myExperiment – A Web 2.0 Virtual Research Environment

David De Roure
Electronics and Computer Science
University of Southampton, UK
dder@ecs.soton.ac.uk
+44 23 8059 2418

Carole Goble
School of Computer Science
The University of Manchester, UK
carole.goble@manchester.ac.uk
+44 161 275 6195

ABSTRACT
e-Science has given rise to new forms of digital object in the Virtual Research Environment which can usefully be shared amongst collaborating scientists to assist in generating new scientific results. In particular, descriptions of Scientific Workflows capture pieces of scientific knowledge which may transcend their immediate application and can be shared and reused in other experiments. We are building the *myExperiment* Virtual Research Environment to support scientists in sharing and collaboration with workflows and other objects. Rather than adopting traditional methods prevalent in the e-Science developer community, our approach draws upon the social software techniques characterised as Web 2.0. In this paper we report on the preliminary design work of myExperiment.

Author Keywords
Scientific Workflow, social networking, Bioinformatics, Taverna, myGrid.

ACM Classification Keywords
H.3.4 [Information Storage and Retrieval]: – Systems and Software; H.4.2 [Information Systems Applications]: – Types of Systems; D.2.2 [Software Engineering]: – Interoperability; J.3 [Life and Medical Sciences]: – Biology and genetics.

INTRODUCTION

e-Science was defined at the launch of the UK e-Science programme as being "about global collaboration in key areas of science and the next generation of infrastructure that will enable it" (John Taylor, Director General of Research Councils). The techniques of e-Science help the scientist deal with increasingly large and increasingly complex scientific applications. Key to this is automation, and several scientific workflow tools have become established as a means of automating the processing of scientific data in a scalable and reusable way.

The myExperiment Virtual Research Environment (VRE) provides a personalised environment which enables users to share, re-use and repurpose experiments. Our vision is that scientists should be able to swap workflows and other scientific objects as easily as citizens can share documents, photos and videos on the Web. Hence myExperiment owes far more to social networking websites such as MySpace (www.myspace.com) and YouTube (www.youtube.com) than to the traditional portals of Grid computing, and is immediately familiar to the new generation of scientists. Where many e-Science projects have focused on bringing computational resources to bear on "reducing time to discovery", we take a holistic view of the scholarly knowledge cycle and focus on reducing "time-to-experiment" and "time-to-citation".

In the next section we describe the nature of workflows and their use within one scientific community, bioinformatics. We then discuss the social context and summarise the results of our design scoping exercise for myExperiment. Finally we review myExperiment against Web 2.0 design patterns. We close by suggesting that other VREs could usefully conduct a similar review.

THE TAVERNA EXPERIENCE

The UK's myGrid project [6] has developed the popular Taverna workflow workbench [5,9], used throughout the world for a whole range of Life Science problems: gene and protein annotation; proteomics, phylogeny and phenotypical studies; microarray data analysis and medical image analysis; high throughput screening of chemical compounds and clinical statistical analysis. Taverna is now part of the Open Middleware Infrastructure Institute UK (http://www.omii.ac.uk) portfolio of supported software development, so that e-scientists can rely upon it as part of their regular collection of tools.

Importantly, Taverna has been designed to operate in the "open wild world" of bioinformatics. For example, the services are expected to be owned by parties other than those using them in a workflow. They are volatile, weakly described and there is no contract in place to ensure quality of service; they have not been designed to work together,

and adhere to no common type system. By compensating for these demands, Taverna has made over 3500 operations available to its users. This has been a major incentive to adoption. Thus, the success of Taverna – measured for example by 30,000 downloads to April 2007 – has largely come about by understanding the needs, fears and reward incentives of its different users (service providers, tool developers and bioinformaticians), working "in the wild".

Workflows are an important new object in science and there is clear evidence of the scientific value of reusing them [14]. We observe that:

- Workflows are valuable knowledge assets in their own right, capturing valuable know-how that is otherwise often tacit.
- Workflows are challenging and expensive to develop – realistic workflows require skill to produce. Consequently, workflow developers need development assistance, and prefer not to start from scratch.
- The reusability of a workflow is often confined to the project it was conceived in, and there are social and technical challenges for workflow discovery, sharing and reuse.
- Workflows and their outcomes need to be bound with their provenance if they are to be trusted and if they are to be interpreted unambiguously and reused accurately. The provenance is often confined to the system from which it originated.
- Workflows matter more than workflow platforms. Users are driven by content not platforms: they will adopt workflows that have the capabilities they need regardless of the platform that executes them.
- Workflows are beginning to be shared on Web pages and Wikis – a recent workflow harvest using Google returned over 400 different workflows publicly available.

It is interesting to note from this that in many ways workflows share characteristics with programs or scripts.

SOCIAL SOFTWARE MOTIVATION

Current e-Science infrastructures provide the capability to combine services from a diverse set of providers in a variety of ways. However, they can only be exploited by a minority of specialists who are familiar with workflow composition systems, programming paradigms, distributed infrastructures and complex problem solving environments.

Many sophisticated individuals and companies are in great need of sharing knowledge and resources, collaborating and generating value-added services – but without the technical expertise they are disenfranchised. Existing communities of practice have the potential to achieve this but lack a means for doing so.

The rise in the "Socio-Web", and now the "Social Grid" has dramatically reminded us that it is people who generate and share knowledge and resources, and people who create network effects in communities. Blog and wikis, shared tagging services, instant messaging, social networks, semantic descriptions of data relationships, etc. are flourishing. By mining the sharing behaviour between users within such a community we can provide recommendations of use. By utilising the structure and interactions between users and workflow tools we can identify what is considered to be of greater value to users. Annotation and indexing may be enhanced by semantic techniques; e.g. [1].

While the technologies and approaches characterised as "Web 2.0" include social networking, there is also a technological perspective which has a profound synergy with myExperiment – it is the relationship between workflows and mashups. We can now see the Web as a planet-wide distributed application platform with a "software-as-a-service" mentality, empowering end users to "mash-up" syndicated content on demand, just in time, by themselves, and to share the results and the mechanisms. This liberation of content and application development creates a vibrant social effect and dramatically accelerates application capability through community network effects. Workflows are, inherently, both mash-ups and content syndication feeds. The culture and practice around code creation and sharing in Web 2.0 is itself a model for workflow sharing in myExperiment.

DESIGNING MYEXPERIMENT

We have held three workshops leading to the initial design of myExperiment: a "portal party" [7] with end-users to establish requirements, followed by two design scoping workshops coupled with presentations from specific end-user groups – we are starting with the life sciences and then extending to chemistry, astronomy and social sciences. The scoping workshops were based on the vision that the myExperiment Virtual Research Environment enables scientists to be (more) creative and to be scientists not programmers. We suggested that the following four requirements must be met for the myExperiment VRE to succeed:

1. It is a social networking environment for sharing any scientific workflows and associated data so that scientists can build on the work of others;
2. It should be very easy to use, effective, extensible and to a large degree self-sustaining (from a support viewpoint);
3. It should be integrated and interoperable, so that workflows can be launched from within the environment – a feature of *in silico* science;
4. It should integrate with the scholarly publishing process.

We summarise the discussions under the following four "design dimensions".

1. Workflow Warehouse or Federation of Repositories?

In one model, myExperiment could be a Web site with its own workflow repository, either constructed as a completely new site or by tailoring existing solutions such as Media Wiki. Alternatively, the various objects (workflows, data, provenance records) could be maintained in distributed repositories. A myExperiment Web site is then

just one of many possible interfaces to this content. In this respect, myExperiment is going beyond what we know today as Web 2.0 because data is often restricted and the software supporting many Web 2.0 sites is proprietary.

We have chosen to build a Web site which can store workflows, thus providing a standalone solution, and which can also participate in a federated repository model. This is achieved through metadata harvesting and repository interoperability protocols such as the Open Archives Initiative (OAI) [10]. This builds on the experience of the publishing ethos of CombeChem [3] and using OAI with scientific data in eBank-UK [4].

2. Social Space or Shoe Shop?

There are accounts of Wikis growing organically in response to the demand and creativity of their users, to the point where their size and ad hoc organisation causes difficulties in performance and navigation. Such an example of growth is OpenWetWare [11], which grew from a Wiki for a lab to support multiple labs in one institution and then transcended the institution.

In contrast, some Web Sites are highly organised and designed to make it very easy for people to find what they want. In a shopping site, for example, the catalogue is carefully maintained while the collective social benefit comes from reviews and recommendations.

The quality and character of descriptive information needed varies according to its function. We decided that social tagging (cf flickr) will assist workflow discovery, but that some aspects of workflows needs to be rigorously described due to the scientific context and for automated use.

Recommender techniques will also help people find workflows in myExperiment; e.g. Amazon-style "People who used this workflow also used this…" or lastfm.com style usage of "workflow playlists". As well as the workflows, this social networking information needs to be handled in a federated manner across multiple myExperiment instances.

3. How open is the content?

The power of myExperiment comes from sharing, but there is clear evidence through existing lab practices that not all users will wish, or indeed be able, to make everything available to everyone. In contrast it is interesting to note that the OpenWetWare experience has created a culture where everything is open; moreover, this is part of its value proposition in the face of competing solutions.

The Web of course has exactly these issues, and its value also comes from open content (e.g. indexable by Google can be aggregated in mashups). Initiatives such as Creative Commons [2] and Science Commons [12] are relevant here.

We decided to support a spectrum of sharing from exposing a workflow for access by others, to giving it to others, to publishing it across a boundary into a group or public domain.

We note that some of these issues may attract new solutions in the context of our work. By tracking provenance, we have a machine-processable record that can assist in mechanisms to deal with ownership and authorship.

4. Integration

Users with no existing mechanism for sharing workflows may welcome a public myExperiment site where they can find and publish workflows. Others may already be publishing workflows on wiki pages in their lab. Should we oblige this latter group to change their practices, or can we bring "myExperiment-ness" to their existing environment? One extreme definition of "using myExperiment" could simply be to work with a core set of file formats and metadata attributes through existing applications.

We decided to provide a public site as well as software for people to build their own "myExperiments", and to make it as easy as possible for existing solutions, such as Wikis, to interact with myExperiment – for example through plug-ins that access the services behind myExperiment.

DESIGN PATTERNS

Having conducted this design exercise driven from user requirements and with an awareness of the Web 2.0 social and technical synergies, how do we measure up against Web 2.0 design patterns? [8]

1. The Long Tail

Our target users are not just the specialist e-Scientists using computing resources to tackle major scientific breakthroughs, but also the large number of scientists conducting the routine processes of science on a daily basis. Through sharing we have the potential to enable smart scientists to be smarter and propagate their smartness, in turn enabling other scientists to become better and conduct better science.

2. Data is the Next "Intel Inside"

myExperiment understands that scientists are focused on data, not software or one particular workflow engine. Workflows are components of customised applications, many of which are data-oriented rather than process-oriented. Users manipulate, through their own applications, the product (data, model) yielded by the workflow. Furthermore, workflows themselves are the data of myExperiment and provide its unique value.

3. Users Add Value

myExperiment makes it easy to find workflows and is designed to make it useful and straightforward to share workflows and add workflows to the pool. To succeed we draw on the insights into the incentive models of scientists gained through experience with Taverna.

4. Network Effects by Default

myExperiment aggregates user data as a side-effect of using the VRE. The ability to execute workflows from myExperiment, and the integration of tools such as Taverna with myExperiment, further enable us to achieve increased value through usage.

5. Some Rights Reserved

myExperiment users require protection as well as sharing, but the environment is designed for maximum ease of sharing to achieve collective benefits – workflows are "hackable" and "remixable". Initiatives such as Science Commons provide a useful context for this.

6. The Perpetual Beta

myExperiment is an online service – indeed a collection of online services – and is continually evolving in response to its users. To support this, the project commenced with developers being embedded in the user community. Through day-to-day contact between designers and researchers, design is both inspired and validated.

7. Cooperate, Don't Control

myExperiment is a network of cooperating data services with simple interfaces which make it easy to work with content. It both provides services and reuses the service of others. It aims to support lightweight programming models so that it can easily be part of loosely coupled systems.

8. Software Above the Level of a Single Device

The current model of Taverna running on the scientist's desktop PC or laptop is evolving into myExperiment being available through a variety of interfaces and supporting workflow execution.

DISCUSSION

Enabling incentive models for sharing within a "community of practice" and supporting an emergent model of sharing is a challenge [13]. The Virtual Organizations of Grid computing often attempt to achieve a similar objective, although they are typically centred on a common technically defined problem and do not focus on social aspects that might involve different incentive structures.

While the Grid community has developed a variety of portal solutions, we have seen the rise of Web 2.0 sites and it is clear that this paradigm demands exploration to support the next generation of scientists. But science is different to shopping, and the challenge is to achieve the benefits that characterise Web 2.0 within the scientific context. This is the experiment that we are conducting in building myExperiment.

Our design workshops and the review against Web 2.0 design patterns have revealed a deep relationship between myExperiment and Web 2.0. The collective benefits of participation arise not only from the users but also from the developers – ease of use and ease of development. Not only is myExperiment something that can be built using the Web 2.0 approach but it can be used this way too, and it sits comfortably in a Web 2.0 context for reuse. e-Science is difficult – workflows and Web 2.0 make it easier.

myExperiment is one case study in one set of communities. However we believe that much of the context and many of the principles we have discussed in this paper are relevant to other Virtual Research Environments. We suggest it may be a useful exercise to review other VREs in a similar way.

ACKNOWLEDGMENTS

We acknowledge the myExperiment team, all who have worked on myGrid and Taverna, and all our users. In particular we would like to acknowledge Tom Oinn for driving Taverna, Robert Stevens for his myExperiment visions, and Andy Brass for his support. This work is funded through UK grants: EPSRC EP/D044324/1, EP/C536444/1, GR/R67743/01 and JISC VRE2 (04/06).

REFERENCES

1. Cayzer, S. Semantic blogging and decentralized knowledge management. Communications of the ACM, 47(12):47–52, Dec. 2004.
2. Creative Commons, http://creativecommons.org/
3. De Roure, D., Frey, J., Three Perspectives on Collaborative Knowledge Acquisition in e-Science, Workshop on Semantic Web for Collaborative Knowledge Acquisition, Twentieth International Joint Conference on Artificial Intelligence, Hyderabad, India, Jan 2007.
4. Monica Duke, Michael Day, Rachel Heery, Leslie A. Carr, Simon J. Coles. Enhancing access to research data: the challenge of crystallography, JCDL 2005 Digital Libraries: Cyberinfrastructure for Research and Education, Denver, Colorado, USA June 7-11, 2005.
5. Goble, CA et al (2007) *Knowledge discovery for in silico experiments with Taverna: Producing and consuming semantics on the Web of Science*, in Semantic Web: Revolutionising Knowledge Discovery in Life Sciences, Baker C. Cheung K-H (eds) Springer 355-396.
6. myGrid: Middleware for in silico experiments in biology. http://www.mygrid.org.uk/
7. myGrid Portal Party Wiki, http://www.mygrid.org.uk/wiki/Portal
8. Tim O'Reilly. What Is Web 2.0 – Design Patterns and Business Models for the Next Generation of Software. http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html
9. Oinn, T et al. (2006). Taverna: lessons in creating a workflow environment for the life sciences. *Concurrency and Computation: Practice & Experience* **18,** 1067-1100.
10. Open Archives Initiative, http://www.openarchives.org
11. OpenWetWare, http://openwetware.org/
12. The Science Commons. http://sciencecommons.org/
13. Virtual Research Communities, OSI e-Infrastructure Report, http://www.nesc.ac.uk/documents/OSI/vrc.pdf
14. Wroe C, Goble CA, Goderis A, Lord P, et al (2007) *Recycling workflows and services through discovery and reuse* Concurrency and Computation: Practice and Engineering **19**(2) 181-194.