myExperiment: Defining the Social Virtual Research Environment

David De Roure[1], Carole Goble[2], Jiten Bhagat[2], Don Cruickshank[1], Antoon Goderis[2], Danius Michaelides[1] and David Newman[1]

[1]*School of Electronics & Computer Science, University of Southampton*
dder@ecs.soton.ac.uk

[2]*School of Computer Science, University of Manchester*
carole.goble@manchester.ac.uk

Abstract

The myExperiment Virtual Research Environment supports the sharing of research objects used by scientists, such as scientific workflows. For researchers it is both a social infrastructure that encourages sharing and a platform for conducting research, through familiar user interfaces. For developers it provides an open, extensible and participative environment. We describe the design, implementation and deployment of myExperiment and suggest that its four capabilities – research objects, social model, open environment and actioning research – are necessary characteristics of an effective Virtual Research Environment for e-research and open science.

1. Introduction

Scientific advance relies on a social process in which scientists share ideas, methods and data. Traditionally this discourse is mediated by the scholarly publishing process, but scientists are increasingly turning to blogs, wikis and social networks to facilitate this process, a phenomenon sometimes characterised as Science 2.0 [1]. With this we also see a movement to *open science* where large scale, open distributed collaboration is enabled by making data, methods and results freely available on the Web.

The purpose of a *Virtual Research Environment* (VRE) is "to provide researchers with the tools and services they need to do research of any type as efficiently and effectively as possible" [2]. Reflecting our observations on the social process of science, we suggest that an effective virtual research environment should provide four key capabilities, and we propose these as the definition of the "Social Virtual Research Environment":

1. It should facilitate the management and sharing of *Research Objects* – these are the digital commodities that are used and reused by researchers, ranging from data and methods to scholarly publications.
2. It should support the *social model*: producers of research objects should have incentives to make them available; consumers need to be able to discover and reuse them; all will benefit from self- and community-curation.
3. It should provide an *open, extensible environment* to permit ease of integration with other software, tools and services, and benefit from participative contribution of software.
4. It should provide a platform to *action* research, for example to deliver research objects to remote services and software. It should be straightforward to create customised, task specific tools and environments.

Capability (1) is a repository function, (2) is characteristic of social web sites and (3) is related to open source community development. (4) is what makes these into a research environment – the research objects are not just stored and exchanged but they are used in the conduct of research (we describe them as *actionable*). We note that the tenets of open science – open data, open access and open source – are consistent with this definition. Implicit in all these capabilities is the notion that the interface, be it human or programmatic, must be familiar and easy to use.

With a view to establishing these four capabilities we have designed and built myExperiment, a social web site for scientists which directly supports their research. The myExperiment.org service went live in November 2007 and has attracted considerable interest. In the period January-July 2008 the site received over 8,500 unique visitors and achieved over 1,000 registered users. myExperiment is distinctive because it

majors on the social dimension, and it can itself be seen as an experiment to explore whether scientific communities share sufficiently in order to benefit from the network effects of a social web site.

In this paper we report for the first time on the construction and usage of the site, and the insights gained into achieving the four capabilities. Section 2 presents the myExperiment project within the framework of the capabilities and positions it with respect to related work. We then look at the software design in Section 3 and the implementation and deployment in Section 4. After an analysis of usage in Section 5 we close in Section 6 by revisiting the four capabilities and reflecting on our experience of 'the myExperiment experiment' at this stage in its development.

2. The myExperiment VRE

Scientific workflows are valuable commodities which require expertise to build [3]. myExperiment was motivated by observing a clear need to share workflows – to reduce reinvention, propagate best practice and enable scientists to concentrate on science – amongst a fairly decoupled community of workflow users. It was also motivated by a frustration with existing systems which: (a) missed the social dimension, merely making things available rather than encouraging and controlling sharing; (b) presented complex user interfaces out of line with the popular web sites that people are using on an everyday basis, thereby demanding further skill. The motivation and rationale for the myExperiment project is discussed in detail in [4-5] and the design principles in [6].

2.1. myExperiment capabilities

myExperiment addresses the four capabilities of our VRE definition as follows:

2.1.1 Research Objects. Our key research objects are scientific workflows and their associated objects (such as data and documentation). We have extra support for specific workflow formats so that we can 'look inside' these compound objects to extract metadata, provide a graphical rendering and possibly identify the services that are used. We already provide this full range of support for Taverna workflows [7] and we are currently developing support for other systems.

Significantly myExperiment also supports research objects which are collections of other objects, because researchers work with collections of items associated with an experiment – for example, a specific version of a workflow together with input and output data, service invocation logs and documentation. They may also collect multiple workflows together for sharing. These collections are manifest to users as *packs* and are a distinctive feature of myExperiment.

As the user communities of myExperiment increase in number and breadth we are developing support for new research objects, such as experimental plans and statistical models.

2.1.2 Social model. To support producers of research objects in contributing to myExperiment we provide members of the site with support for credit and attribution, and fine control over the visibility and sharing of research objects. Early user feedback revealed this to be the most critical factor in making a social web site *acceptable* for use by scientists.

Other members of the site 'consuming' the research objects can view, download, tag, review and 'favourite' them, which aids their discovery and enhances reputation of the producers. Additionally, content exposed publicly is discoverable through search engines.

Unless they are maintained, workflows and other research objects can cease to be reusable over time – they effectively 'decay', though in fact it is their context that is changing. For example, a recent change in gene identifiers by one service provider led to a myExperiment announcement for users of the affected workflows. Useful workflows will be curated by the community that uses them, and the original authors are also encouraged to curate because they are getting credit for use of their work. Workflow decay is a difficult problem and myExperiment provides a new approach through community curation.

2.1.3 Open environment. myExperiment has paid as much attention to its developer community as it has to designing the user interface.

By creating tools to manage the API, the exposed functionality is highly customisable in response to requirements. The API has enabled new interfaces to be built, such as Google Gadgets and Facebook Apps. It also enables existing interfaces to incorporate myExperiment functionality, such as a wiki or the Taverna workflow workbench.

myExperiment always prefers reuse to reinvention and can easily access other services. It is designed to be part of the scholarly knowledge cycle and is compatible with Open Archives Initiative protocols. While it provides a workflow repository function, much of the associated information – such as data and publications – may be held in other repositories, so myExperiment makes it easy to refer to external content.

In addition to the API, the myExperiment codebase is open source and can be used by anyone to set up their own myExperiment instance.

2.1.4 Actioning research. myExperiment is designed to call upon external services to process research objects. Taverna workflows are executed by myExperiment submitting a collection of research objects for remote processing to an enactor, and the results are automatically collected back into myExperiment. A similar mechanism could run simulations or statistical models, for example.

The service could be local to the user, perhaps in their laboratory, or potentially 'in the cloud'. This latter possibility is significant because researchers are then able to access remote services without any requirement for local software installation.

As well as bringing this capability to the user through the myExperiment interface, the API is designed so that developers are easily able to build 'functionality mashups' over myExperiment for rapid prototyping of tools to support researchers. These may be prescriptive interfaces for specific tasks, such as running preconfigured workflows.

2.2. Related work

A number of systems already provide some of the VRE capabilities we have discussed. No single system however combines all aspects. Table 1 shows representative examples from workflow management systems and community networking sites.

Table 1 VRE capabilities found in existing systems.

	Research Object	Social Model	API	Action
Kepler	Workflow	Yes	No	Yes
Inforsense	Workflow	No	No	Yes
Galaxy	Workflow	No	No	Yes
Facebook	None	Yes	Yes	No
Epernicus	None	Yes	No	No
OpenWetWare	Protocol	Yes	No	No
Nanohub	Simulation	Yes	No	Yes
myExperiment	Workflows	Yes	Yes	Yes

Workflow management systems already make workflows available for sharing, through repository stores for workflows developed as part of projects or communities. Unlike myExperiment, they are tied to a particular type of workflow and do not offer programmatic access to the workflows. For example, the *Kepler Hydrant* (www.hpc.jcu.edu.au/hydrant) is a site (under development) for sharing Kepler workflows. It supports workflow execution and allows users to assign permissions to other users.

Inforsense's commercial *Customer Hub* (www.chub.inforsense.com) has the ambition of enabling Inforsense workflow users to share best practices and leverage community knowledge. However, it does not rely on the social model.

Galaxy (galaxy.psu.edu) provides a public site where biologists can run analyses and for developers it provides an open-source framework for tool and data integration. It does not provide social infrastructure to support sharing of workflows.

Social networking sites such as *Facebook* (www.facebook.com) do not support research objects, and the handling of attribution and licensing may not be adequate for scientists. Facebook supports the development of plug-in applications. Similarly, science-specific social networks, like Epernicus (www.epernicus.com) only support the social part of the VRE function.

The research objects of *OpenWetWare* (openwetware.org) are protocols used in biology labs and, through use of a wiki, OpenWetWare supports the social model and open environment. However, it does not itself intend to be a platform for conducting computationally-intensive research.

Finally, *Nanohub* (www.nanohub.org) is a good example of a VRE that takes a portal approach. It focuses on the nanotechnology domain and provides web-based resources for research, education and collaboration. It also provides simulation tools that can be accessed from the browser. In terms of social infrastructure it provides workspaces, online meetings and user groups. In contrast, myExperiment deliberately set out to build a Web 2.0 site which would be familiar to users, choosing a Web application framework (Ruby on Rails) rather than a portal framework. It offers a rich API and remote execution. myExperiment is designed to provide services to a portal and also to be used as a Web 2.0 'skin' over existing portal services.

3. Software design

myExperiment is designed around a set of entities which are reflected in the internal data model, the user interface (see Figure 1) and the external open data representations. These were derived through extensive user consultation with focus groups and interface mock-ups.

3.1. The entities in myExperiment

myExperiment is being extended to support a variety of research objects in different domains. The current research objects are:

- *Workflows* – compound objects which contain services, the workflow graph, workflow-specific metadata and additional information dependent on the workflow system. Workflows are versioned, and each workflow has usage statistics such as the number of viewings and the number of downloads.
- *Packs* – collections of research objects to form aggregate entities. In addition to objects on the current server, packs can also contain links to objects on other servers.
- *Files* – binary objects that are uploaded to myExperiment and are opaque to the system.
- *Groups* – collections of registered users. The person who creates a group controls its membership through invitations and requests. The data model supports relationships between groups.

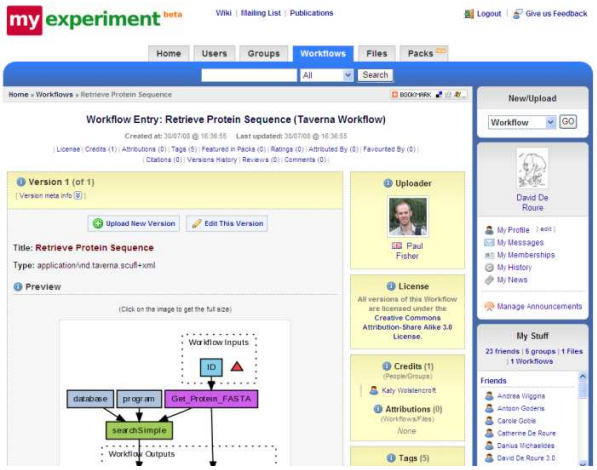

Figure 1. Viewing a workflow in myExperiment. The main navigation tabs and search box are at the top. To the right of the workflow is its 'social metadata' and the user's contextual sidebar.

The social model is supported through a number of entities. Foremost is the member – these are users who have registered with myExperiment and can find or contribute research objects, create tags, comments and reviews etc. They can also form friendships with other members. A great many users of myExperiment are not members – they are simply people browsing the site for publicly available content or following links from search engines such as Google.

The other 'social metadata' entities, many of which are visible in Figure 2, are:

1. Attributions – So that members can show what a research object is based on.
2. Creditations – So that members can give credit to others. By default, the uploader is given credit.
3. Favourites – To support reputation and provide incentive, members can identify their favourite research objects. The list of favourites is visible to other members.
4. Ratings – A simple 5 star rating system to assist with recommendation.
5. Reviews – Explanations to augment ratings.
6. Citations – Publication information associated with research objects.
7. Comments – To enable members to comment on research objects.
8. Tags – To annotate research objects for ease of discovery. The owner of the "tagging" (the association of the tag to the object) is recorded.
9. Policies – see below for Ownership, Sharing and Permissions.

Broadly the user interface reflects the same set of entities and is designed to make it as easy as possible for consumers to find research objects (by search or navigation) and for producers to contribute. Workflows, users, groups and packs have their own pages which become the root for pivoting and browsing. Mechanisms for tagging and commenting etc are consistent across these pages. Only content that is authorised to be shown is visible to the current user. As a result, many parts of the user interface are very dynamic, with different content, features and actions shown/enabled based on the current user context.

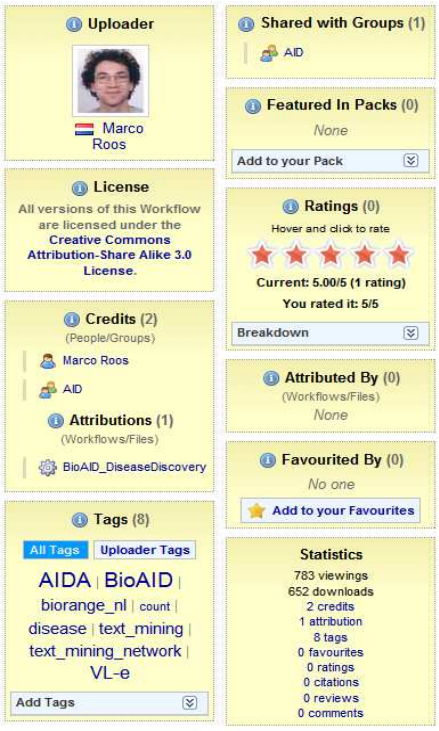

Figure 2. The 'social metadata' associated with a workflow.

Sometimes the user interface does not directly reflect the underlying data model. In particular, the underlying sharing and permissions model is highly object oriented, with User, Group, Contribution, Policy and Permission objects all working together. However, users are presented with 'canned' options that allow quick selection of the most appropriate sharing option, for example: 'anyone can view and download'; 'anyone can view, but only my Friends are allowed to download'.

3.2. Encapsulated myExperiment Objects

As we developed the myExperiment software a recurrent feature request was the ability to upload other content apart from workflows and then link together different pieces of content for a specific purpose. This reflects the fact that scientists work with collections of research objects, such as the input and output data for a workflow or a collection of workflows.

Hence we set out with a more general notion of research object, which captures aggregations of objects and also encompasses the other forms of data in myExperiment – for example members, groups, tags and the social network. We call these objects EMOs (Encapsulated myExperiment Objects). EMOs are represented in RDF and we have developed a myExperiment ontology which uses Dublin Core metadata for research objects and FOAF for the social network information. To meet the versioning requirements of scientists, EMOs carry information which enables mutable content to be validated.

EMOs can be exported in any format to support integration. To interwork with repositories we have adopted the Object Reuse and Exchange representation from the Open Archives initiative (www.openarchives.org/ore), which is based on named RDF graphs.

4. Implementation

4.1. System Architecture

The architecture of one instance of myExperiment is shown in Figure 3. For ease of use, all the interfaces to myExperiment functionality are accessed via the HTTP protocol. For end users we provide an HTML based web interface. External applications can also access the other interfaces, in particular the managed RESTful API (see next section).

In line with our open environment capability, the database server, search server and external workflow enactors are all separate systems to which the main application connects. The interfaces are accessed via a

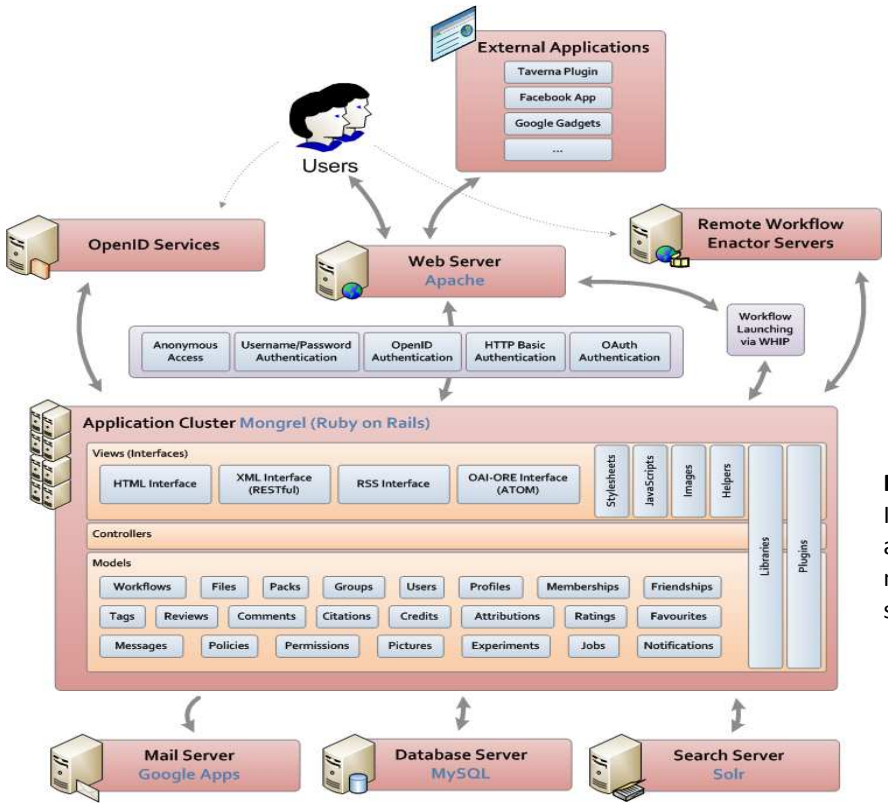

Figure 3. Implementation architecture of a myExperiment server instance.

web server that handles load balancing over a cluster of mongrel application servers. Ultimately scalability will also be achieved by federating multiple instances of myExperiment.

myExperiment is built in the Ruby on Rails web application framework and follows the Model View Controller abstractions set out in Rails. In particular, the models follow the active record pattern as provided by the ActiveRecord library. By keeping with the architectural design of Rails we were able to leverage many of its capabilities to build features for users rapidly.

Various mechanisms for authentication are provided based on the interfaces used. For end users, authentication can be via external OpenID services or the internal username/password mechanism.

4.2. Managed REST API

To support the open and extensible environment we provide data access using basic REST principles, and in line with the community we are increasingly adopting Atom as a means of delivering content and synchronising with peer services. These interfaces have wide adoption in the developer community.

Though Ruby on Rails provides a mechanism for automatically providing REST access, we decided to manage the API separately so that we could respond to the requirements of API users, while also being independent of codebase evolution. Hence the REST API is driven by an XML specification that can be loaded and edited within Microsoft Excel. This allows us to create an independent API specification with the added benefit that it is in one place instead of spread across many model files. It also assists in generating documentation and tests.

Elements of the myExperiment data model have been revealed via the REST API on a case by case basis. Currently, the exposed entities include: workflows, files, users, groups, tags, messages, citations, reviews, comments, ratings and packs.

Given that control of visibility is crucial to myExperiment, we need a means of authenticated API access. This is achieved by using the OAuth protocol, whose purpose is not just to authenticate that a user has given a service consumer access to a service provider; it is a specific key that may have certain privileges assigned to it. With OAuth, a user can create several keys which could be used with one service, and each of those keys may have a different set of privileges.

A developer community is growing up around the API, with projects developing Google gadgets, Facebook Apps, a plugin for the Taverna workflow system, mashups over myExperiment services and a Silverlight interface. It is also being used to incorporate myExperiment functionality in systems such as Wikis. The developer community uses the myExperiment developer wiki to collaborate, following our own principle of supporting the social model.

4.3. Deployment

In response to 24x7 demand, the myExperiment.org servers are hosted in a commercial collocation company with service availability that exceeds university targets. The service is hosted on two servers: a web frontend and a database backend. The frontend consists of the Apache webserver and a cluster of Ruby on Rails processes, running on separate ports using the Mongrel Cluster software.

Static content such as CSS stylesheets, Javascript files and images are served directly by Apache, whereas for dynamic content (HTML and XML), Apache makes connections to the Ruby on Rails processes using the load balancing and proxy Apache modules. The database, which is a major component of the Ruby on Rails system, is hosted on the second server in the form of MySQL. This second server also runs the Solr search server, which is a Java implementation of the Lucene search library running as a Java servlet in Tomcat. To ensure service reliability, CPU load, memory and disk usage is monitored using the Nagios monitoring tools, which also check for correct and timely response of the entire service by making web requests as if it were an external user.

The agile ‘perpetual beta’ development process [8] requires frequent updates to be rolled out to the main myExperiment.org service. This is aided by maintaining a separate server for final testing of code, which allows preview and test of new features and checking for performance regressions with automated tools. A test server containing a recent snapshot of the public data from the live site is also provided to developers writing applications that make use of the myExperiment API.

4.4. Evolution

After 12 months of development by two core developers the myExperiment codebase is now quite a sizeable Ruby-on-Rails application. The models and controllers are approximately 14.5 thousand lines of ruby code and the views are about 12.5 thousand lines of HTML.

The software base comprising the myExperiment VRE is now being extended in an open manner across several projects, notably the BioCatalogue project (see biocatalogue.org) which provides service catalogues. It is also in use in the SKUA astronomy project (see

myskua.org) and the NEMA music analysis project (see nema.lis.uiuc.edu). We have three engagements with the open science research community in chemistry, looking at blogging the lab [9], repository integration and sharing of experimental plans. Other new research objects include statistical models in conjunction with the social statistics community.

5. Analysis of usage

Analysis of myExperiment.org usage statistics over the period January-July 2008 demonstrates: (i) a rapidly growing community, (ii) extensive use of contributed research objects and (iii) the development of social groups.

(i) Community size. At the time of writing, myExperiment.org has 1051 activated accounts. There has been a steady growth in the user base during 2008, with about 10-20 new users registering a week. Spikes in registrations are due to Taverna workshops that use myExperiment to host their tutorial materials and conferences. 38% of the registered users are regular visitors.

In a seven month period the site received approximately 60000 page views in 13500 visits by 8581 unique visitors. The figures are collected using Google Analytics and do not include accesses made via the API. It is interesting to note that the number of unique visitors is much larger than the number of registered users This suggests that the publicly visible content on the site is of value to a wider audience.

(ii) Use of research objects. myExperiment.org hosts three types of research objects: workflows, files and packs.

There are 329 workflows and a further 132 workflows that are revised versions. Workflows were downloaded a total of 50934 times, with three workflows commanding over a 1000 downloads each. Figure 4 shows a general overview of workflow popularity based on downloads. Over time we might expect a larger number of workflows appearing with a smaller number of downloads. The present figure is explained by the strong differences found in documentation of workflows – the less documentation, the fewer downloads.

In terms of permissions, 280 (85%) of the workflows are publicly visible whereas 252 (76%) are publically downloadable. 40% of the workflows with restricted access are entirely private to the user and for the remaining the user has elected to share with individual users and groups. 36 workflows (over 10%) have been shared with the owner granting edit permissions to specific users and groups. In addition there are 53 instances where users have noted that a

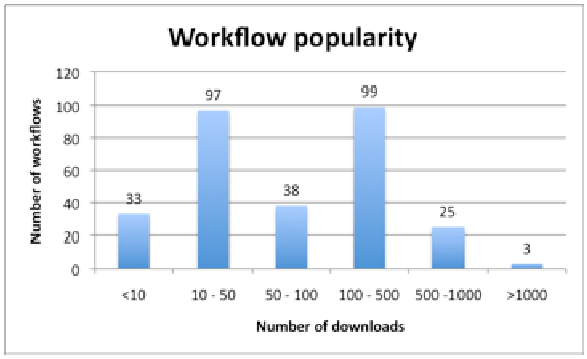

Figure 4. Workflow popularity based on downloads.

workflow is based on another workflow on the site. This indicates that the site is supporting collaboration amongst its users and that they are willing to contribute derived works.

We have also investigated how users discover workflows using the site, finding that an enthusiastic core is willing to share quality workflows but expects credit for doing so, acting as provider to the wider community [10].

Plain files are used much less with only 109 uploads. However, 70% were added after the introduction of packs where users are making use of packs to associate documentation, example inputs and outputs and other files with workflows. Analysing the use of packs would be premature given their recent introduction on the site. To date some 20 packs have been created.

(iii) Social group development. The Groups mechanism has been used to form teams of workflow builders, to help organise events, to collaborate between projects and to locate peers with similar interests. myExperiment currently counts over 100 groups, ranging from two to 20 members.

6. Conclusion

This paper has presented the 'myExperiment experiment'. Based on our experience so far we can reflect on the four capabilities introduced in Section 1:

1. *It should facilitate the sharing of Research Objects.* There is clear evidence of sharing, both in the numbers of downloads of workflows and the extent to which objects are made visible to others. We do not yet have sufficient data to analyse packs, but we anticipate that they will be particularly informative in terms of seeing what users choose to share through this mechanism.

2. *It should support the social model*. We currently have a small number of producers and a large number of consumers. There is clear evidence that consumers are benefitting from the site. Although the producers benefit in the longer term from credit and increased reputation, informally they report that having a popular workflow is a mixed blessing because they receive many requests for assistance. We will monitor usage to see if the asymmetry shifts, whether consumers are able to self-help and to what extent curation occurs. We will also see how we can gain value from the large number of non-members accessing research objects.
3. *It should provide an open, extensible environment.* The API has been very successful: the number of people developing over it quickly exceeded the size of the core development team. It is not sufficient simply to make the API available, but rather we took steps to support the community, such as the creation of the developers' wiki. The automation of API management has paid off the initial investment, providing agile response as API features are requested. Repository integration and federation are currently in development.
4. *It should provide a platform to action research.* We support basic workflow enactment, but if a sophisticated interactive interface is required then users need to use existing tools. Many of our user engagements indicate a very clear desire to run workflows from a web browser rather than installing software to do so.

These achievements have come at a cost. myExperiment is a sizeable Ruby on Rails application. We feel the adoption of this platform has had crucial benefits in terms of the developers spending more time with users and has also assisted in terms of managing the live system. However, we have had to do extensive analysis to manage the scalability of the service, which we are addressing through the federation model. Management of the codebase is now taking more effort when more pervasive changes are required to the underlying entity model, partly due to content in place.

We set out to see if scientists will engage with a social web site VRE and share, and we have now demonstrated that in the right circumstances they do. So far our definition of VRE capabilities is upheld. It is interesting to compare these with the design patterns of Web 2.0 [11] – our definition is consistent but we add the ability for the content to be "actionable". This has been reinforced by our researchers' clear desire for web-based interfaces, and we suggest that this is in line with open science and with the anticipated shift towards cloud computing.

Acknowledgements

The design of the myExperiment codebase and user interface has been significantly influenced by our 'friends and family' including Mark Borkum, Simon Coles, Catherine De Roure, Paul Fisher, Jeremy Frey, Antoon Goderis, Duncan Hull, Yuwei Lin, Savas Parastatidis, Meik Poschen, Rob Procter, Marco Roos, Robert Stevens, Franck Tanoh, Katy Wolstencroft. Thanks to the myGrid team, especially David Withers for supporting the Taverna integration, and to Ian Taylor and Andrew Harrison from the Triana team. myExperiment is funded by the JISC Virtual Research Environments programme and Microsoft Technical Computing Initiative.

References

[1] Waldrop, M. Mitchell, "Science 2.0: Great New Tool, or Great Risk?", Scientific American, Published online January 9, 2008 on http://www.sciam.com/article.cfm?id=science-2-point-0-great-new-tool-or-great-risk

[2] Borda, Ann, et al. Report of the Working Group on Virtual Research Communities for the OST e-Infrastructure Steering Group. London, UK, Office of Science and Technology, 46pp. 2006.

[3] Gil, Y., Deelman, E., Ellisman, M. et al. "Examining the Challenges of Scientific Workflows". IEEE Computer 40(12): 24-32. 2007.

[4] De Roure, D., Goble, C. and Stevens, R., "Designing the myExperiment Virtual Research Environment for the Social Sharing of Workflows," *IEEE International Conference on e-Science and Grid Computing*, pp.603-610, 10-13 Dec. 2007

[5] De Roure, D., Goble, C. and Stevens, R.. "The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows", Future Generation Computer Systems, published online July 2008.

[6] De Roure, D. and Goble, C. Six Principles of Software Design to Empower Scientists. IEEE Software. In Press, 2009.

[7] Oinn, T., Greenwood, M., Addis, M.et al. "Taverna: lessons in creating a workflow environment for the life sciences," Concurrency and Computation: Practice and Experience 18, 10 Aug. 2006, 1067-1100.

[8] Lin, Y., Poschen, M., Procter, R. et al. "Agile Management: Strategies for Developing a Social Networking Site for Scientists," in 4th International Conference on e-Social Science, 18-20 June 2008, Manchester, UK.

[9] Neylon, C. Openwetware blog. See http://blog.openwetware.org/scienceintheopen/

[10] Goderis, A., De Roure, D., Goble, C., Bhagat, J., Cruickshank, D., Fisher, P., Michaelides, D. and Tanoh, F. "Discovering Scientific Workflows: The myExperiment Benchmarks," IEEE Transactions on Automation Science and Engineering . (Submitted 2008)

[11] O'Reilly, T. What is Web 2.0? http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html