e-Science and the Web

David De Roure,
University of Southampton, UK

e-Science projects have generated new ways of thinking, new expertise and methods, and a new collaborative infrastructure.

In the last decade science has experienced a step change in problem-solving ability, brought about by the increasing digitization and automation of scientific practice. We call this e-Science, and the Web plays a crucial role in its success.

The e-Science revolution has partly resulted from the deluge of data from new experimental methods and technologies. These include high-throughput DNA sequencing, combinatorial chemistry, industrial-scale laboratory automation, sensor networks, and Earth observation.

Data collection is fast and parallel, and our rapidly evolving high-performance computing infrastructure must be able to tackle the resulting data tsunami to enable scientific discovery. If researchers have to triage data to a manageable search space, they might miss the right pieces or fail to see patterns in the bigger picture. The preponderance of data is causing a shift in scientific methods, from scientific hypotheses driving the collection of data to data driving the formation of hypotheses.

This can be characterized as the "Big Science" view of e-Science: scientists working with massive computational power and volumes of data to achieve breakthroughs in the modeling of everything from storms to earthquakes to fly brains to nanoscale transistors.

To harness the distributed resources required for this unprecedented scale of operation, the cyberinfrastructure was born. Big Science is perhaps best exemplified by the grid infrastructure for CERN's Large Hadron Collider, where 300 Gbytes per second of raw data from detectors filters down to tens of Tbytes per day for distribution to thousands of physicists around the world.

Big Science is a great success, but it's only part of the story. There's another revolution going on.

SCIENCE ON THE WEB

Digitization is changing how we do science, as if we've invented new scientific instruments, but it's also changing who uses these techniques. It's not just the heroic few: go into any department on a university campus today, whatever the discipline, and you'll see some aspect of research being conducted on computers.

The adoption of wikis, blogs, and collaborative Web technologies in the pursuit of science has ushered in Science 2.0. Moreover, the "open science" movement, which advocates that the methodology, data, and results of experiments be freely available, encourages massively distributed collaboration. In fact, we should say "research" rather than science, because the Web is agnostic about research discipline: it's as much a home for digital arts and humanities as digital science and engineering.

In a further democratizing step, the Web enables citizens to participate more directly in research. "Citizen scientists" count birds in their backyards or report on astronomical phenomena in the nighttime sky, providing a new data collection instrument for scientists and a new scale of distributed expertise. For the social scientist, the Web is a powerful survey device—for example, crowdsourcing is being used for real-time geographic surveys on everything from the impact of the credit crunch to congestion charging—as well as an opportunity to obtain more data than ever before on how people interact. Studying this data is establishing new methods for scientists that in turn yield insights that inform both science and society.

WHEN INFRASTRUCTURES COLLIDE

Sometimes there seem to be separate initiatives: a top-down creation and rollout of cyberinfrastructure versus the natural evolution of the Web ecosystem, with high-end

Published by the IEEE Computer Society

0018-9162/10/$26.00 © 2010 IEEE

researchers using cyberinfrastructure and the long tail using the Web. Big Science practitioners combat the learning curve of advanced technologies, while everyday researchers and amateur scientists simply choose what's available and useful.

It may not have been their intention, but infrastructure providers tend to take a "build it and they will come" approach, and then wonder why people don't come. The answer, according to several UK studies, is that users have been neglected: the software, service, and tool providers must think about rolling in users, not just rolling out technology. Some initiatives may be guilty of adopting a "technological determinism" viewpoint—that the inexorable progress of technology shapes how we do science—but in fact, scientists and technology coshape research tools and techniques, and this symbiosis has flourished on the Web.

To fully realize its potential, technology must be as easy as possible to use—we need "access ramps" for users and developers. The Web is the biggest, most successful, and most programmable distributed systems architecture ever. It's the favored means for disseminating and discovering information, for collaboration, and increasingly for distributed applications. It buzzes with content and programs created by both experts and novices.

Domain-specific computing specialists can readily mold the Web to meet specialists' requirements, but how do we extend Big Science thinking and capability so that any researcher can do data-intensive research? One way is through utility computing—computational resources on demand, like electricity—which is increasingly being realized through cloud computing. This was one of the grid's original goals, and another is the notion of virtual organizations: flexible assemblies of resources and people to meet the needs at hand.

To be empowered, researchers must have that power of assembly, and therein lies a key e-Science challenge: How can researchers assemble resources and express those assemblies for reproducible and repurposable research?

SERVICE-ORIENTED SCIENCE

Our infrastructure and middleware efforts have partly been driven by the vision of a massively service-oriented future—that one day we will choose from millions of services and compose them dynamically to tackle research problems. We're well on our way: the seekda Web service catalogue (http://webservices.seekda.com) carries 28,000 services from more than 7,000 providers, and in the life sciences domain, BioCatalogue (www.biocatalogue.org) provides a registry curated by service providers, experts, and users.

Scientists and technology coshape research tools and techniques, and this symbiosis has flourished on the Web.

Scientific workflow systems provide a means of composing these services, to conduct in silico experiments and data analysis pipelines. The various workflow systems that have emerged from e-Science are some of its most successful outcomes, catering to Big Science as well as empowering individual researchers in labs around the globe.

Workflows remove the drudgery of routine manual processing, deliver systematic pipelines to deal with the data deluge, provide a repeatable experimental record to facilitate interpretation and reuse, and enable scientists to share experimental methods. Meanwhile, workflow systems liberate workflow designers from low-level programming concerns and deal with the increasing numbers of services and resources, at the same time generating a research agenda in large-scale service description and matchmaking.

As service-oriented architecture (SOA) provisioning matures, another assembly technology has come into the ascendant: the mashup. Critics of mashups argue that unlike workflows, which are well-engineered, declarative templates that capture processes for reuse, mashups are fragile, imperative hacks for human consumption. In fact, both artifacts are fragile: they don't decay but rather fail because the surrounding service landscape (whether accessed in REST or SOAP) is in flux. Further, they aren't competing solutions for data integration but solve two different but important problems: workflows bundle services together for reuse in the emerging landscape of increasing scale and automation, while mashups are a powerful means of rapid application assembly to assist scientists.

SCIENTIST-ORIENTED SCIENCE

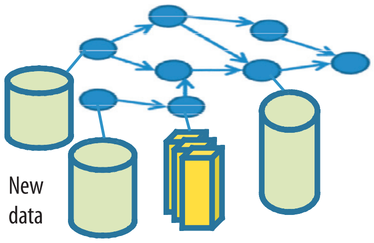
In 2000, John Taylor, director general of Research Councils at the UK Office of Science and Technology, described e-Science as "global collaboration in key areas of science and the next generation of infrastructure that will enable it" (www.rcuk.ac.uk/escience/news/firstphase.htm). Climate change research, for example, must interlink data, models, and expertise in disparate areas from atmospheric chemistry and soil science to hydrology and oceanography.

The key to any collaboration, with people we do or don't know, is sharing information, techniques, and expertise. Some of the tools for sharing are already in Web users' hands, but just because the tools exist doesn't mean scientists will use them.

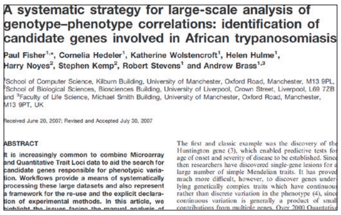
One project that recognizes the social dimension of e-Science is the social website myExperiment (www.myexperiment.org). Codesigned by

Paul Fisher is a bioinformatician studying disease in African cattle.

1 Paul **designs** a workflow and **executes** it over shared Web services.



2 Paul **publishes** the workflow and results on the Web and the paper online.



A systematic strategy for large-scale analysis of genotype–phenotype correlations: identification of candidate genes involved in African trypanosomiasis

Paul Fisher[1,*], Cornelia Hedeler[1], Katherine Wolstencroft[1], Helen Hulme[1], Harry Noyes[2], Stephen Kemp[2], Robert Stevens[1] and Andrew Brass[1,3]

[1]School of Computer Science, Kilburn Building, University of Manchester, Oxford Road, Manchester, M13 9PL, [2]School of Biological Sciences, Biosciences Building, University of Liverpool, Crown Street, Liverpool, L69 7ZB and [3]Faculty of Life Science, Michael Smith Building, University of Manchester, Oxford Road, Manchester, M13 9PT, UK

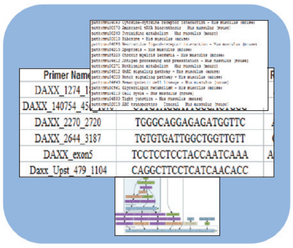
Received June 20, 2007; Revised and Accepted July 30, 2007

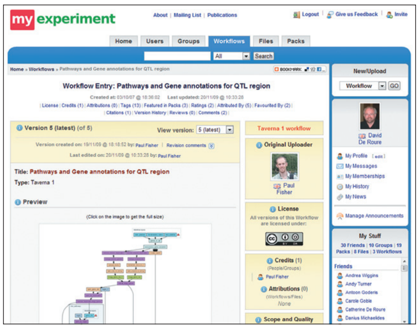
ABSTRACT

It is increasingly common to combine Microarray and Quantitative Trait Loci data to aid the search for candidate genes responsible for phenotypic variation. Workflows provide a means of systematically processing these large datasets and also represent a framework for the re-use and the explicit declaration of experimental methods. In this article, we highlight the issues facing the manual analysis of

The first and classic example was the discovery of the Huntington gene (3), which enabled predictive tests for age of onset and severity of disease to be established. Since then researchers have discovered single-gene lesions for a large number of simple Mendelian traits. It has proved much more difficult, however, to discover genes underlying genetically complex traits which have continuous rather than discrete variation in the phenotype (4), since continuous variation is generally a product of small contributions from multiple genes. Over 2000 Quantative

3 The data and workflow are **discovered** by others for **reuse** in other areas of science.





myexperiment

About | Mailing List | Publications

Logout | Give us Feedback | Invite

Home | Users | Groups | Workflows | Files | Packs

All | Search

Home » Workflows » Pathways and Gene annotations for QTL region

Workflow Entry: Pathways and Gene annotations for QTL region

Created at: 03/10/07 @ 18:36:02 Last updated: 20/11/09 @ 10:33:28

| License | Credits (1) | Attributions (0) | Tags (13) | Featured in Packs (3) | Ratings (2) | Attributed By (5) | Favourited By (2) | Citations (1) | Version History | Reviews (0) | Comments (2) |

Version 5 (latest) (of 5) View version: 5 (latest)

Version created on: 19/11/09 @ 18:18:52 by: Paul Fisher | Revision comments

Last edited on: 20/11/09 @ 10:33:28 by: Paul Fisher

Title: Pathways and Gene annotations for QTL region

Type: Taverna 1

Preview

(Click on the image to get the full size)

Taverna 1 workflow

Original Uploader

Paul Fisher

License

All versions of this Workflow are licensed under:

Credits (1)

(People/Groups)

Paul Fisher

Attributions (0)

(Workflows/Files)

None

Scope and Quality

New/Upload

Workflow GO

David De Roure

My Profile [edit]

My Messages

My Memberships

My History

My News

Manage Announcements

My Stuff

30 Friends | 10 Groups | 19 Packs | 8 Files | 3 Workflows

Friends

Andrea Wiggins

Andy Turner

Antoon Goderis

Carole Goble

Catherine De Roure

Danius Michaelides

4 The workflow is tagged, reviewed, and **curated** by its user community and by specialists.

Figure 1. e-Science in action. The sharing of methods builds reputation and enables community curation in data-intensive science.

scientists, myExperiment uses a Web 2.0 approach to provide a medium where scientists can safely publish their workflows and other artifacts and share them with others. To meet the unique needs of its user base, the website provides support for credit, attributions, licensing, and fine control over privacy—all of which are essential to researchers. Significantly, the scale of user participation brings the prospect of social curation of workflows to combat the inexorable problem of decay.

myExperiment could have been yet another repository to share anything, but it instead chose to offer a service for which there was an urgent need. Building good workflows is difficult, especially in a diverse and distributed community, and the site tackles this head-on. As scientists share new objects on myExperiment—from experimental plans for a chemistry lab to scripts and statistical models—the focus remains on methods, as Figure 1 shows. This is intrinsic to the incentives that enable the site to succeed: by sharing methods, researchers gain expertise and reputation, and the community gains in shared know-how and new capacity.

There is an e-Science message for repositories, too: in a world slowly embracing data curation, myExperiment provides an approach for curating methods. The data deluge brings a method deluge, too, and this valuable resource must not be neglected.

RECORD AND REUSE

At some level, e-Science is fundamentally about recording information, be it data from devices or results of experiments, and then reusing it. The big challenge is making it available for both anticipated and unanticipated reuse.

A particularly exciting opportunity has grown up alongside e-Science. The "linked data" movement, emerging alongside the Semantic Web, has established guidelines to make it as easy as possible to connect related data that wasn't previously linked. Not only is there an increasing number of public data providers using linked data, but the tooling for consuming it is improving—a researcher can now easily build a script or workflow that draws upon multiple data sources and integrates them.

"Record and reuse" has been achieved through academic papers up until now, and they're very usable by humans; in fact, they're increasingly read by machine, too, with growing sophistication. But what is their digital equivalent? It isn't a PDF, but rather the sharable collection of data and methods to support the emerging scholarly knowledge cycle of data-intensive and open research.

myExperiment and related "e-laboratory" projects suggest that records of research should be

- *Replayable*—go back and see what happened. Whether observing the planet, the population, or an automated experiment, data collection can occur over milliseconds or months. The ability to replay the experiment, and to focus on crucial parts, is essential to understanding what happened.

- *Repeatable*—run the experiment again. There must be enough information for the original researcher or others to be able to repeat the experiment, perhaps years later, to verify the results or validate the experimental environment. This also helps scale to the repetition of processing demanded by data-intensive research.
- *Reproducible*—conduct an independent experiment to reproduce the results. Others should be able to start with the description of the experiment and see if they can replicate the results. This is one of the scientific method's central tenets.
- *Reusable*—use as part of new experiments. One experiment may call upon another, and assembling methods in this way makes it possible to conduct research, and ask research questions, at a higher level.
- *Repurposable*—reuse the pieces in a new experiment. A black box experiment is only reusable as a black box. By opening the lid we find parts (and combinations of parts) available for reuse, and the way they're assembled is a clue to their reassembly.
- *Reliable*—be robust under automation. This applies to the robustness of science provided by systematic processing with humans out of the loop, and to the comprehensive handling of failure demanded in complex systems where success may be the exception and not the norm.

How do we achieve this? In the open repositories world, a new standard called Object Reuse and Exchange is using RDF (Resource Description Framework) graphs to describe collections of things—like all the pieces that make up an experiment—even if they're distributed across the Web. Hence we're moving toward self-describing, digital scholarly artifacts, and before long, these are what researchers will share.

The term e-Science emphasizes scientific ambitions: success isn't measured by the uptake of technologies but rather research outcomes and the impact they have on our understanding of the universe, the discovery of new drugs, or social policy. Using the Web, e-Science projects have—like moonshots—generated new ways of thinking, new expertise and methods, and a new collaborative infrastructure of shared services, data, and software.

David De Roure is a professor of computer science in the School of Electronics and Computer Science at the University of Southampton, UK. Contact him at dder@ecs.soton.ac.uk.

Editor: Simon S.Y. Shim, Dept. of Computer Engineering, San Jose State Univ., San Jose, CA; simon.shim@sjsu.edu