^{my}Experiment: Social Networking for Workflow-using e-Scientists

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

David De Roure
School of Electronics and Computer Science
University of Southampton
Southampton, UK, SO17 1BJ
++44 23 8059 2418
dder@ecs.soton.ac.uk

ABSTRACT

We present the Taverna workflow workbench and argue that scientific workflow environments need a rich ecosystem of tools that support the scientists' experimental lifecycle. Workflows are scientific objects in their own right, to be exchanged and reused. ^{my}Experiment is a new initiative to create a social networking environment for workflow workers. We present the motivation for ^{my}Experiment and sketch the proposed capabilities and challenges. We argue that actively engaging with a scientist's needs, fears and reward incentives is crucial for success.

Categories and Subject Descriptors

D.2.2 [**Software Engineering**]: – *Interoperability*

General Terms

Management, Design, Experimentation, Human Factors.

Keywords

Workflow, social networking, Bioinformatics, Taverna, myGrid, myExperiment, e-Science, scientist

1. INTRODUCTION

The UK's ^{my}Grid project (http://www.mygrid.org.uk) has developed the popular Taverna workflow workbench [1], 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.

Taverna was designed from the outset to suit the work-a-day bioinformatician in a normal, not especially well-resourced, research laboratory; to ease and automate the routine burden of plumbing together the myriad of data resources and analytical tools publicly available and privately developed, and hence release scientists to, well, do science.

Importantly, Taverna has been designed to operate in the "open wild world" of bioinformatics. For example, the services steps executed are expected to be owned by parties other than those using them in a workflow. They are volatile, scruffy and have no contract with their users for reliability. 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 has largely been down to understanding the needs, fears and reward incentives of its different users (service providers, tool developers and bioinformaticians), working "in the wild".

1.1 More than Plumbing

Workflows are effectively plumbing between services. Plumbing is not enough. A workflow environment needs an ecosystem of other tools that support the whole scientific method [2] such as: (a) **designing and running** workflows using a Graphical User Interface designed for expert bioinformaticians, by bioinformaticians; (b) **providing service management** for new kinds of services, legacy services and monitoring continued accessibility of the services; (c) **discovering and publishing** services and workflows using semantic descriptions rich enough to be valuable but simple enough to be captured; and **provenance logging** the process history of a workflow's run and the provenance of the outcomes of the workflow runs.

1.2 Workflows are Scientific Assets

Part of the purpose of promoting workflow-based e-Science was to create an ethos where workflows were recognized as scientific objects in their own right, like data and articles.

Workflows are know-how. Workflows capture valuable know-how that is otherwise often tacit. Workflows are expensive and difficult to develop. Such hard-won assets should be pooled to be drawn upon to be reused, repurposed and recycled by others [3].

Workflows are protocols. They are explicit and precise descriptions of a scientific protocol that at least should enable outcomes to be unambiguously interpretable, and at best make experiment repeatable, or perhaps even reproducible.

As Taverna has become adopted by different communities we have seen over 400 different Taverna workflows appear on the web. Research groups have begun to build their own wikis to publish workflows and their own portals to launch them outside the Taverna environment. Scientists have begun to trade workflows informally through emails. Trainers have begun to request "workflow packs" for Bioinformatics 101 or ask for "certified workflows" from experts in the field.

Workflow e-scientists should also be able to describe and swap workflows, publications, experiences, services, and scientific

WORKS'07, June 25, 2007, Monterey, California, USA.
ACM 978-1-59593-715-5/07/0006.

gossip (a.k.a. insights) as easily as citizens can share documents, photos and videos on the Web.

2. ^{my}Experiment

^{my}Experiment (http://myexperiment.org) is a new initiative from the ^{my}Grid project to create a Virtual Research Environment which makes it easier for workflow workers to gossip about and exchange workflows, regardless of the workflow system – Taverna, Kepler, Triana, ActiveBPEL etc. Scientists rarely care about the workflow engine they use: they typically care about the workflow itself, its function and the services it uses. We envisage:

A gossip shop to share and discuss workflows and their related scientific objects such as provenance logs and semantic descriptions. ^{my}Experiment draws upon social networking websites such as MySpace (http://www.myspace.com) and YouTube (http://www.youtube.com), immediately familiar to our new generation of scientists. We want to promote social tagging, recommendations, ratings etc;

A market place to share, re-use and repurpose workflows, reducing time-to-experiment, sharing expertise and avoiding unnecessary reinvention. Our scientists should be able to "shop" for workflows (and services) like they shop on the web using interfaces that are appealing and not styled on 1970s library catalogues;

A seamless gateway to other established environments, for example: depositing into data repositories; searching digital libraries and publishing to journals. Our scientists work within a larger scholarly lifecycle where running workflows is just a part;

A platform to launch workflows, whatever their system, and handle the provenance logs and data that arise. We hope that our scientists will use whatever workflow is appropriate for their applications– kind of "workflow mashing".

We do not envisage one ^{my}Experiment, but many, set up by different groups, by different communities, by different geographical locations, and even by individuals. Thus we also add the challenge of an inherently federated world between different ^{my}Experiment installations. We will need a new metadata platform such as S-OGSA, proposed by the OntoGrid project [4].

^{my}Experiment throws up a range of technical, political and social challenges, including:

- The building of workflow warehouses vs federating the repositories underpinning the various instances of myExperiment, using the Open Archives Initiative protocols.
- The spectrum of a free-form social space with a social discourse and folksonomy-based tagging to an organised rich "shopping" site using curated semantics that can be effectively navigated by users.
- Handling workflow scientific objects such as provenance logs as workflows and data are exported outside their originating systems, and handling a wide range of identity schemes such as DOIs and LSIDs.
- Confidence and safety. How do we deal with quality, reliability, validation, Intellectual Property, ownership, secrecy? How do we handle open vs protected content? How do we handle private local data mashed with a public environment?
- Workflow hosting for running different workflow systems, and desktop integration, e.g. Google Gadgets;
- Enabling scientists to add value from the start by designing for mash-ups through web services interfaces, the content syndication of scientific objects, and the re-use of the services of other, social tagging etc;
- Socialisation of a community for content, discussion and use.

We have started to develop two pilots for ^{my}Experiment, for Life Sciences and for Chemistry, with two more pilots planned for Social Science and Astronomy.

2.1 e-Science is me-Science

The success of ^{my}Experiment, as it was with Taverna, will depend on understanding the reward incentives of scientists to share and collaborate. Science is a competitive business. The rewards for a scientist are reputation and being first to a discovery; their greatest fears are to be misrepresented and being beaten into second place by a competitor. Thus, e-Science is, inherently, me-Science. The "selfish scientist" will participate if it is to their competitive advantage to do so. ^{my}Experiment aims to support the individual and promote the community through "tribal bonding" within communities and crossing those tribal boundaries to benefit from the expertise of others. We want to encourage meta-workflows, the sharing of invisible tacit know how and know who, best practice dissemination, viral exchange and meme creation, and foster of emergent cultures. This is "OurScience".

The challenge is how we work with the inherent self-interest of the scientist to gain trusted and enthusiastic participation in an inherently altruistic activity that relies in the network effects of many members.

3. ACKNOWLEDGMENTS

We acknowledge all who have worked on ^{my}Grid, Taverna and ^{my}Experiment, and all our users. In particularly we would like to acknowledge Tom Oinn for driving Taverna, Robert Stevens for his ^{my}Experiment 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)

4. REFERENCES

[1] Oinn, T et al. (2006). Taverna: lessons in creating a workflow environment for the life sciences. *Concurrency and Computation: Practice & Experience* **18,** 1067-1100.

[2] 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.

[3] 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

[4] Corcho O, Alper P, Kotsiopoulos I, Missier P, Bechhofer S, Goble CA (2006) *S-OGSA: a Reference Semantic Grid Architecture* in Elsevier Journal Web Semantics 4(2), 81-15