Scientific Social Objects

The Social Objects and Multidimensional Network of the myExperiment Website

David De Roure
Oxford e-Research Centre
University of Oxford
Oxford, UK
david.deroure@oerc.ox.ac.uk

Sean Bechhofer and Carole Goble
School of Computer Science
The University of Manchester
Manchester, UK
firstname.lastname@manchester.ac.uk

David Newman
Electronics and Computer Science
University of Southampton
Southampton, UK
drn@ecs.soton.ac.uk

Abstract— **Scientific research is increasingly conducted digitally and online, and consequently we are seeing the emergence of new digital objects shared as part of the conduct and discourse of science. These *Scientific Social Objects* are more than lumps of domain-specific data: they may comprise multiple components which can also be shared separately and independently, and some contain descriptions of scientific processes from which new objects will be generated. Using the myExperiment social website as a case study we explore Scientific Social Objects and discuss their evolution.**

Keywords—Scientific Social Object; workflow; Research Object

I. Introduction

Routine research practice in many disciplines has entered the "Science 2.0" world where we have new mechanisms for sharing [1] and also new objects to share. Research tools produce and consume data, together with metadata to aid interpretation and reuse. We also have the scripts and experiment plans that support automation, and the records that make the results interpretable and reusable. Our new objects include data, metadata, scripts, workflows, provenance records and ontologies, and our tools for sharing include the array of collaboration tools from repositories, blogs and wikis to social networking, instant messaging and tweeting that are available on the Web today. Where researchers come together around these objects they become *Scientific Social Objects.*

In this paper we focus on one of these new objects, the computational scientific workflow [2], as a case study. Scientific workflow systems are used to conduct automated data analysis, predictions and validations, and have emerged as a key part of today's data-intensive research environment. A workflow itself provides a transparent encoding of a particular process that is then shared in order to support reproducible science and the spread of knowledge and expertise.

The myExperiment website (www.myexperiment.org) was designed to make it easy to share these workflow objects – a kind of flickr or youtube but for workflows [3]. It has successfully adopted a Web 2.0 approach in delivering a social website where scientists can discover, publish and curate scientific workflows and other objects. 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, licensing and privacy. Since its launch at the end of 2007, myExperiment has over 4000 registered users, thousands more downloading public content, and with nearly 2000 workflows it provides the largest collection available. It is, however, characteristically a 'boutique' site with a specialist audience.

We propose that workflows, and myExperiment as a resource, provide a useful case study in scientific social objects. Workflows are indeed social objects that are shared and used by researchers, but significantly they are composite objects containing heterogeneous components which can be shared separately and independently. Since they capture process they are also prescriptions for the creation of other objects. This distinguishes them from an object type such as a photo or collection of photos.

The next section illustrates the workflow as a social object, and in Section III we describe the multiple interlinked networks in myExperiment. We then explain how these can be explored through myExperiment's SPARQL query interface in order to support further study. Section V characterises a future scientific social object that we call a *Research Object*. Finally we outline current work in the Wf4Ever workflow preservation project.

II. Workflows as Social Objects

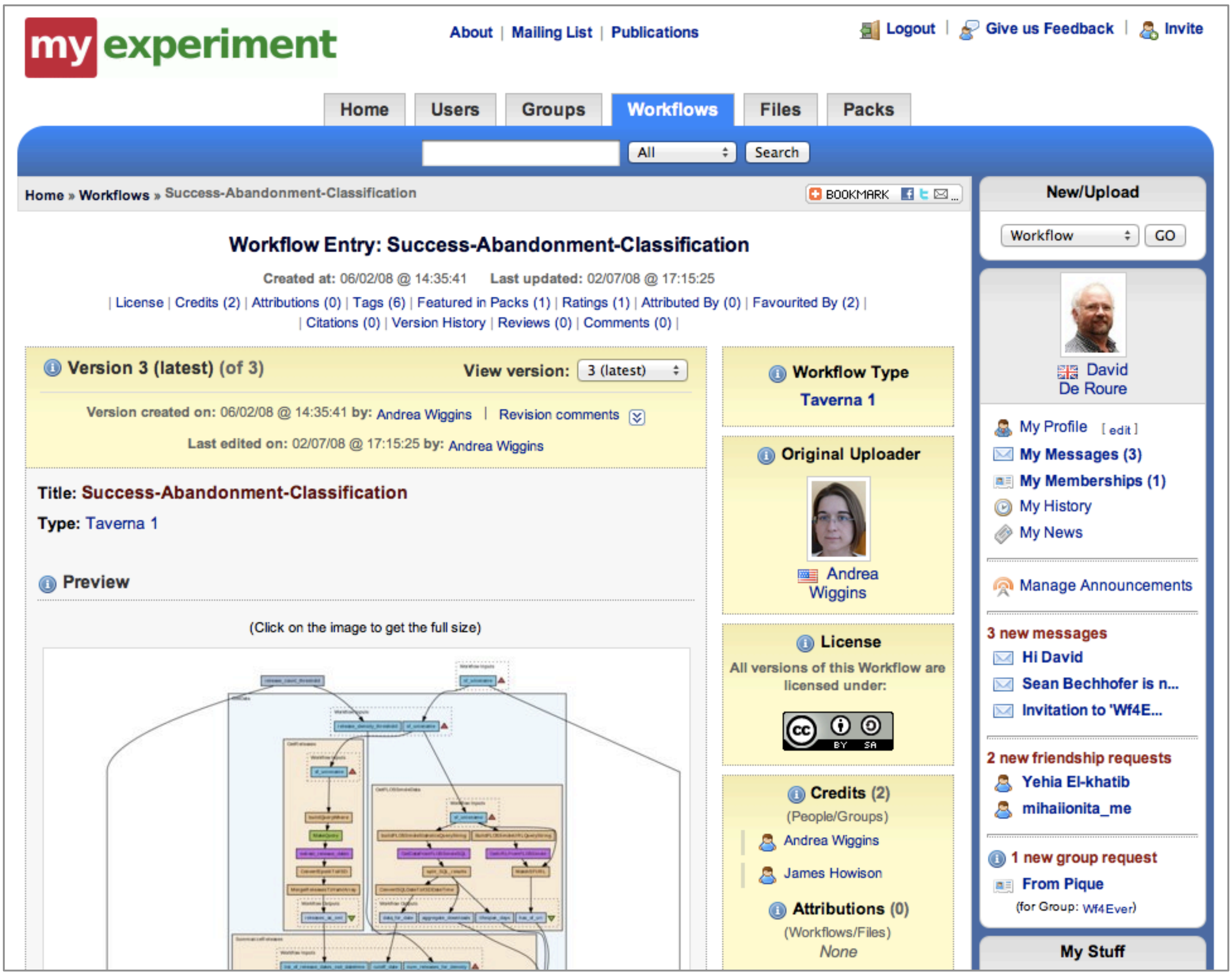
Workflows are social objects in that they form connections between people in many different ways. Within the context of the myExperiment site these collaborations are asynchronous, with researchers collaborating around a workflow over a period of time, perhaps on an ad hoc basis.

First it is useful to understand the anatomy of a workflow: it is a precise, executable description of a scientific procedure – a multi-step process to coordinate multiple tasks, like a script. Each task represents the execution of a computational process, such as running a program, submitting a query to a database, submitting a job to a computational facility, or invoking a service over the Web to use a remote resource. Data output from one task is consumed by subsequent tasks according to a predefined graph topology that orchestrates the flow of data. The tasks might be local or they may occur remotely hosted by third parties.

The following scenarios illustrate three typical interactions that come about through workflows as social objects:

1. A user makes a runnable workflow publicly available and publishes its URI in a paper. Others find and use it, perhaps creating a new version which credits the original creator. They might also contact the user for help in using the workflow or to give feedback.

myExperiment is supported by UK JISC. Wf4Ever is funded by the Seventh Framework Programme of the European Commission (Digital Libraries and Digital Preservation area ICT-2009.4.1 project reference 270192).



About | Mailing List | Publications

Logout | Give us Feedback | Invite

Home | Users | Groups | Workflows | Files | Packs

All | Search

Home » Workflows » Success-Abandonment-Classification

BOOKMARK

Workflow Entry: Success-Abandonment-Classification

Created at: 06/02/08 @ 14:35:41 Last updated: 02/07/08 @ 17:15:25

| License | Credits (2) | Attributions (0) | Tags (6) | Featured in Packs (1) | Ratings (1) | Attributed By (0) | Favourited By (2) |
| Citations (0) | Version History | Reviews (0) | Comments (0) |

Version 3 (latest) (of 3) View version: 3 (latest)

Version created on: 06/02/08 @ 14:35:41 by: Andrea Wiggins | Revision comments

Last edited on: 02/07/08 @ 17:15:25 by: Andrea Wiggins

Title: Success-Abandonment-Classification

Type: Taverna 1

Preview

(Click on the image to get the full size)

Workflow Type
Taverna 1

Original Uploader
Andrea Wiggins

License
All versions of this Workflow are licensed under:

Credits (2)
(People/Groups)
Andrea Wiggins
James Howison

Attributions (0)
(Workflows/Files)
None

New/Upload
Workflow | GO

David De Roure

My Profile [edit]
My Messages (3)
My Memberships (1)
My History
My News

Manage Announcements

3 new messages
Hi David
Sean Bechhofer is n...
Invitation to 'Wf4E...

2 new friendship requests
Yehia El-khatib
mihaiionita_me

1 new group request
From Pique
(for Group: Wf4Ever)

My Stuff

Figure 1. myExperiment website, showing a workflow (left) with its 'social metadata' (middle) and the user's social network (right).

2. A user finds a workflow by searching myExperiment and needs help in its use. They can see who created the workflow, with which groups it is shared and by whom it is favourited or rated.

3. A user finds a workflow and tries to use it but there are difficulties with a particular task in the workflow. They search for other workflows which use the same task in order to find others who may help.

These scenarios demonstrate that myExperiment provides multiple routes of connection between people around the workflow as a social object. In the latter case it is a task within the workflow that connects people. However workflows are not the only social objects; for example, a scientist may be analysing data and may bring both workflows and other scientists together around that data.

III. The Networks

myExperiment has a multidimensional network that links people and the objects that they share, and some of those objects are themselves networks as is the case with workflows. Rather than treating this as one big network, here we suggest six categories of network that are superimposed in myExperiment.

A. Friends, groups, ownership and credit

myExperiment provides notions of friends and groups, familiar from other social websites. The groups have administrators, and visibility and sharing of objects can be finely controlled. As well as being owned by people and groups, objects give explicit credit to those who were involved in creating them – this consideration of credit and attribution is critically important in the scientific context. These are the main social graphs of myExperiment and clearly evident in the interface, which is illustrated in Figure 1.

B. Workflows

A workflow can be viewed as a network of tasks, any of which may also appear in other workflows. Workflows can call other workflows, and inputs and outputs may also be typed so that they can be matched up. Hence there is a 'workflow network'.

Furthermore the workflow consumes and produces data, and so a workflow execution ('run') can produce a provenance graph which describes the sources of information and processes involved in producing a particular output. As a record of a particular experiment, the provenance graph could itself be a social object.

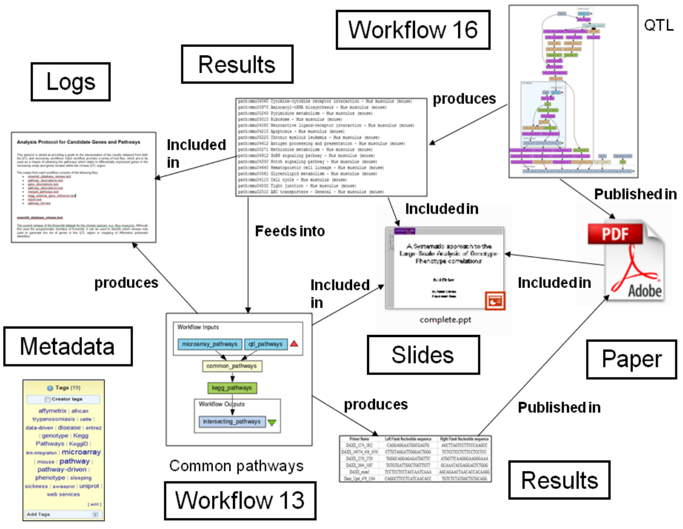


Figure 2. A myExperiment pack with annotated relationships.

C. Packs

myExperiment users were quick to recognise that a workflow can be enriched as a social object by bundling it with some other pieces which make up the “experiment”. Hence we developed support for *packs* – collections of items, both inside and outside myExperiment, which can be shared as one bundle. For example, a pack might contain workflows, example input and output data, results, logs, PDFs of papers and slides (see Figure 2) – such a pack captures an experiment and is reusable and repurposeable. Approximately 10% of the contributions to myExperiment are packs.

A pack is also a network – it is essentially a bundle of annotated URIs with relationships between them. Packs together form a ‘pack network’ by pointing at each other, by sharing components or by being shared.

D. Tags and other annotations

Annotation of Social Objects, through tagging, reviews and favouriting by multiple users, leads to another superimposed network on the site. Folksonomy-based tagging creates an emergent network of social objects linked by common tags, while we also have controlled vocabularies and some semi-automated tagging as part of the workflow curation process. The act of tagging is seen as significant and tags have an owner attached.

E. Citation network

myExperiment’s network is also interlinked with external networks, as illustrated by case 1 above. Workflows and packs on myExperiment are referred to by URIs which then appear in research publications; the publications are themselves linked by co-authorship and citation networks. Hence the social objects on myExperiment participate in these bibliographic networks.

This network is only partially stored on myExperiment but it is important: we track it carefully by running queries to identify myExperiment citations and then logging these, as well as inviting people to inform us of publications. myExperiment links both in and out of external repositories, and there is ongoing work integrating with dlibra and EPrints.

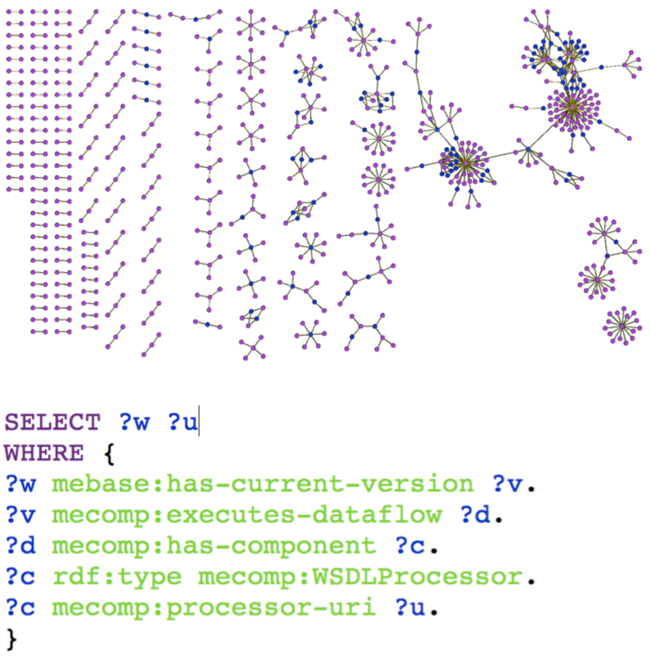
F. Service network

A great many of the workflows on myExperiment make use of remote web services and thus form a service network, which itself has been the subject of analysis [4]. These services are entities in their own right, many stored in myExperiment’s sister site www.biocatalogue.org which provides a community-curated registry of Web Services in the life sciences [5]. A task in a workflow links to a service description in Biocatalogue and hence into the associated network of users and curators.

IV. Querying the Networks

The networks described above are published in RDF (Resource Description Framework) and follow Linked Data practice. Every myExperiment entity, whether it be a Workflow, Pack, User, Group, etc. has its own *Non-Information Resource* URI to identify it. The structure of myExperiment RDF is defined by ontology modules that can be assembled to build the complete myExperiment Ontology [6]. This set of modules borrows classes/properties from FOAF, SIOC, Dublin Core, Creative Commons and OAI-ORE. Depending on the workflow system in question it is possible to access the workflow graph: the majority of myExperiment workflows are in the Taverna [7] system and available as RDF.

All myExperiment's public RDF data can queried using the query language SPARQL at myExperiment's SPARQL Endpoint, which implemented using the 4store RDF database and reasoner. There is also a tutorial available on http://rdf.myexperiment.org/howtosparql It is relatively easy to query the networks described above. For example, Figure 3 shows a visualisation of the services network based on Taverna workflows with the associated SPARQL query. Further queries and visualisations can be found on http://wiki.myexperiment.org/index.php/Vis



```
SELECT ?w ?u
WHERE {
?w mebase:has-current-version ?v.
?v mecomp:executes-dataflow ?d.
?d mecomp:has-component ?c.
?c rdf:type mecomp:WSDLProcessor.
?c mecomp:processor-uri ?u.
}
```

Figure 3. Querying and visualising the services network using the SPARQL endpoint and the Cytoscape network analysis tool.

V. From Packs to Research Objects

As the design of the myExperiment site has co-evolved we observe that packs are becoming a social object in their own right. In some ways they have a role like papers, in capturing materials, method and results and supporting reproducibility. Currently they provide a machine-readable supplement to the academic paper, but as the utility of scientific social objects increases it is interesting to speculate how and when they might replace it.

To consider this evolution we have generalised the notion of packs to a future scientific social object which we call the *Research Object*. Through a series of discussions about the affordances of these social objects[1] we propose the following dimensions [8].

- *Reusable*. The key tenet of Research Objects is to support the sharing and reuse of data, methods and processes. Thus our Research Objects must be reusable as part of a new experiment or Research Object. This is black box reuse as a whole or single entity.
- *Repurposeable*. Reuse may also involve the reuse of constituent parts of the Research Object, for example taking a study and substituting alternative services or data for those used in the study. To facilitate such a disaggregation and recombination, Research Objects should expose their constituent pieces.
- *Repeatable*. There should be sufficient information in a Research Object for the original researcher or others to be able to repeat the study, perhaps years later. Information concerning the services or processes used, their execution order and the provenance of the results will be needed.
- *Reproducible*. To reproduce (or replicate) a result is for a third party to start with the same inputs and methods and see if a prior result can be confirmed. Reproducibility is key in supporting the validation and non-repudiation of scientific claims.
- *Replayable*. If studies are automated they might involve single investigations that happen in milliseconds or protracted processes that take months. Either way, the ability to replay the study, and to study parts of it, is essential for human understanding of what happened.
- *Referenceable*. If research objects are to augment or replace traditional publication methods, then they (and their constituent components) must be referenceable or citeable.
- *Revealable*. The issue of provenance, and being able to audit experiments and investigations is key to the scientific method. Third parties must be able to audit the steps performed in the research in order to be convinced of the validity of results.
- *Respectful*. Explicit representations of the provenance, lineage and flow of intellectual property associated with an investigation are needed.

Although not explicit in this list, it is the nature of research that these objects need to be interpreted and reused across laboratory, community and disciplinary boundaries, and for his reason it is also constructive to consider Research Objects as *Boundary Objects* [9].

VI. Conclusion and Future Work

Scientific social objects are becoming crucial to data-intensive research, and myExperiment provides a useful case study (a social probe) into how researchers work with workflow objects in particular. Although workflows are a specific kind of object they may help us define scientific social objects in general; for example, some of the aspects of the myExperiment network may be more generally applicable, especially with respect to the composite nature of SSOs, inclusion of process descriptions and potentially their executability.

In the Wf4Ever project (http://www.wf4ever-project.org) we are focusing on workflow preservation, building on this experience with myExperiment. Packs have demonstrated a significant role in workflow reuse and curation, so the project is further developing the notion of Research Objects in this context. Wf4Ever also features an important strand of activity in recommender systems, which could draw heavily on the multidimensional network in order to assist users in their interactions with scientific social objects.

We invite others to join in the analysis of the myExperiment networks and hope that it may further the study of scientific social objects.

References

[1] B. Shneiderman. Science 2.0. Science 7 March 2008: **319** (5868), 1349-1350. doi:10.1126/science.1153539.

[2] Y. Gil, E. Deelman, M. Ellisman, T. Fahringer, G. Fox, C. Goble, M. Livny, L. Moreau and J. Myers. Examining the challenges of scientific workflows. IEEE Computer, 40:24-32, Dec. 2007. doi: 10.1109/MC.2007.421

[3] D. De Roure, C. Goble, and R. Stevens. The design and realisation of the myExperiment virtual research environment for social sharing of workflows. Future Generation Computer Systems, 25(5):561-567, 2009. doi:10.1016/j.future.2008.06.010

[4] W. Tan, J. Zhang and I. Foster. Network Analysis of Scientific Workflows: a Gateway to Reuse. IEEE Computer, 43(9): 54-61, 2010. doi:10.1109/MC.2010.2622010

[5] J. Bhagat, F. Tanoh, E. Nzuobontane, T. Laurent, et al. BioCatalogue: a universal catalogue of web services for the life sciences. Nucl. Acids Res. (2010) 38 (suppl 2): W689-W694. doi: 10.1093/nar/gkq394

[6] D. Newman, S. Bechhofer and D. De Roure. myExperiment: An ontology for e-Research. In: Workshop on Semantic Web Applications in Scientific Discourse in conjunction with the International Semantic Web Conference, October 2009, Washington DC, US

[7] D. Hull, K. Wolstencroft, R. Stevens, C. Goble, et al. Taverna: a tool for building and running workflows of services. Nucleic Acids Research, 34(suppl 2):W729-W732, 1 July 2006. doi: 10.1093/nar/gkl320

[8] S. Bechhofer, J. Ainsworth, J., Bhagat, I. Buchan, P. Couch, D. Cruickshank, et al. Why linked data is not enough for scientists. In IEEE Sixth International Conference on e-Science, pages 300-307, 2010. Doi: 10.1109/eScience.2010.21

[9] S.L. Star and J.R. Griesemer JR. Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. Social Studies of Science 19 (3): 387–420. doi:10.1177/030631289019003001

[1] An earlier list of twelve 'R dimensions' can be found in the article "Replacing the Paper: The Twelve Rs of the e-Research Record" on http://blogs.nature.com/eresearch/