Lessons from myExperiment: Research Objects for Data Intensive Research

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

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 discover, publish and curate scientific workflows and other artefacts. While it shares many characteristics with other Web 2.0 sites, myExperiment's distinctive features to meet the needs of its research user base include support for credit, attributions and licensing, fine control over privacy, a federation model and the ability to execute workflows. Figure 1 shows a workflow in myExperiment, with its associated "social metadata".

myExperiment now has over 2000 registered users, with thousands more downloading public content, and the largest public collection of workflows, for systems which include Microsoft's Trident. Created in close collaboration with its research users [2], myExperiment gives important insights into emerging research practice. As it moves into its second phase we see new forms of sharable Research Object which challenge traditional scholarly publishing and provide an important basis for data-intensive science. To support this, semantic technologies are increasingly coming into play to maximise the potential for reuse and repurposing of experiments.

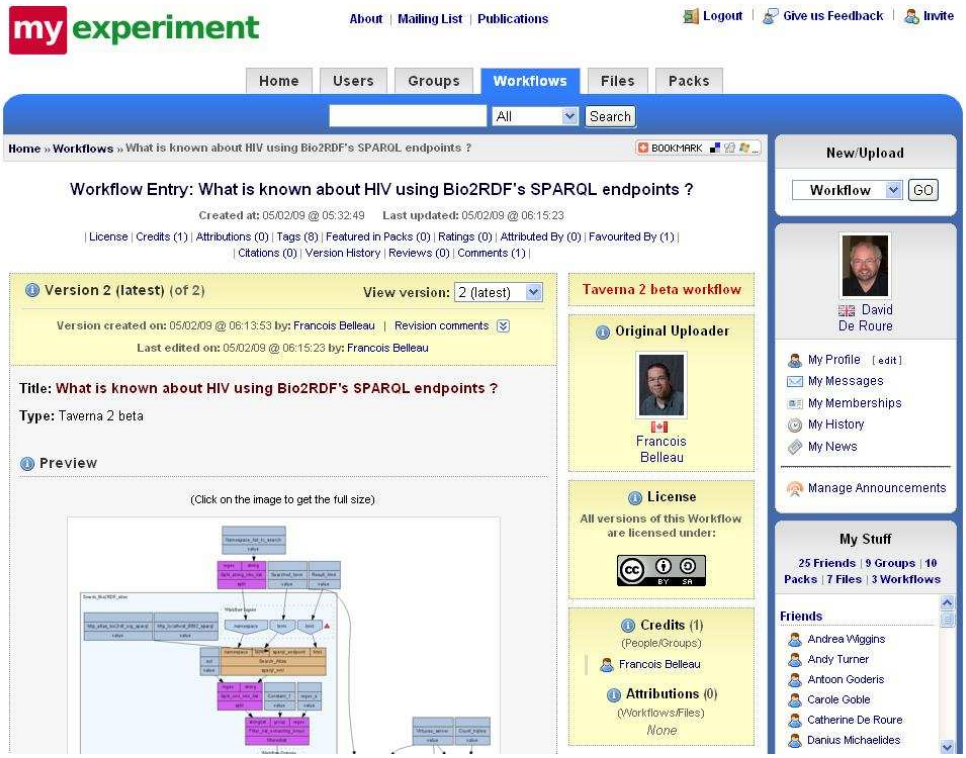

Figure 1: This Taverna workflow on myExperiment builds a mashup and queries it using SPARQL

2. From workflows to Packs

The Web 2 design patterns [3] tell us "Data is the next Intel Inside. Applications are increasingly data-driven. Therefore for competitive advantage, seek to own a unique, hard-to-recreate source of data." We followed this principle in myExperiment by focusing on scientific 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 workflow systems including Taverna and Trident.

Significantly, we have also recognised that a workflow can be enriched as a sharable item by bundling it with some other pieces which make up the "experiment". We observe that researchers do not work with just one content type and moreover that their data is not in just one place – it is distributed, and sometimes quite messy too. Hence 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 and is reusable and repurposeable. Packs are created using the shopping basket (or wishlist) metaphor and can be exported using the Object Reuse and Exchange representation which is gaining increasing adoption in the open repositories community.

3. Research Objects for Data-Intensive Science

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 to support data intensive research. They have important properties:

- **Replayable** – go back and see what happened. Experiments are automated and may occur in milliseconds or in months. Either way, the ability to replay the experiment, and to study parts of it, is essential for human understanding of what happened.
- **Repeatable** – run the experiment again. There's enough in a Research Object 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 needed for the scale of data intensive science.
- **Reproducible** – run new experiment to reproduce the results. To reproduce (or replicate) a result is for someone else to start with the same materials and methods and see if a prior result can be confirmed.
- **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.
- **Repurposeable** – reuse the pieces in a new experiment. An experiment which is a black box is only reuseable 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 re-used.
- **Reliable** – robust under automation, which brings systematic and unbiased processing, and also "unattended experiments" – human out the loop. In data-intensive science, Research Objects promote reliable experiments, but also they must be reliable for automated running.

We believe that in the fullness of time, objects such as these will replace academic papers as the entities that researchers share, because they plug straight into the tooling of e-Research. This means it is Research Objects rather than papers that

will be collected in our repositories, and as well as a workflow repository, myExperiment has become a prototypical Research Object repository. In the second phase of myExperiment we are integrating with EPrints in Southampton and Fedora in Manchester and we will pursue this line of exploration.

4. A Semantic approach

To achieve these properties a Research Object must be self-contained and self-describing – containing enough metadata to have all the above characteristics and have maximal potential for re-use, whether anticipated or unanticipated. To explore this we have developed a prototype service (rdf.myexperiment.org) which makes myExperiment content available according to the myExperiment data model. This uses a modularised ontology drawing on Dublin Core, FOAF and OAI Object Reuse and Exchange representation. We are building on the SWAN-SIOC work and Science Collaboration Framework at Harvard [4], and also with the Open Provenance Model (OPM) [5], and we are complying with the linked data model to make sure our Research Objects are as re-usable as possible.

rdf.myexperiment.org makes myExperiment content available through a SPARQL endpoint and this 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 the "cooperate don't control" principle of Web 2.0.

This brings myexperiment into the fold of the other SPARQL endpoints in e-Science, especially in the healthcare and life sciences area [6], 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. We believe this is an important glimpse of future research practice: the ability to assemble with ease experiments that are producing and consuming this form of rich content.

References

[1] De Roure, D., Goble, C., Aleksejevs, S., Bechhofer, S., Bhagat, J., Cruickshank, D., Fisher, P., Hull, D., Michaelides, D., Newman, D., Procter, R., Lin, Y. and Poschen, M. (2009) Towards Open Science: The myExperiment approach. Concurrency and Computation: Practice and Experience. (In Press)

[2] 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

[3] 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

[4] Sudeshna Das, Lisa Girard, Tom Green, Louis Weitzman, Alister Lewis-Bowen and Tim Clark "Building biomedical web communities using a semantically aware content management system," Briefings in Bioinformatics 2009 10(2):129-138; doi:10.1093/bib/bbn052

[5] Luc Moreau, Paul T. Groth, Simon Miles, Javier Vázquez-Salceda, John Ibbotson, Sheng Jiang, Steve Munroe, Omer F. Rana, Andreas Schreiber, Victor Tan, László Zsolt Varga: The provenance of electronic data. Commun. ACM (CACM) 51(4):52-58 (2008)

[6] François Belleau, Marc-Alexandre Nolin, Nicole Tourigny, Philippe Rigault, Jean Morissette: Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. Journal of Biomedical Informatics 41(5): 706-716 (2008)