Lessons from myExperiment: Two insights into emerging e-Research practice

David De Roure

University of Southampton
Southampton, UK

dder@ecs.soton.ac.uk

Carole Goble

University of Manchester
Manchester, UK

carole.goble@manchester.ac.uk

Abstract. The design of the myExperiment social web site for scientists adopted the Web 2.0 design principles, investigating the question: does Web 2.0 work for scientists? Two years later the site has thousands of users and the largest public collection of scientific workflows of its kind. Here we reflect on two of the Web 2.0 design principles and the insights we have gained into Web 2.0 for researchers. They reveal new forms of sharable object and new ways of assembling data and services.

1. Introduction

The myExperiment Virtual Research Environment [1] 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, the myExperiment's distinctive features to meet the needs of its research user base are support for credit, attributions and licensing, fine control over privacy, a federation model and the ability to execute workflows.

The myExperiment design was consistent with the Web 2.0 design patterns described by O'Reilly [2], a process we reported at the outset of the project [3]. A prototype was launched in July 2007 to gather requirements, and the current www.myexperiment.org site went live after a further 5 months of development and has since evolved substantially in "perpetual beta". It now has 1800 registered users, with thousands downloading public content, and some 600 workflows for multiple workflow systems. Figure 1 shows a workflow in myExperiment, with its associated "social metadata".

It is useful at this stage to reflect on the Web 2.0 design patterns and how they work in the context of e-Research. Here we reflect on two of them and present the insights we have gained through 'the experiment that is myExperiment'. We believe that both of these give us important glimpses of future practice in e-Research.

2. Data is the Next "Intel Inside"

This design patterns tell us "Applications are increasingly data-driven. Therefore for competitive advantage, seek to own a unique, hard-to-recreate source of data." We followed this principle by focusing on scientific workflows, and in fact on one particular system – the Taverna workflow workbench [4] – where we knew there was already a very real need to discover and share workflows. Our "unique selling point" was (and is) that we are *the* place to go find workflows (like photos on flickr, movies on youtube and slides on slideshare). Over the course of time, myExperiment has embraced several other workflow systems.

An alternative strategy would have been to create yet another social networking site for scientists, or to support sharing of arbitrary objects, but for adoption it was more important to start with a specific group of users and meet their requirements [5]. This focus on the specific rather than the generic is at odds with conventional wisdom in computer science and software engineering. It will come as no surprise that the myExperiment architecture is generalisable to share anything – the emphasis on the specific occurs through the codesign work with the users to ensure the site and service is useful to them.

As new objects are shared on myExperiment – from experimental plans for the chemistry lab through to scripts and statistical models – we have maintained a focus on *methods*. This is intrinsic to the incentives that enable myExperiment to succeed: by sharing methods the researchers gain in expertise and reputation, and the community gains in shared know-how and new capacity. It turns out that this also addresses a crucial gap in our thinking about resources, where we perhaps have an over-emphasis on data: with the data deluge that characterises data-intensive science there is also a deluge of the methods used to process it, and we hear of data curation but what of the curation of process and method, and the role of method in curation? [6]

Significantly, we have recognised that researchers do not work with just one content type and that their data is distributed, and so we have also developed support for "packs" – collections of items, both inside and outside myExperiment, which can be shared as one. For example, a pack might contains workflows, example input and output data, results, logs, PDFs of papers and slides – such a pack captures an experiment, can be validated, is self-described, reusable and repurposeable. Packs are created using the shopping basket (or wishlist) metaphor and can be exported using the Object Reuse and Exchange RDF representation which is gaining increasing adoption in the open repositories community.

As we have studied the use cases for packs, and how packs are actually being used, we have recognised the emergence of a new form of digital object – the "Research Object". Research objects are an evolution of packs and provide the sharable, reusable digital objects that enable research to be replayable, repeatable, reproducible and reliable (i.e. unbiased and systematic). It seems entirely likely that, in the fullness of time, objects such as these will replace academic papers as the entities that researchers share, because they plug straight in to the tooling of e-Research. It is Research Objects rather than papers that will be collected in our repositories, and the 'experiment that is myExperiment' will continue to explore this in its next phase.

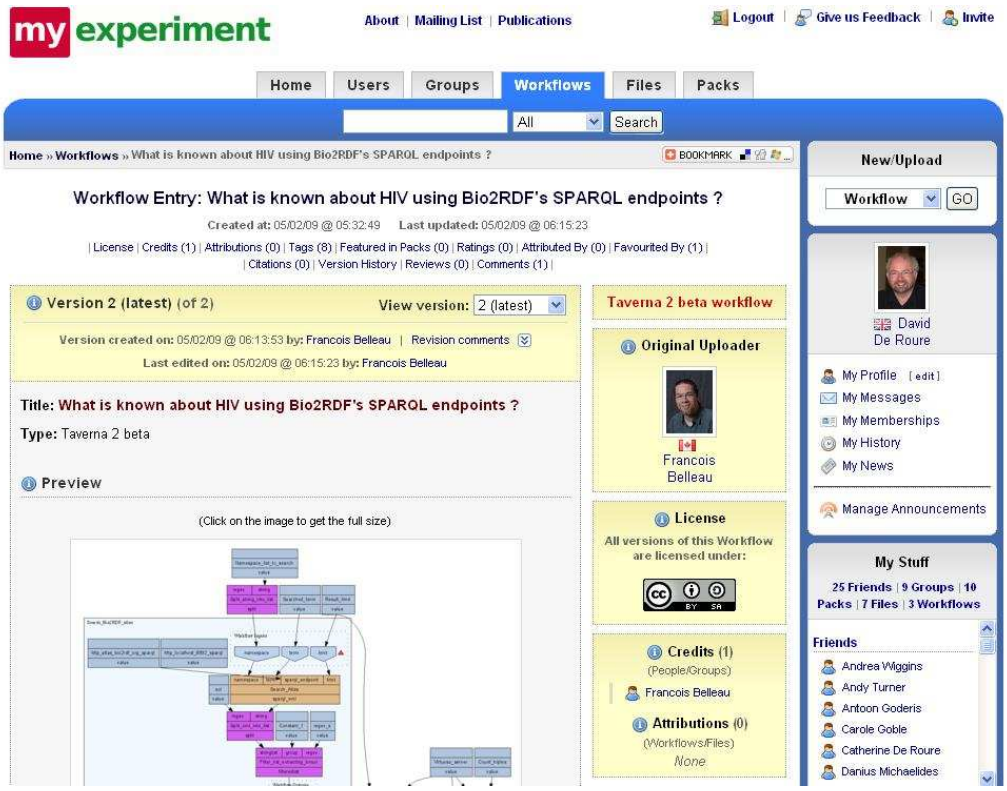

Figure 1: This Taverna workflow on myExperiment builds a mashup and queries it using SPARQL

3. Cooperate, Don't Control

This design pattern says "Web 2.0 applications are built of a network of cooperating data services. Therefore offer web services interfaces and content syndication, and re-use the data services of others. Support lightweight programming models that allow for loosely-coupled systems."

Significantly, this pattern enabled us to bring myExperiment functionality through to users in their existing environment and thereby minimise the impact on their working practice – our notion of adoption is not only to bring the users to myExperiment but to bring myExperiment to the users. Consequently we have put as much effort into supporting the RESTful API for developers as we have the HTML interface for users of the site. The API itself is highly managed, so that we can maintain and evolve it independently of changes to the myExperiment codebase. Alternative interfaces to myExperiment now include Google Gadgets, Facebook, Windows 7 and a bespoke Silverlight interface.

We have also produced a service called rdf.myexperiment.org for the Semantic Web community, which makes myExperiment content available through a SPARQL endpoint. Although it started as an exercise as an adjunct to the main service, it has become the subject of significant interest within the community. It is effectively a generic API whereby the user can specify exactly what information they want to send and what they expect back – instead of asking us to provide this in the API. In some ways it has the versatility of querying the myExperiment database directly, but with the significant benefit of a common data model which is independent of the codebase, and through use of OWL and RDF it is immediately interoperable with available tooling. Exposing our data in this way is an example of "cooperate don't control".

This design pattern also operates at another level in myExperiment: the workflows we are sharing are part of the loose coupling of systems in order to conduct research. This shift to RDF and SPARQL endpoints is an important new development and we are not alone. A number of RDF-based services already exist, especially in the healthcare and life sciences area, and we are beginning to see workflows that use these – as exemplified by the work of Francois Belleau in Figure 1 (see myexperiment.org for further examples). In minutes a user can assemble a pipeline which integrates data and calls upon a variety of services from search and computation to visualisation. While the linked data movement has persuaded public data providers to deliver RDF, we are beginning to see assembly of scripts and workflows that consume it – and the sharing of these on myExperiment.

4. Conclusion

It is an instructive exercise to revisit the Web 2.0 design patterns in the light of the myExperiment experience, and here we have touched on just two. Both are significant because they show emerging research practice which could really take e-Research to the next level: a new kind of object, and a new means of assembly. It is also interesting to observe that our Web 2 VRE began with no explicit intentions of using Semantic Web technologies and yet they have surfaced both in Research Objects and in our world of loosely coupled SPARQL endpoints. In its next 2-year phase, myExperiment will be integrating with institutional repositories and further investigating the notion of the Research Object and the emerging practices of e-Research.

References

[1] De Roure, D., Goble, C. and Stevens, R. (2009) "The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows," Future Generation Computer Systems 25, pp. 561-567. doi:10.1016/j.future.2008.06.010

[2] O'Reilly, T. (2005) What Is Web 2.0? "Design Patterns and Business Models for the Next Generation of Software," September 2005. http://oreilly.com/web2/archive/what-is-web-20.html

[3] De Roure, D. and Goble, C. (2007) "myExperiment – A Web 2.0 Virtual Research Environment." International Workshop on Virtual Research Environments and Collaborative Work Environments, May 2007, Edinburgh, UK. http://eprints.ecs.soton.ac.uk/13961/

[4] T. Oinn, M. Greenwood, M. Addis, N. Alpdemir, J. Ferris, K. Glover, C. Goble, A. Goderis, D. Hull, D. Marvin, P. Li, P. Lord, M. Pocock, M. Senger, R. Stevens, A. Wipat, and C. Wroe,(2006) "Taverna: lessons in creating a workflow environment for the life sciences," Concurrency and Computation: Practice and Experience, vol. 18, iss. 10, pp. 1067-1100, 2006.

[5] De Roure, D. and Goble, C. (2009) "Software Design for Empowering Scientists," IEEE Software, vol. 26, no. 1, pp. 88-95, January/February 2009. doi:10.1109/MS.2009.22

[6] De Roure, D. and Goble, C. (2008) "Curating Scientific Web Services and Workflows." EDUCAUSE Review, vol. 43, no. 5 (September/October 2008). http://www.educause.edu/library/erm0857