The impact of workflow tools on data-centric research

Carole Goble and David De Roure

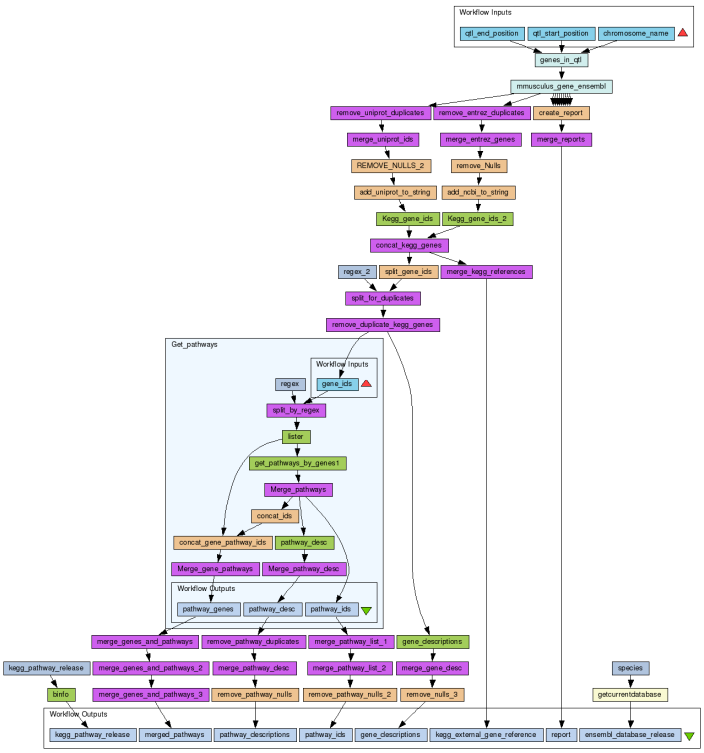
The scientific data landscape is expanding rapidly both in scale and diversity. For example, in the Life Sciences data sources are growing in number and are only partly coordinated [Romano], so discovery and integration tasks are significant. High throughput gene sequencing platforms are capable of generating terabytes of data in a single experiment, and data volumes are set to increase further with the trend to industrial-scale automation. Science itself draws more broadly across data sources: modern bioinformatics draws insights from combining "omic" data (proteomic, metabolomic, transcriptomic, genomic) as well as data from other disciplines such as chemistry, clinical medicine and public health. Data embraces all types: structured database records to published articles of text; raw numeric data to descriptive interpretations using controlled vocabularies through to images. We are in an era of data-centric scientific research, where hypotheses are not only tested through directed data collection and analysis but also generated from combining and mining the pool of data already available [Kell].

Data generation on this scale must be matched by scalable processing methods. Preparation, management and analysis of data have become rate limiting steps. Workflows [Taylor] provide a systematic means of processing data and capturing that process for reuse. Workflow Management Systems (WfMSs) are evolving to provide the capacity increase in the analysis pipeline by harnessing increasing capability of underlying computational and storage resources. Thus workflow is becoming established as a paradigm for systematically and automatically managing the data preparation and analysis pipelines, and the preferred vehicle for computational knowledge extraction and for enabling science at a large scale.

A workflow is a precise description of a scientific procedure – a multi-step process to coordinate multiple tasks. Each task represents the execution of a computational process, such as running a program, a query to a database, a submission of a job to a compute cloud or Grid, or the invocation of service over the Web to use a remote resource. Data flows from the outputs of one task to be consumed by subsequent tasks according to a pre-defined graph topology which "orchestrates" the flow of data. WfMSs provide a platform that executes the workflow on behalf of applications and handles common cross-cutting concerns: memory, storage and execution optimisation, concurrency and parallelisation, monitoring, debugging, process logging and data provenance tracking, data movement and data reference management, data streaming and staging policies, security, service invocation, failure recovery, deployment over different platforms etc. They are required to support long-running processes in volatile environments and thus must be robust and capable of fault tolerance and recovery in the face of error. WfMSs also provide a design suite for authoring and sharing workflows and preparing the components that are to be incorporated as executable steps.

Open Source WfMSs include Taverna (www.taverna.org.uk), Kepler (kepler-project.org), Pegasus (pegasus.isi.edu) and Triana (www.trianacode.org). Figure 1 presents an example workflow encoded in the Taverna WfMS. These are general systems capable of assimilating a range of components from different disciplines (biology, chemistry, astronomy, earth sciences etc) over various distributed computing infrastructures such as the Web, the Cloud or the Grid. WfMSs such as the LONI pipeline for neuroimaging (pipeline.loni.ucla.edu) and the commercial Pipeline Pilot (accelrys.com/products/scitegic) for drug discovery are geared towards specific applications and optimised to support specific libraries of pre-prepared

components. Pegasus (pegasus.isi.edu) and DAGman (www.cs.wisc.edu/condor/dagman) have been used for a series of large-scale e-Science experiments, such as the provisioning of compute cycles to hazard curves prediction models for earthquake rupture forecasts using sensor data in the SCEC Cybershake project (epicenter.usc.edu/cmeportal/CyberShake.html). Each WfMS has its own language, design suite and software components and vary in their execution model and the kinds of components they coordinate [Deelman]. Sedna is one of the few to use the industry standard Business Process Execution Language (BPEL) for scientific workflows [Wassermann].



This workflow searches for genes which reside in a Quantitative Trait Loci (QTL) region in the mouse, Mus musculus. The workflow requires an input of: a chromosome name or number; a QTL start base pair position; QTL end base pair position. Data is then extracted from BioMart to annotate each of the genes found in this region. The Entrez and UniProt identifiers are then sent to KEGG to obtain KEGG gene identifiers. The KEGG gene identifiers are then used to search for pathways in the KEGG pathway database.

Figure 1: A Taverna workflow that chains together several internationally distributed datasets to identify candidate genes implicated in the parasitic disease

Workflows shoulder the burden of systematically, accurately and repeatedly running routine and often complex and laborious tasks. These tasks can be experiment or data specific such as gathering data from sensors or other instruments, cleaning, validation, pre-processing, and normalisation. For example the Pan-STARRS astronomy survey (pan-starrs.ifa.hawaii.edu) uses Microsoft's Trident system workflows (www.microsoft.com/mscorp/tc/trident.mspx) to load and validate telescope detections running ~30TB/year. Workflows have generally proved a well-suited method of keeping data collections and warehouses current, reacting to changes in the underlying data sets. The Nijmegen Medical Centre cleaned up the tGRAP G-protein coupled receptors mutant database using a suite of text mining Taverna workflows. Pipelines for data movement, aggregation and integration assist not only the consumer in assimilating new content but also the data service provider in providing clean, robust and validated data services, stimulating the need for community standards in data formats and interfaces.

At a higher level, a workflow is an explicit, precise and modular expression of an *in silico* or "dry lab" experimental protocol, not just for data assembly but also for codifying data mining, knowledge discovery pipelines and parameter sweeps across predictive algorithms. For example, LEAD workflows (portal.leadproject.org) are driven by external events generated by data mining agents monitoring collections of instruments for significant patterns to trigger a storm prediction analysis. 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 Trypanosomosis; interlinked data sets are chained together by the dataflow. In this instance, the automated and systematic processing by the workflow overcame the inadequacies of manual data triage – prematurely excluding data from analysis to cope with the quantity – and delivered new results [Fisher].

Workflows are not new. In the "pre-industrial-scale data" era, they were laborious manual processes, fraught with error and shortcut temptations. Processes were hard to replicate and results were hard to compare or interpret in the absence of accurate logs or data provenance tracking. Customised applications embed specific workflows within software which automates consistent processing but often at the cost of process transparency, and they are hard to adapt or to reuse in other applications. Script-driven software separates the scripted process from the application, making it explicit and open to configuration and scrutiny. However, scripting solutions require that the application handles the cross-cutting "plumbing" activities that a WfMS manages on an application's behalf, creating a means to share workflows between applications and a flexible mechanism for rapid application development.

WfMSs liberate the implicit workflow embedded in an application into a specification that is explicit and reusable over a common software machinery and shared infrastructure. Workflows also have the potential to liberate the scientist by dealing with the drudgery of routine data processing and freeing the scientist to concentrate on the science and accelerate the creation of results. Expert informaticians use the WfMSs directly as means to develop workflows for handling infrastructure; expert scientific informaticians use WfMS to design and explore new investigative procedures, while a larger tranche of scientists use pre-cooked workflows with restricted configuration constraints launched from within applications or hidden behind web portals.

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 and when 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 dynamically to meet the requirements at hand. In a landscape of data in considerable flux, workflows provide us with robustness, accountability and full audit. By combining workflows and their execution records with published results we have a means to promote transparent and comparable research where outcomes carry the provenance of their derivation, with potential acceleration of scientific discovery.

To accelerate experimental *design*, workflows are reusable with reconfiguration and repurposable as new components or templates. Creating workflows requires specialist

expertise that is hard-won and may be outside the skill-set of the researcher. They are often complex and challenging to build [Goderis], because they are essentially forms of program that require some understanding of the datasets and tools they manipulate. Hence there is significant benefit in establishing shared collections of workflows, containing standard processing pipelines for immediate reuse or for repurposing in whole or part. They represent collaborations of people and resources – aggregations of expertise. If exchanged, they are a means of propagating technique and best practice: specialists create the application steps; experts design workflows and set parameters; and the inexperienced punch above their weight by using sophisticated protocols.

The myExperiment project (www.myexperiment.org) has demonstrated that by adopting social content sharing tools for repositories of workflows we can harness a social infrastructure that enables social networking around workflows and provides community support for social tagging, comments, ratings and recommendations, social network analysis and reuse mining (what is used with what, for what and by whom), and remixing of new workflows from previously deposited ones. This is made possible by the scale of participation in data-centric science, which is a powerful instrument we may also bring to bear on challenging problems that do not yield to engineering solutions. 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.

In fact workflows enable data-centric science to be a collaborative endeavour at multiple levels. They enable scientists to collaborate over shared data and over shared services; for example, Taverna gives non-developers access to sophisticated codes and applications, without the need to install and operate them, and enables a scientist to use the best applications and not just the ones they know. Multi-disciplinary workflows promote even broader collaborations. In this sense a WfMS is a framework to reuse a community's tools and datasets that respects the original codes and overcomes heterogeneous coding styles. Initiatives such as the BioCatalogue of Life Science Web Services (www.biocatalogue.org) and the component registries deployed in 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 usage of data sources, services and methods: this enables monitoring of resources and recommendation of common practice and optimisation.

Although the impact of workflow tools on data-centric research is potentially profound – scaling processing to match the scaling of data – there are many challenges over and above engineering issues inherent in large-scale distributed software [Gil, Deelman]. Today it is necessary to select from a confusing number of workflow platforms, of various capabilities and purposes. Workflows are often challenging and expensive to author and run, with languages often at an inappropriate level of abstraction requiring too much knowledge of 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, which can be difficult and expensive to produce; if the services or infrastructure decay so does the workflow, and debugging failing workflows is a neglected but crucial issue. Contemporary workflow platforms fall short of adequately supporting rapid deployment into the user applications that consume them, and legacy application codes have to be integrated and managed.

In summary, workflows have a four-fold impact on data-centric research. The first is the shift in scientific practice; for example, in data-driven hypothesis [Kell], data analysis yields results to be tested in the laboratory. The second is the potential for empowering scientists to be the authors of their own sophisticated data processing pipelines without waiting for software developers to produce the tools they need. The third is the systematic production of data that is comparable and verifiably attributable to its source. Finally, people speak of a data deluge [Bell], and data-centric science could be characterised as being about the primacy of data as opposed to the primacy of the academic paper or document [Erbach], but it brings with it a method deluge: workflows illustrate primacy of method as another crucial paradigm in data-centric research.

References

P. Romano, "Automation of in-silico data analysis processes through workflow management systems." Brief Bioinform, vol. 9, no. 1, pp. 57-68, January 2008.

Douglas B. Kell, Stephen 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 26(1) pp. 99-105, 2004.

Taylor, I.J.; Deelman, E.; Gannon, D.B.; Shields, M. (Eds.) Workflows for e-Science Scientific Workflows for Grids 2007, ISBN: 978-1-84628-519-6

Wassermann, B., Emmerich, W., Butchart, B., Cameron, N., Chen, L. and Patel, J. Sedna: a BPEL-based environment for visual scientific workflow modelling. In: Taylor, I.J. and Deelman, E. and Gannon, D.B. and Shields, M., (eds.) Workflows for e-Science. Springer London, London, UK, pp. 428-449, 2007.

Ewa Deelman, Dennis Gannon, Matthew Shields and Ian Taylor Workflows and e-Science: An overview of workflow system features and capabilities Future Generation Computer Systems Volume 25, Issue 5, May 2009, pp. 528-540.

Fisher, P., Hedeler, C., Wolstencroft, K., Hulme, H., Noyes, H., Kemp, S., Stevens, R. & Brass A. A Systematic Strategy for Large-Scale Analysis of Genotype-Phenotype Correlations: Identification of candidate genes involved in African Trypanosomiasis. Nucleic Acids Res. 2007;35(16):5625-33. 2007.

Goderis, A., Sattler, U., Lord, P. & Goble, C. Seven Bottlenecks to Workflow Reuse and Repurposing in The Semantic Web – ISWC 2005 pp. 323-337, 2005.

Gil, Y., Deelman, E., Ellisman, M., Fahringer, T., Fox, G., Gannon, D., Goble, C., Livny, M., Moreau, L. & Myers, J. Examining the Challenges of Scientific Workflows, Computer. 40, pp. 24-32, 2007.

Bell G, Hey T, Szalay A. Computer science. Beyond the data deluge. Science. 2009 Mar 6;323(5919):1297-8.

Gregor Erbach, Data-centric view in e-Science information systems, Data Science Journal Vol. 5 pp.219-222, 2006.