e-Science and the Web

by David De Roure, July 2009 (revised 9 August 2009)

As science progresses we witness evolution and revolution in scientific understanding. Coupled intricately with this is an evolution and revolution in scientific techniques and methods: our ability to solve scientific problems advances as new methods lead to new understanding, and this in turn generates new methods. In the last 10 years science has experienced a step change in problem-solving ability, brought about by the increasing digitisation and automation of scientific practice. This has in part been achieved through the engagement between scientists and providers of advanced techniques in data, computation, and communication. We call this e-Science, and the Web plays a crucial role in its success.

This revolution in technique and method has partly come about due to the deluge of data from new experimental methods. These include high-throughput DNA sequencing technologies, combinatorial chemistry, industrial-scale laboratory automation, sensor networks, and Earth observation. Data collection is fast and parallel, and we need our rapidly evolving high performance computing infrastructure to tackle the resulting data tsunami. If we fail then scientific discoveries will be missed, because we have to triage data to a manageable search space and then we might miss the right pieces or fail to see patterns in the bigger picture. Thus science is accelerated and practice evolves: the preponderance of data is causing a change in scientific methods, from scientific hypothesis driving the collection of data to data driving the formation of hypotheses.

This can be characterised as the "Big Science" view of e-Science: scientists working with heroic computational power and volumes of data, targeting breakthroughs in the modelling of everything from storms and earthquakes to fly brains and nanoscale transistors. Out of these needs was born the cyberinfrastructure to harness the distributed resources needed for this scale of operation. This approach is famously exemplified by the Grid infrastructure for the incredible data output from the Large Hadron Collider at CERN, where 300GB per second of raw data from detectors is filtered down to tens of terabytes per day for distribution to thousands of physicists around the world. Big Science is an important part of the story and a great success, but it is only part of the story. There is another revolution going on.

Science on the Web

This lurch into the digital world is changing the way we do science, as if we've invented some new scientific instruments, but importantly it's also changing *who* can use these techniques. It's not just the heroic few. Go into any department on a university campus today and, whatever the discipline, some aspect of research will be conducted on a computer. This scale of participation is new, and so is the breadth of disciplines that are touched. The next generation of researchers is the Web generation and they've never known life without Google, brought up in a digital world of collaborative tools and sharing. The power that results from this scale of participation, and the potential for new research across disciplines, is phenomenal.

Researchers have new digital tools and artefacts to work with, and some of our science is massively more collaborative and multidisciplinary. The traditional scholarly knowledge lifecycle handles papers and some supplemental materials, yet does not deal with the new objects of digital scholarship or the scale of collaboration. Naturally we turn to other solutions. With the adoption of Wikis, Blogs and collaborative Web technologies in the pursuit of science we have entered the era of *Science 2.0*. Moreover the *Open Science* movement shares the spirit of open source and advocates that the methodology, data and results

of experiments should be freely available, permitting massively distributed collaboration. In fact, we should say “research” rather than science, because the Web is agnostic about research discipline: it is as much a home for digital arts and digital humanities as digital science and engineering.

In a further democratising step, the Web provides a means for citizens to participate more directly in research. It brings new opportunities in data collection: thousands of people have counted birds in their backyards, and crowd-sourcing is being used for real-time geographic surveys from credit crunch to congestion charging. For the scientist this is indeed a new instrument, and for the social scientist it is an amazing survey device as well as giving more data than ever before on how people are interacting. Studying this data is establishing new methods for social scientists and in turn giving insights to inform science and society.

When infrastructures collide

Sometimes these seem like separate initiatives: a top-down creation and rollout of cyberinfrastructure versus the natural evolution of the Web ecosystem, with high-end researchers using cyberinfrastructure and the long tail using the Web. The practitioners of Big Science have been resourced to combat the learning curve of advanced technologies, while everyday researchers simply choose what is available and useful. This may not have been the intention, but infrastructure providers take a “build it and they will come” approach, 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 need to think about rolling-in users and not just rolling out technology. Some initiatives are perhaps guilty of adopting a “technological determinism” viewpoint – that the inexorable progress of technology is shaping how we do science – when in fact our research tools and techniques are co-shaped by scientists and technology, and this co-constitution has flourished on the Web.

The moral of this story is that the technology must be as easy as possible to use, and scientists should feel empowered to do so with the same fluency they enjoy with the other apparatus of their professions. This explains the pattern of the technology adoption. The Web is simply the biggest, most successful, most usable and most programmable distributed systems architecture ever. It is the favoured infrastructure for disseminating and discovering information, for collaboration and increasingly for distributed applications. It is buzzing with content and programs created by experts and novices. Domain-specific computing specialists can readily mould it to meet the requirements of their science users: it is a perpetual beta e-Infrastructure that meets scientists’ needs in an agile fashion.

So how do we bring some of that big science thinking, capability and resource to the everyday researcher? One way is utility computing – processor power on demand like electricity – and to a growing extent this is being realised by cloud computing, which sits very comfortably in the Web ecosystem. This was one of the original visions of the Grid, and another is the essential notion of *virtual organisations* – flexible assemblies of resources and people to meet the needs at hand. This assembly story is key. For researchers to be empowered we must give them that power of assembly, and therein lies one of the most important computer science challenges in e-Science today: how can researchers assemble resources, and how do they express those assemblies for reproducible and repurposable research?

Service-oriented science

Our infrastructure and middleware efforts have partly been driven by a vision of a massively service-oriented future – that one day we will choose from millions of services and compose them

dynamically to tackle our research problems. We are on our way: the SeekDa service catalogue carries 28,000 Web Services from over 7,000 providers, and in the life sciences domain the Biocatalogue provides a registry which is curated by service providers, experts and users.

Scientific workflow systems give us a means of composing these services, to conduct *in silico* experiments and data analysis pipelines. The various workflow systems that have emerged from the e-Science community are perhaps one its most successful outcomes, catering for big science as well as empowering individual researchers scattered in labs around the globe. Workflows are powerful at multiple levels: they relieve the scientist from the drudgery of routine manual processing, deliver systematic pipelines to deal with the data deluge, provide a repeatable record of the experiment to facilitate interpretation and reuse, and enable scientists to share their experimental methods. Meanwhile the workflow systems liberate the workflow designer 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 we step toward this greater maturity in SOA provision we see another assembly technology in the ascendant: the mashup. The apparent collision of workflows and mashups as competing solutions for data integration has been an interesting debate, in which workflows are portrayed as well-engineered declarative templates which capture processes for reuse, while mashups are seen as fragile imperative hacks for human consumption. In fact, both artefacts are fragile: they don't decay but rather fail because the service landscape around them (whether accessed in REST or SOAP) is in flux. Furthermore they 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

Now let's look at this from the perspective of the scientists entering this world of new resources, services, tools and techniques – and new challenges. An early definition of e-Science described it as "global collaboration in key areas of science and the next generation of infrastructure that will enable it". An excellent example is climate change research, in which we need to interlink data, models and expertise in previously disparate areas from atmospheric chemistry and soil science to hydrological models and the oceans. We've looked at data and computation, but what about the social dimension? The key to collaboration, be it local or global, with people we know or people we don't, is sharing information, techniques and expertise, and some of the tools for sharing are already in the hands of the users of the Web. But just because the tools exist doesn't mean scientists will use them. Again it is a socio-technical issue, and we tackle it by going on a journey with the scientists.

One such journey is the myExperiment project, a social web site for scientists which has been codesigned with its users. myExperiment has successfully adopted a Web 2.0 approach in delivering a social web site where scientists can safely publish their scientific workflows and other artefacts, share them with groups and find those of others. While it shares many characteristics with other Web 2.0 sites, myExperiment's distinctive features to meet the needs of its research user base are support for credit, attributions, licensing, and fine control over privacy – all of which are essential for the research users. Very 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 set up as yet another repository to share anything, but it chose to focus on a service for which there was an urgent need. Building good workflows is difficult, especially in a diverse and distributed community, and myExperiment tackled this head-on. As new objects are shared on myExperiment – from experimental plans for the chemistry lab through to scripts and statistical models – it has maintained a focus on *methods*. This is intrinsic to the incentives that enable the site to succeed: by sharing methods the researchers gain expertise and reputation, and the community gains in shared know-how and new capacity. There is an e-Science message there for repositories too: in a world slowly embracing data curation, myExperiment provides an approach for curating methods. This is important: the data deluge brings a method deluge too, and this valuable resource must not be neglected.

Record and reuse

At some level, much of 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 from 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 a workflow which draws upon multiple data sources and integrates them.

"Record and reuse" is what academic papers have done up till now, and they are very usable by humans; in fact they are increasingly read by machine too, with growing sophistication. But what is their digital equivalent – not a PDF, but rather the sharable collection of data and methods to support the emerging scholarly knowledge cycle of data-intensive and open research? Research in myExperiment and related "e-laboratory" projects suggests that records of research should have six key properties:

- **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 for human understanding of what happened.
- **Repeatable** – run the experiment again. Enough information for the original researcher or others to be able to repeat the experiment, perhaps years later, in order to verify the results or validate the experimental environment. This also helps scale to the repetition of processing demanded by data intensive research.
- **Reproducible** –an independent experiment to reproduce the results. To reproduce (or replicate) a result is for someone else to start with the description of the experiment and see if a result can be reproduced. This is one of the tenets of the scientific method as we know it.
- **Reusable** – use as part of new experiments. One experiment may call upon another, and by assembling methods in this way we can conduct research, and ask research questions, at a higher level.
- **Repurposable** – reuse the pieces in a new experiment. An experiment which is a black box 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 are assembled is a clue to how they can be reassembled.
- **Reliable** – robust under automation. This applies to the robustness of science provided by systematic processing with human-out-the-loop, and to the comprehensive handling of failure demanded in complex systems where success may be the exception not the norm.

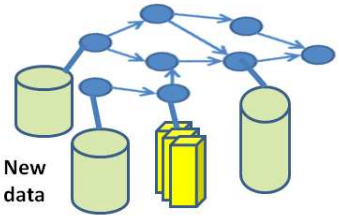
How do we achieve this? Again the Semantic Web has spawned a solution. In the Open Repositories world, a new standard called Object Reuse and Exchange is using RDF (Resource Description Framework) graphs to describe collections of things – like all the pieces that make up an experiment – even if they are distributed across the Web. Hence we move towards self-describing, digital scholarly artefacts, and before long it is these that researchers will share rather than their papers.

Conclusion

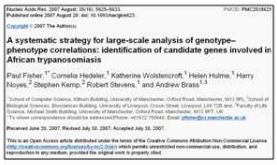
The term e-Science was chosen to emphasise scientific ambitions: the real measure of success of e-Science is not the uptake of the technologies but rather the new research outcomes and the impact these have in fundamental understanding of the universe, discovery of new drugs or changes in social policy from climate change to health. On their way to these outcomes, e-Science projects have – like moonshots – generated new ways of thinking, new expertise and methods, a new collaborative infrastructure of shared services, data and software and a Pandora's box of research questions. The Web is a fantastic melting pot for all of this, an ecosystem where cyberinfrastructure, citizens and scientists collaborate and compete, and where society and technology meet in creating new instruments and new outcomes with completely new means of impact.

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

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

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

e-Science in action: the sharing of methods builds reputation and enables community curation in data-intensive science.