipelines, and the preferred vehicle for computational knowledge extraction.

Workflows Defined

A workflow is a precise description of a scientific procedure–a multi-step process to coordinate multiple tasks, like a sophisticated script [7]. Each task represents the execution of a computational process, such as running a program, submitting a query to a database, submitting a job to a compute cloud or grid, or invoking a service over the Web to use a remote resource. Data output from one task is consumed by subsequent tasks according to a predefined graph topology that "orchestrates" the flow of data. Figure 1 presents an example workflow, encoded in the Taverna Workflow Workbench [8], which searches for genes by linking four publicly available data resources distributed in the USA, Europe and Japan: BioMart, Entrez, Uniprot, and KEGG.

Workflow systems broadly have three components: an execution platform, a visual design suite and a development kit. The platform executes the workflow on behalf of applications and handles common crosscutting concerns, including: (a) *invocation* of the service applications, handling the heterogeneity of data types and interfaces on multiple computing platforms; (b) *monitoring and*

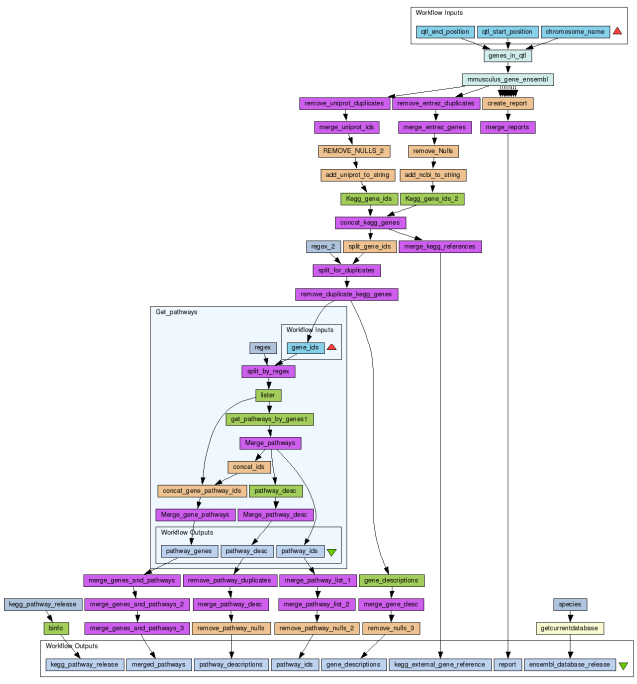

Figure 1. A Taverna workflow that connects several internationally distributed datasets to identify candidate genes that could be implicated in resistance to African trypanosomiasis [11].

recovery from failures; (c) *optimization* of memory, storage and execution including concurrency and parallelization; (d) *data handling*: mapping, referencing, movement, streaming and staging; (e) *logging* of processes and data provenance tracking; and (f) *security* and monitoring of access policies. Workflow systems are required to support long-running processes in volatile environments and thus must be robust and capable of fault tolerance and recovery. They also need to evolve continually to harness the growing capabilities of underlying computational and storage resources, delivering greater capacity for analysis.

The design suite provides a visual scripting application for authoring and sharing workflows and preparing the components that are to be incorporated as executable steps. The aim is to shield the author from the complexities of the underlying applications and enable them to design and understand workflows without recourse to commissioning specialist and specific applications or hiring software engineers. This empowers scientists to build their own pipelines when they need them and how they want them. Finally, the development kit enables developers to extend the capabilities of the system and enables workflows to be embedded into applications, web portals or databases. This embedding is transformational: it has the potential to incorporate sophisticated knowledge seamlessly and invisibly into the tools that scientists use routinely.

Each workflow system has its own language, design suite, and software components, and the systems vary in their execution models and the kinds of components they coordinate [9]. Sedna is one of the few to use the industry-standard Business Process Execution Language (BPEL) for scientific workflows [10]. General purpose open source Workflow systems include Taverna, Kepler, Pegasus, and Triana. Other systems, such as the LONI Pipeline for neuroimaging and the commercial Pipeline Pilot for drug discovery, are more geared toward specific applications and are optimized to support specific component libraries. These focus on interoperating applications; other workflow systems target the provisioning of compute cycles or submission of jobs to Grids. For example, the Pegasus and DAGMan have been used for a series of large-scale e-science experiments such as prediction models in earthquake forecasting using sensor data in the Southern California Earthquake Center (SCEC) CyberShake project.

Workflow Usage

Workflows liberate scientists from the drudgery of routine data processing so they can concentrate on scientific discovery. They shoulder the burden of routine tasks, they represent the computational protocols needed to undertake data-centric science, and they open up the use of processes and data resources to a much wider group of scientists and scientific application developers.

Workflows are ideal for systematically, accurately, and repeatedly running routine procedures: managing data capture from sensors or instruments; cleaning, normalising and validating data; securely and efficiently moving and archiving data; comparing data across repeated runs; regularly updating data warehouses. For example, the Pan-STARRS astronomical survey uses Microsoft's Trident system workflows to load and validate telescope detections running at about 30 TB per year. Workflows have also proved useful for maintaining and updating data collections and warehouses by reacting to changes in the underlying datasets. For example, the Nijmegen Medical Centre rebuilt the tGRAP G-protein coupled receptors mutant database using a suite of text-mining Taverna workflows.

At a higher level, a workflow is an explicit, precise, and modular expression of an *in silico* or "dry lab" experimental protocol. Workflows are ideal for gathering and aggregating data from distributed datasets and data-emitting algorithms—a core activity in dataset annotation, data curation, and multi-evidential, comparative science. In Figure 1, disparate datasets are searched to find and aggregate data related to metabolic pathways implicated in resistance to African trypanosomiasis; interlinked datasets are chained together by the dataflow. In this instance, the automated and systematic processing by the workflow overcame the inadequacies of manual data triage—which leads to prematurely excluding data from analysis to cope with the quantity—and delivered new results [11].

Beyond data assembly, workflows codify data mining and knowledge discovery pipelines and parameter sweeps across predictive algorithms. For example, LEAD workflows are driven by external events generated by data mining agents that monitor collections of instruments for significant patterns to trigger a storm prediction analysis and the Jet Propulsion Laboratory uses Taverna workflows for exploring a large space of multiple parameter configurations of space instruments.

Finally, workflow systems liberate the implicit workflow embedded in an application into an explicit and reusable over a common software machinery and shared infrastructure. Expert informaticians use workflow systems directly as means to develop workflows for handling infrastructure; expert *scientific* informaticians

use them to design and explore new investigative procedures; a larger group of scientists uses precooked workflows with restricted configuration constraints launched from within applications or hidden behind Web portals.

Workflow-Enabled Data-centric Science

Workflows offer techniques to support the new paradigm of data-centric science. They can be replayed and repeated. Results and secondary data can be computed as needed using the latest sources, providing virtual data (or on-demand) warehouses by effectively providing distributed query processing. *Smart reruns* of workflows automatically deliver new outcomes when fresh primary data and new results become available—and also when new methods become available. The workflows themselves, as first-class citizens in data-centric science, can be generated and transformed dynamically to meet the requirements at hand. In a landscape of data in considerable flux, workflows provide robustness, accountability, and full auditing. By combining workflows and their execution records with published results, we can promote systematic, unbiased, transparent and comparable research in which outcomes carry the provenance of their derivation. This can potentially accelerate scientific discovery.

To accelerate experimental *design*, workflows can be reconfigured and repurposed as new components or templates. Creating workflows requires expertise that is hard won and often outside the skill set of the researcher. Workflows are often complex and challenging to build because they are essentially forms of programming that require some understanding of the datasets and the tools they manipulate [12]. Hence there is significant benefit in establishing shared collections of workflows that contain standard processing pipelines for immediate reuse or for repurposing in whole or in part. These aggregations of expertise and resources can help propagate techniques and best practices. Specialists can create the application steps, experts can design the workflows and set parameters, and the inexperienced can benefit by using sophisticated protocols.

The myExperiment [13] social web site has demonstrated that by adopting content-sharing tools for repositories of workflows, we can enable social networking around workflows and provide community support for social tagging, comments, ratings and recommendations, and mixing of new workflows with those previously deposited. This is made possible by the scale of participation in data-centric science, which can be brought to bear on challenging problems. For example, the environment of workflow execution is in such a state of flux that workflows appear to decay over time, but workflows can be kept current by a combination of expert and community curation.

Workflows enable data-centric science to be a collaborative endeavor on multiple levels. They enable scientists to collaborate over shared data and shared services granting non-developers access to sophisticated code and applications without having to install and operate them. Consequently scientists can use the best applications, not just the ones with which they are familiar. Multidisciplinary workflows promote even broader collaboration. In this sense, a workflow system is a framework for reusing a community's tools and datasets that respects the original codes and overcomes diverse coding styles. Initiatives such as the BioCatalogue registry of life science Web services and the component registries deployed at SCEC enable components to be discovered. In addition to the benefits that come from explicit sharing, there is considerable value in the information that may be gathered just through monitoring the use of data sources, services, and methods. This enables automatic monitoring of resources and recommendation of common practice and optimization.

Although the impact of workflow tools on data-centric research is potentially profound—scaling processing to match the scaling of data—many challenges exist over and above the engineering issues inherent in large-scale distributed software [14]. There are a confusing number of workflow platforms with various capabilities and purposes and little compliance to standards. Workflows are often difficult to author, using languages that are at an inappropriate level of abstraction and expecting too much knowledge of the underlying infrastructure. The reusability of a workflow is often confined to the project it was conceived in—or even to its author—and it is inherently only as strong as its components. Although workflows encourage providers to supply clean, robust, and validated data services, component failure is common. If the services or infrastructure decay, so does the workflow. Unfortunately debugging failing workflows is a crucial but neglected topic. Contemporary workflow platforms fall short of adequately supporting rapid deployment into the user applications that consume them, and legacy application codes need to be integrated and managed.

Conclusion

Workflows affect data-centric research in four ways. First, they shift scientific practice. For example, in a data-driven hypothesis [15], data analysis yields results that are to be tested in the laboratory. Second, they have the potential to empower scientists to be the authors of their own sophisticated data processing pipelines without having to wait for software developers to produce the tools they need. Third, they offer systematic production of data that is comparable and verifiably attributable to its source. Finally, people speak of a data deluge [13], and data-centric science could be characterized as being about the primacy of data as opposed to the primacy of the academic paper or document [16], but it brings with it a method deluge: workflows illustrate *primacy of method* as another crucial paradigm in data-centric research.

References

[1] D. B. Kell and S. G. Oliver, "Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era," *BioEssays*, vol. 26, no. 1, pp. 99–105, 2004.

[2] A. Halevy, P. Norvig, and F. Pereira, "The Unreasonable Effectiveness of Data," *IEEE Intell. Syst.*, vol. 24, no. 2, pp. 8–12. 2009.

[3] C. Anderson, "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete," *Wired*, vol. 16, no. 7, June 23, 2008, www.wired.com/science/discoveries/magazine/16-07/pb_theory.

[4] Galperin MY. The Molecular Biology Database Collection: 2007 update. Nucl Acids Res. 2007 January 12, 2007;35(suppl_1):D3-4.

[5] Goble C, Stevens R (2008), The State of the Nation in Data Integration in Bioinformatics, J Biomed Inform. 41(5):687-93.

[6] I. J.Taylor, E. Deelman, D. B. Gannon, and M. Shields, Eds., *Workflows for e-Science: Scientific Workflows for Grids*. London: Springer, 2007.

[7] P. Romano, "Automation of in-silico data analysis processes through workflow management systems", *Brief Bioinform*, vol. 9, no. 1, pp. 57–68, Jan. 2008.

[8] T. Oinn, M. Greenwood, M. Addis, N. Alpdemir, J. Ferris, K. Glover, C. Goble, A. Goderis, D. Hull, D. Marvin, P. Li, P. Lord, M. Pocock, M. Senger, R. Stevens, A. Wipat, and C. Wroe, "Taverna: lessons in creating a workflow environment for the life sciences," *Concurrency and Computation: Practice and Experience*, vol. 18, no. 10, pp. 1067–1100, 2006.

[9] E. Deelman, D. Gannon, M. Shields, and I. Taylor, "Workflows and e-Science: An overview of workflow system features and capabilities," *Future Gen. Comput. Syst.*, vol. 25, no. 5, pp. 528–540, May 2009.

[10] B. Wassermann, W. Emmerich, B. Butchart, N. Cameron, L. Chen, and J. Patel, "Sedna: a BPEL-based environment for visual scientific workflow modelling," in I. J. Taylor, E. Deelman, D. B. Gannon, and M. Shields, Eds., *Workflows for e-Science: Scientific Workflows for Grids*. London: Springer, 2007, pp. 428–449.

[11] P. Fisher, C. Hedeler, K. Wolstencroft, H. Hulme, H. Noyes, S. Kemp, R. Stevens, and A. Brass, "A Systematic Strategy for Large-Scale Analysis of Genotype-Phenotype Correlations: Identification of candidate genes involved in African Trypanosomiasis," *Nucleic Acids Res.*, vol. 35, no. 16, pp. 5625–5633, 2007.

[12] A. Goderis, U. Sattler, P. Lord, and C. Goble, "Seven Bottlenecks to Workflow Reuse and Repurposing in The Semantic Web," *ISWC 2005*, pp. 323–337.

[13] D. De Roure, C. Goble, and R. Stevens, "The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows," *Future Gen. Comput. Syst.* vol. 25, pp. 561–567, 2009.

[14] Y. Gil, E. Deelman, M. Ellisman, T. Fahringer, G. Fox, D. Gannon, C. Goble, M. Livny, L. Moreau, and J. Myers, "Examining the Challenges of Scientific Workflows," *Computer*, vol. 40, pp. 24–32, 2007.

[15] G. Bell, T. Hey, and A. Szalay, "Computer science. Beyond the data deluge," *Science*, vol. 323, no. 5919, pp. 1297–1298, Mar. 6, 2009.

[16] G. Erbach, "Data-centric view in e-Science information systems," *Data Sci. J.*, vol. 5, pp. 219–222, 2006.