CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE

Concurrency Computat.: Pract. Exper. 2000; **00**:1–6

Towards Open Science: The myExperiment approach

D. De Roure[*], Carole Goble[†], Sergejs Aleksejevs[†], Sean Bechhofer[†], Jiten Bhagat[†], Don Cruickshank[*], Duncan Hull[†], Yuwei Lin[†], Danius Michaelides[*], David Newman[*], Rob Procter[†]

[*] *School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK.*

[†] *School of Computer Science, The University of Manchester, Manchester M13 9PL, UK*

SUMMARY

By making scientific content more reusable, and providing a social infrastructure which facilitates sharing, the human aspects of the scholarly knowledge cycle may be accelerated and thereby reduce 'time-to-discovery'. We introduce the notion of the Research Object – the work objects that are built, transformed and published in the course of scientific experiments – and suggest that by encapsulating methods with results we can achieve more repeatable, reusable and replayable science. We then present myExperiment, a social web site for discovering, sharing and curating Research Objects. We describe how myExperiment facilitates the management and sharing of research workflows, supports a social model for content curation tailored to the scientist and community, supports open science by exposing content and functionality into the users' tools and applications, and introduces a more general notion of the e-Laboratory.

KEY WORDS: *Scientific Workflow, Research Object, Web 2.0, Data Curation, e-Laboratory*

1. INTRODUCTION

1.1 Motivation

To accelerate the time to discovery of new research results we must look at the human component of the discovery cycle. Scientific advance relies on a social process in which scientists share hypotheses, insights and results, and the data and methods that support these. Traditionally scholarly discourse and dissemination has focused on the publication of peer reviewed articles, mediated by the scholarly publishing process and established structures such as conferences. In recent years it has become a widespread and successful practice to use the Web as a distributed dissemination platform for research resources: in addition to websites that provide research portals and repositories, there are now tens of thousands of publically available web services across business and science [ref]. There has also been an expansion in the kinds of scientific commodities being published, for example:

- *Primary and secondary scientific data sets*, along with standard metadata sufficient to support their interpretation [ref], although the tying together published results with the data that upon which they are based is still poorly supported – This often goes under the title "supplementary data" but there are major, unsolved issues to do with persistence [D13,D 14]
- *Algorithms, software tools, scripts and procedures*, such as community services like the nanoHUB gateway [ref], which hosts user-contributed resources in the nanotechnology domain, and OpenWetWare [ref] which provides an exchange for techniques in biological sciences;

This latter point is the focus of our work. Scientists need to share (and find) not just the digital materials of scientific research but also the *digital methods and processes*: the protocols, plans, and standard operating procedures of bench science and the scripts, workflows and provenance records of e-Science. Methods are

scientific commodities in their own right, with associated intellectual property, metadata, and life cycles; and as with data and articles, subject to their own forms of authorship, credit and reuse criteria:

- *By pooling and sharing methods* we have the potential to accelerate science by exchanging know-how and best practice, avoiding reinvention and hence reducing time-to-experiment [ref]. Moreover, participating scientists are not always organised into predetermined Virtual Organisations but form fluid opportunistic groupings amongst decoupled strangers.
- *By combining methods with results* we can accelerate discovery by enabling transparent, comparable and reproducible science [ref]. By packaging and aggregating methods with data, results, publications, tutorials, simulations, logs, tags and people (experts, members, groups) and sharing these across applications as publication units we can work towards an open *e-Laboratory* that is outside any specific application.

A case in point is the Scientific Workflow. The Web provides a platform for delivering not just documents and data but also services which support the research process: Scientific workflows are the means to compose these, providing descriptions of processes that specify the co-ordinated execution of multiple tasks so that, for example, data analysis and simulations can be repeated and accurately reported. Alongside experiment plans, Standard Operating Procedures and laboratory protocols, automated workflows are one of the most recent forms of scientific digital methods, and one that has gained popularity and adoption in a short time [1]. However, they are often complex and challenging to build, and can require specialist expertise that is hard-won and may be outside the skill-set of those who need [ref CCPE reuse paper]. The suite of scientific workflows in [trps] took a bioinformatics expert six months and over 40 versions to develop; however, once developed they were immediately reusable by other, perhaps less experienced, e-Scientists in turn accelerating their research [ref].

1.1 Research Objects

It follows that the key to accelerating the human part of the lifecycle is to make the digital resources as reusable as possible and to include methods as well as outputs. Hence we introduce the notion of *Research Objects* (ROs) – the work objects that are built, transformed and published in the course of scientific experiments. They are compound objects that group together resources used in a scientific investigation – an aggregation of datasets, analysis methods, workflows, results, and electronic records that represent a narrative about an investigation, experiment, question or process, and the corresponding metadata. A digital resource in its native application format, like a document, script or spreadsheet, can be seen as a very basic research object, but it becomes considerably more reusable when augmented with the knowledge of the context of its use.

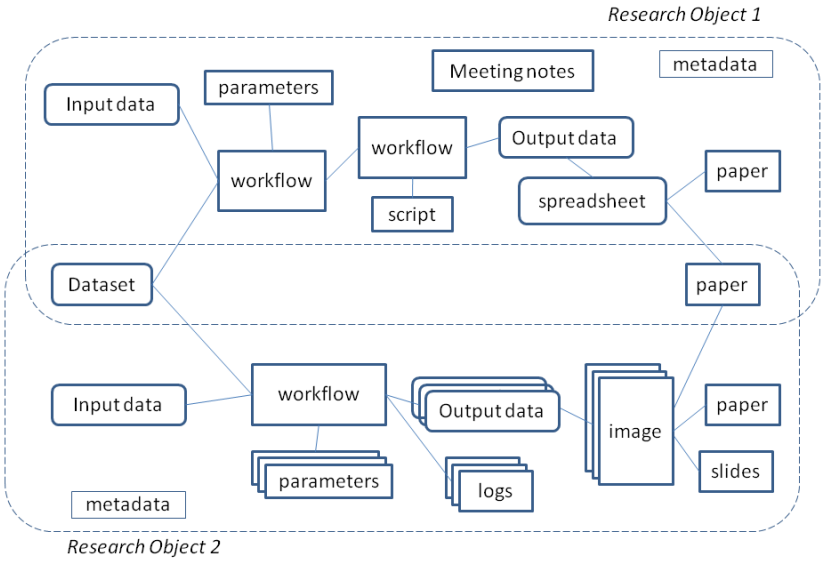

Figure 1. Two Research Objects (ROs) representing data analyses

Figure 1 gives an example of two ROs representing analyses of the same dataset. In Research Object 1, the dataset is processed by the first workflow in conjunction with some additional input data and parameters, and then passed for statistical analysis by a workflow with a R processor, the results of which are discussed in a meeting, stored and placed in a spreadsheet which is then used in two papers. The analysis in research Object 2 is more automated, with multiple runs of a single analysis pipeline using different parameter sets, resulting in

visualisations which are used in papers and slides, together with secondary data and the logs which explain exactly which services were used. Note that ROs are descriptions of aggregations, the components of which may be distributed and may be shared with other ROs: both ROs in this figure contain the original dataset and a paper that includes the results of both analyses.

Research Objects have five key characteristics:

1 **Encapsulated**. They contain typed interrelationships and dependencies between resources and are in turn labeled and identifiable as an individual resource. We reuse vocabularies and schemas such as FOAF and SIOC for representing the social network; Creative Commons for contribution licenses and Dublin Core for common metadata properties.

2 **Distributed**. They are structured collections of references to locally managed and externally located resources, which has implications for reliability, consistency, mixed stewardship, versioning and identity resolution.

3 **Annotated**. They carry metadata concerned with their provenance profile, lifecycle profile, sharing profile (permissions, licensing, downloads, views), curation profile (tags, comments, ratings) and usage profile (coreferencing, co-searching etc).

4 **Repeatable**. They capture information about the lifecycle of the investigation (for example provenance information about analyses), facilitating the ability of experiments to be *repeatable* (without change), *reusable* (with reconfiguration), *replayable* and/or *repurposable* (as new components or templates) [ref antoon].

5 **Interoperable**. They are publishable and exchangeable units that facilitate interoperability; for example, by using the OAI-ORE standards we increase interoperability and facilitate the consumption of Research Objects in between applications.

In [Barend Mons], the problem of "knowledge burying" is highlighted, where knowledge about investigations or experiments is published in paper form, and text mining techniques are required to extract this knowledge, leading to inefficient transfer of information (See Figure burying). A view of "Research Object as publication", packaging and associating data, results and methods as part of the publication process helps to overcome some of these issues by ensuring that information and knowledge is not lost during that publication process.

Research Objects are core components produced and consumed by an e-Laboratory, and the currency of exchange between e-Laboratories. We discuss the notion of an e-Laboratory in Section 5.

1.2 Social infrastructure for collective intelligence

Accelerating time to experiment requires *social infrastructure*. A data and method deluge demands new techniques, especially in the context of open science, where primary research data are posted which can be added to/interpreted by anybody who has the necessary expertise and who can therefore join the collaborative effort [Wikipedia def]. The Open Science movement [ref], though currently niche, vocally advocates the large scale, open distributed collaboration is enabled by making data, methods and results freely available on the Web. The new instrument that we bring to bear on this challenge is provided by *society itself* – it is the scale of community participation and the network effects that this brings. This instrument offers new ways of tackling difficult challenges; for example, the ‘decay’ over time of research objects as methods become obsolete or data outdated can be addressed by community curation.

Hence there is great potential in providing social tools to support the scientific process and the sourcing, sharing and continued curation of scientific research objects [wikinomics article]. This is possible because increasingly (a) the various research objects are born or available digitally and (b) a new generation of scientists who are digitally native. Scientists are just beginning to use blogs, wikis and social networks to facilitate more rapid and immediate sharing of research, a phenomenon sometimes characterised as Science 2.0 [ref, ref]. We propose that:

- *By adopting social content sharing tools* for Research Object repositories we can harness a social infrastructure that enables social networking around scientific objects and provides community support for social tagging, comments, ratings and recommendations and social network analysis and reuse mining (what is used with what, for what and by whom), and remixing of new research objects from previously deposited ones. We can take advantage of popular and familiar user interfaces of social content sharing sites such as Flickr, YouTube and Slideshare [refs].
- *By adopting an open, extensible and participative development environment* for research object repositories functionality can become readily available for reuse by others and draw on other services as much as

possible. Open science is the process of opening up content (sharing research objects in controlled and appropriate ways) and opening up applications (sharing research objects and the functionality of their repositories with applications). We should not oblige the scientist to come to a repository, but rather make it as easy as possible to bring the content to the scientist's own environment. This is essential for adoption [ref], which in turn is essential to build a community and catalyse community network effects.

1.3 myExperiment

We have put this thinking into practice in the creation of myExperiment [ref], a socially-sourced content repository that supports the sharing and curating of methods-based Research Objects used by scientists, specifically focused on scientific workflows and experiment plans. For researchers it provides 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. This paper describes "the experiment that is myexperiment" by examining its three key capabilities:

- *Facilitates the management and sharing of research workflows.* The public repository (myexperiment.org) has established a significant collection of scientific workflows, spanning multiple disciplines (biology, chemistry, social science, music, astronomy) and multiple workflow systems, which has been accessed by over 16,000 users worldwide. At the time of writing[1] the public site has over 600 different workflows (200+ versions), drawn from XX workflow management systems including Taverna [ref], Kepler [ref], Triana [ref], and Trident [ref]). There are 1600 registered users. In section 2 we introduce myExperiment, briefly present our development methodology and compare our work with other method repositories.

- *Supports a social model for content curation tailored to the scientist and community.* Producers of research objects should have incentives to share and consumers need to be able to discover and reuse them; all should benefit from self- and community-curation. myExperiment has proved to be a fruitful study environment for social scientists [ncess ref]. In section 3 we describe the social model that myExperiment implements and discuss it in practice as identified by a user study that has shadowed and steered the development of the repository. In particular we show that the content is roughly split into a market and a toolbox; and that sharing is desirable and possible but anonymous reuse is challenging. We compare our social content approach to other content services in science and outside science.

- *Supports open science by exposing its content and functionality into user tools and applications and absorbing other interfaces.* myExperiment provides an open, extensible environment to permit ease of integration with other software, tools and services, and benefit from participative contribution of software. In contrast to social web sites like Facebook and mySpace, developers can download, reuse and repurpose myExperiment itself, and the codebase is evolving as it is used across multiple projects. In section 4 we show how, by exposing the myExperiment functionality, new interfaces have been built and existing interfaces have incorporated myExperiment functionality, including plug-ins to the Taverna workbench, Facebook applications, an iGoogle gadget-based research dashboard and a Silverlight interface, and Chemistry Electronic Lab Notebooks and 'blogging the lab' [ref]. We describe how myExperiment makes Research Objects accessible and *actionable* beyond the core repository using Semantic Data Web techniques, social networking practices and standard APIs from a range of communities.

- *Establishes Research Objects and the e-Laboratory.* We conclude in section 5 by discussing myExperiment's role as a first step towards a general notion of Research Objects and a greater vision of interoperable e-Laboratories. We envisage that the scholarly publishing process will evolve to support this more general notion of scientific research object, which will facilitate reusable and reproducible research.

2. MYEXPERIMENT – A COLLABORATIVELY SUPPORTED WORKFLOW REPOSITORY

myExperiment was motivated by an observed need to share workflows; see [4-5] for more on our rationale and [6] for our design methods. We set out to build an attractive and immediately understandable rich web experience that uses the metaphors and behaviours of the popular social content tools used in everyday life but is closely tailored to the different needs of scientific researchers – for example, careful attention to issues of attribution, credit, licensing and fine control over sharing. The system provides a distinctive combination of several facets which themselves are demonstrated by other systems:

- *A federated set of repositories for methods-based Research Objects*. myExperiment implements a realisation of the Research Object as a *Pack* of resources relevant to scientific workflows or plans, independent of any workflow system. Our public web site is one instance of myExperiment; other instances are being customised and instantiated for Astronomy and Chemistry. The architecture and adoption of

[1] 25th March 2009

persistent, real URLs for ROs, standard protocols and rich RESTful APIs supports federation, interoperability and inter-system referencing/bookmarking. Other workflow repositories like *Kepler's Hydrant* (www.hpc.jcu.edu.au/hydrant) and *Inforsense's* commercial *Customer Hub* (www.chub.inforsense.com) are tied to a particular workflow system and do not offer programmatic access to the workflows. Pipeline Pilot, a popular workflow engine for cheminformatics, allows sharing of workflows through its "Accelrys community" website at http://accelrys.org/ [D15]

- *An open Virtual Research Environment for Social curation of Research Objects*. myExperiment is not intended to be a general social networking environment for scientists like Twine, SciSpace, BioMedExperts or Nature Networking. The focus is on social networking around shared artifacts (i.e. the Research Objects). In this way it is more like the social bookmarking systems like CiteULike and Connotea, but with a much wider and richer remit than published articles, or social content systems like YouTube, SlideShare and Flickr. As one would expect of any social content system, myExperiment supports:

- *An execution platform for Research Objects*. In the same way that *Kepler's Hydrant* supports workflow execution, so myExperiment provides a platform for executing or rendering the contents of its Research Objects. We 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, for example, 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. This is in contrast to *Nanohub* (www.nanohub.org) a portal that 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. *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 but it does not provide social infrastructure to support sharing of workflows. 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.

2.1 The myExperiment Research Objects

Viewed as a repository we can classify myexperiment content into four categories:

- *Primary content*: these are the chief scientific commodities that are deposited, published and exchanged. There are currently two categories: workflows, represented natively in various XML formats and associated thumbnail images dependent on their system, and files. The SysMO project has extended content to include Standard Operating Procedures (structured documents) and spreadsheets. All primary content has a unique, persistent URL.

- *External content*: these are references to content that is not deposited within the myExperiment server. This includes content on third part systems (e.g. videos, powerpoint slides, documentation, web sites etc). References to external content that is outwith the control of myExperiment raises issues of versioning and availability. Effectively, myExperiment is a mixed stewardship system in that responsibility for the stewardship of its content is distributed and outsourced.

- *Compound content* – these are the compound structured Research Objects that gather content into heterogeneous collections, called *packs*. For example: the Taverna workflow introductory pack of deposited example workflows and example data and references to eternally held manuals and user guides; the SysMO project pack of useful deposited workflows and test data that would be of value to those working in Systems Biology: the Trident workflows.

- *Metadata content* – this is the metadata attributed to the three prime content types above that describes (a) the interrelationships between the prime content and (b) key properties of the content for discovery and curation purposes. In addition to information about creation, version and description, the metadata includes citation (attribution to other research objects upon which this is based), credit to people or groups, and community contributed metadata such as tags, comments and review threads, ratings, recommendations and favouriting by registered members.

In combination these support Research Objects. For example, a workflow is treated as an aggregation of services plus associated metadata, and a pack is similarly treated as an aggregation but may have external parts; a file is local and opaque but augmented with metadata. A URL to a RO takes the user to a web page carrying all the information about the object, its components and, where appropriate, provides native content for download. Section 4 described how Research Objects can be exported in a standard RDF representation.

2.2 Design and implementation

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 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.

The interfaces are accessed via a 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. The agile 'perpetual beta' development process [IEEE software] 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. The software is released under the BSD open source licence, available from http://wiki.myexperiment.org.

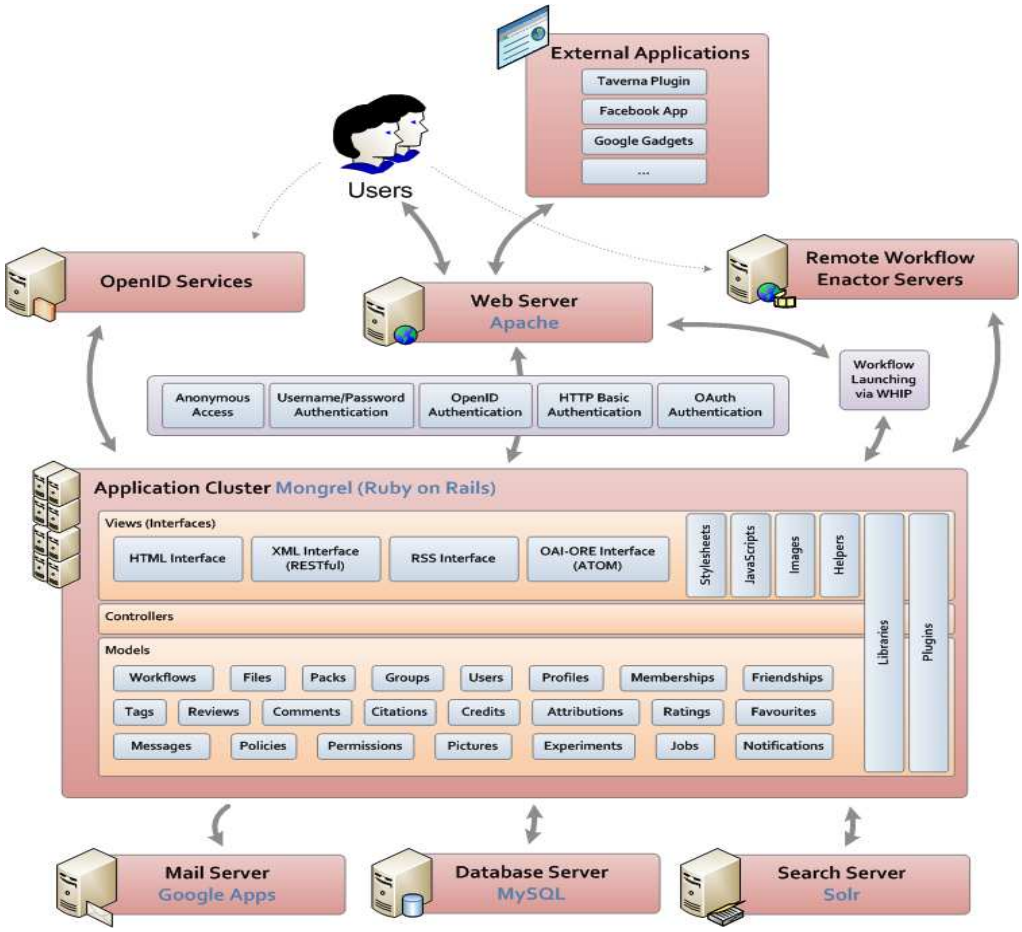

Figure 2: The myExperiment architecture

3. SUPPORTING COMMONS-BASED PRODUCTION

myExperiment relies on self-deposition by workflow designers and commons-based curation by a community of users. It is not required to login to myExperiment to browse, view and download any of the publicly published content; but it is necessary to do so to deposit content; annotate content and view restricted content. Thus we distinguish between *contributors* who create and deposit content; *editors/curators* who maintain and add to content; and *users* who take content but do not add to it or curate it. To be a contributor or curator requires membership. At the time of writing, myExperiment.org has 1626 activated membership 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. 34% of the registered users are return visitors[2]. In a one month period[3] the site received 13681 page views in 13500 visits by 2397 unique visitors. As with other social content sites, the number of unique visitors is much larger than the number of registered members, and a small fraction of members contribute content or actively curate their own or others' content. The figures do suggest that the publicly visible content on the site is of value to a wide audience, but that audience is not interested in content deposition.

Figure 3 illustrates our content and curation cycle, and the various stakeholders involved.

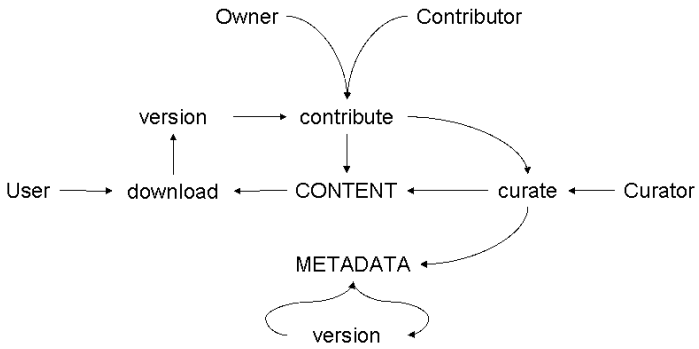

Figure 3. The content and curation lifecycle in myExperiment

In partnership with the National Centre for e-Social Science in the UK, we have conducted an ongoing investigation of our users' sharing and re-use practices, their motivations and their concerns. We conducted a series of interviews with registered users to provide a longitudinal perspective over a period of 24 months. Interviewees were selected on the basis of their activity profiles, including workflows uploaded/downloaded; number of friends; group membership; group moderation and discipline, and recruited either via myExperiment or by "snowball sampling" (i.e., users suggested by interviewees). To date, we have conducted 34 interviews with 27 users; one user has been interviewed three times; five users interviewed twice, and the rest have been interviewed once. All interviewees report successfully using myExperiment for publishing and disseminating workflows, citing *personal benefits* of convenience and dissemination, and *collaborative benefits* of sharing scholarly work and benefiting from network effects.

3.3. Community Contributed Content – "Just Enough Sharing"

Commons-based content production requires built-in incentive models for contribution. Scientists will share when there is a competitive advantage that does not damage their own competitive edge [cite Heather Pwowars work on sharing]. We identified several key drivers which have led us to create a "just enough sharing" model where control is placed in the hands of the contributor. *Credit and attribution* support and fine control over the visibility and sharing of research objects were identified early on to be the most critical factor in making a social web site acceptable for use by scientists. The myExperiment contribution model supports this need, which allows the contributor (the owner or a third party) to control the view/download and edit permissions on content Credit and attribution propagate through versions and attribution chains, though this raises the issue of "workflow drift" when the workflow has evolved to the point that it has become a new workflow. We also support creative commons licensing with clear copyright statements.

Professional reputation building is crucial to a scientist. Credit and attribution mechanisms provide one means to build a citation profile. Other methods include accurate records of downloads and views, and records of who

[2] Figures are collected using Google Analytics and do not include accesses made via the API

[3] Feb 16th 2009-March 18th 2009

viewed; the former we report, the latter we do not for privacy reasons. *Professional reputation protection* is the flip side, as scientists are concerned that their work may be misinterpreted, misused or open to unwelcome scrutiny. Consequently, we provide mechanisms for contributing rich metadata to describe how to use deposited workflows, examples, example data, references to documentation and papers etc. We also encourage an ethos of constructive comment through discussion threads. Reputation protection also raises the issue of *liability*; that is concerns that workflows might be flawed or be poorly used and their authors liable for subsequent flawed results. Thus liability disclaimer policies are important to reassure contributors, though they do not reassure consumers, as are take-down policies for workflows that have been contributed but not by their authors and possibly against their wishes.

Premature publication and thus being "scooped" by giving away valuable insights and know-how to rivals is a real obstacle to sharing. myExperiment supports incremental publication model by which a contributor can deposit their content embargoed (effectively using the site as a private archive) and reveal content to selected members and groups and finally publicly when the time is appropriate. Some communities go so far as to install their own private instance of myExperiment that supports their own policies; again our federated design means there is a path for later publication to the public instance.

At the time of writing, of the 648 workflows, ??? are publicly visible whereas 493 are publically downloadable. ??% of the workflows with restricted access are entirely private to the contributor and for the remaining they elected to share with individual users and groups. ?? workflows (over ??%) 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 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. The most viewed workflows has 2924 views and the most downloaded 3493 downloads - hence the pollution!!! 106 workflows have never been downloaded (which of these are real?).

3.3. Collaborative Curation

Unless they are annotated with metadata, workflows (and other ROs) are difficult to find, correctly interpret and understand and use without resorting to contact with the author (who may or may not be the contributor). The idea is that useful ROs will be curated by the community that uses them, and original authors are encouraged to curate because they are getting credit for use of their work. Through user feedback, blogging, e-tracking, recommendations and "folksonomy-based" tagging and so forth we leverage community to collaboratively self-manage these shared assets.

Quality and sufficiency of good documentation is accepted as a key requisite for facilitating sharing and re-use. The metadata needed to find a workflow is much less rich than the metadata needed to actually use it. From our interviewees' comments, the community is still learning what constitutes good documentation for workflow discovery and sharing, missing out core descriptions such as input and output data types and formats and making too many assumptions. Metadata is time-consuming to produce, and requires an author to imagine what an unknown stranger with unknown skills would need to know. *Social solutions to incomplete documentation* exploit the social networking and commenting facilities to start up a dialogue with the contributors, forming collaborations.

Contributor-dominated curation dominates in that the majority of metadata is supplied at the point of contribution and by the contributor. Little is supplied post-contribution and only a small number of registered users curate or edit metadata associated with workflows they have not contributed. This is in line with the finds of other social content sites. *Tagging practices* are evolving and have yet to establish best practice. The vocabularies used for tags can quickly become unruly without the enforcement of controlled terms and practices. Tag clouds and suggested tags are used, auto-tagging through workflow-specific parsers help, and tags tend to be objective ("text mining" rather than subjective ("nice"). However, tags are not sufficiently discriminatory; tagging practice needs to be established and standardized and tag terms need to be harmonised controlled.

Content decay surveillance is necessary as 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 the way genes are identified by one service provider led to a myExperiment announcement for users of the affected workflows. myExperiment provides a content surveillance and notification forum to channel changes the majority of which can be automated, though not all.

Incentives for curation are similar but subtly different to those of content deposition. We need to encourage both contributors and potential editors to add metadata and continue to add metadata and we need to automatically gather information (usage, co-usage patterns, etc). The more we gather incidentally the better. The rewards and fears discussed in section 3.2 apply, so we need to create reputations for best curated or most effective curator; nanoHub has pioneered competitive curation using real prizes, and other proposals include "strong password"

bars and metadata league-tables. Comments are actively used but, disappointingly, ratings are not. We speculate this is due to a number of things: reticence to publicly criticise; poor metadata leads to inability to effectively rate; and the requirement to return to the site to make the rating. We thus need to gather curation metadata at the point of use (for example while running a workflow in a system) and through other systems (for example, social book marking systems or Google gadgets). This latter point is one of the drives for or open platform, discussed in section 4. Finally, we built a critical mass of curated content by cultivating core groups of discipline-specific active advocates and employing expert curators whose role is to annotate and maintain content and set up the curation pipelines for content that is not of their making. The phenomenon of a coterie of editors, sometimes self-appointed, is common in social content sites such as Wikipedia, and is crucial to building consumer confidence.

3.4. Reuse and Re-Sharing workflows

Reuse happens – lets make this upbeat cos the rest is critical! – how much reuse? The incidence of attribution is low and anecdotally we observe that users download workflows and use them but do not return to post comments, nor do they return to re-contribute adaptations that would attract attribution. Unsurprisingly, the ability to find workflows is directly correlated to the quality of their metadata.

Two distinct myExperiment communities have emerged when it comes to workflow re-use, which we characterise as *supermarket shoppers* and *tool builders*. Workflow consumers prefer larger workflows ready to be down loaded and enacted; workflow authors prefer smaller, modularized workflows which can be assembled and customized. Workflow consumers see myExperiment as a workflow 'supermarket' whereas workflow builders see it as a 'toolbox'. Larger workflows are usually specific and complex, more likely to be difficult to understand and yet poorly documented and thus difficult to adapt; smaller workflows are typically self-contained, coherent units undertaking one task. *Domain parochialism* suggests that workflows do not easily migrate across domains, reflecting distinctive 'patterns' to research processes in different domains. Many interviewees also commented that their research is relatively advanced or is too specialised for many workflows to be directly helpful to them. This may reflect that the myExperiment community is still evolving and, as yet, is populated by early adopters, such that effects normally attributable to social networks have yet to make themselves felt. Both these points have implications on contribution in encouraging better quality metadata, encouraging contributors to adopt better workflow design practices that enable them to be reusable, and give them the tooling to support this. Designing a good workflow is hard enough; designing one to be reusable is much harder. It is an aim of myExperiment that by gathering cohorts of workflows we can mine patterns and improve design [ref Groth and Gil 09, ICDT09].

Although *anonymous reuse* (i.e. the author was not contacted by the user) is observed (how ,many?) for "toolbox" workflows, *negotiated reuse* has emerged as common practice for the "supermarket" workflows. This is in part because of a lack of adequate documentation and the complexity of the workflows, but is also underpinned by the social interaction, enabling users and authors to communicate, and a desire for control on the part of authors. Returning to an earlier point, the author needs to trust that a user will use their workflow properly and one way to control this is to force them to communicate by making the workflow attractive but un-reusable without communication. This may be tacit behaviour as popular workflow authors complain about the increase in communication traffic that they encouraged, although this in turn leads to improvements in the metadata for those workflows. The flip side of author trust is *consumer reassurance* to satisfy a potential user that a particular workflow matches what they are looking for and works reliably. Discussion with the author is one direct method; peer review, usage popularity validation authorities; and judgement based on the quality and richness of the Research Object are all evidence. There are a number of different theoretical approaches to the study of trust (Axelrod, 1997; Kipnis, 1996; Luhmann, 1979; 1990), Luhmann suggests that most approaches fail to pay attention to "the social mechanisms which generate trust" (1990:95). The key point is that there is no single mechanism which, a priori, guarantees a solution. What is important, as we see from our study, is that there are a range of 'trust affordances' to hand when trust becomes a practical issue for myExperiment users. myExperiment facilitates social interaction, enabling users to switch relatively seamlessly from workflow to workflow author and back again.

4. OPEN SCIENTIFIC PLATFORMS

The myExperiment website provides is designed to be easy to use for those discovering and sharing ROs, but we do not oblige the user to go to the site: we have given just as much attention to designing the site so that its functionality can be delivered directly to the user in their work environment. This is achieved by providing interfaces for developers, and also by making the myExperiment public content available in RDF so that we can add knowledge directly to the Web.

4.1 API

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. 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.

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.

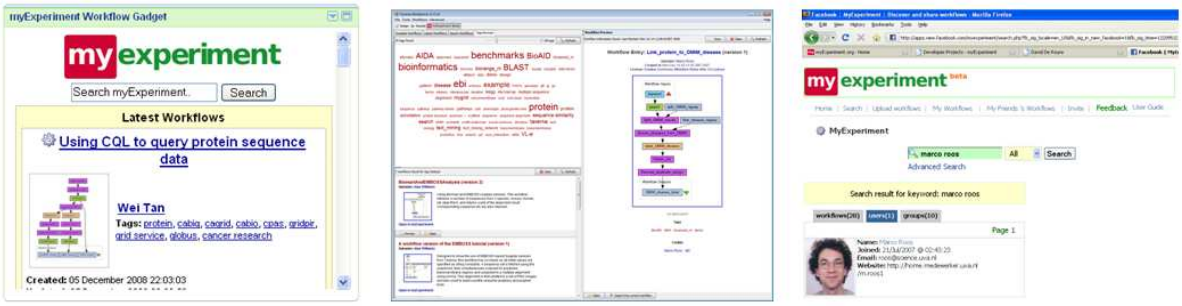

Figure 4: Interfaces to myExperiment that use the API – Google Gadget, Taverna plugin and Facebook application

A developer community is growing up around the API, developing new user interfaces and bringing myExperiment through into existing interfaces. Three interfaces are illustrated in figure 3.

Developing new Interfaces. We have several exercises in building entirely new user interfaces to myExperiment's functionality. Firstly we have built Google Gadgets for myExperiment, creating separate interfaces to myExperiment capabilities. Secondly we are building functionalit mashups, using Silverlight to build a rich similarity search and socially-driven workspace mashup that uses the myExperiment API together with other common data sources like Google Search, Google Scholar, CiteULike, Connotea, PubMed and so on. Silverlight, in a similar vein to Adobe Flash, is an extension to the browser in which rich content and functionality can be provided to users. Our search mashup presents a clean interface that allows a user to focus on discovery without being distracted by the other features of myExperiment. We have used the keyword search and tag cloud functionality (via the API) to allow discovery of all public content from the myExperiment.org repository.

Bringing myExperiment to existing interfaces. We have integrated with the Taverna workflow workbench by building a Taverna plugin for myExperiment, so that Taverna users can access the myExperiment capabilities from within the Taverna environment. We have also integrated myExperiment as an application inside Facebook. We are currently integrating with Microsoft's Trident Scientific Workflow Workbench [6], and for this we have developed preliminary support in myExperiment for sharing Windows Workflow Foundation (WWF) workflows. Finally we are working in conjunction with our open science colleagues in chemistry to bring myExperiment together with work on Electronic Lab Notebooks and 'blogging the lab' [7].

4.2 Publishing knowledge to the Web in RDF

myExperiment's ability to share information is one of its key advantages when it comes to closing the experimental lifecycle loop. However it is important to consider the mechanisms for how this information is shared. The myExperiment RESTful API has already demonstrated how a machine-oriented sharing mechanism

can allow the development of new interfaces in the form of "mashups" and "gadgets". The RESTful API although extensible is quite rigid and requires any linking-up to be performed client-side, RDF provides a framework so this can be executed server-side.

myExperiment publishes all its public data as RDF at http://rdf.myexperiment.org/. RDF has a very simple subject-predicate-object (triple) structure that facilitates linking-up but these relationships can be formalized using a meta-structure provided by a schema or ontology. By formalising, additional information can be inferred rather than having to define it explicitly. myExperiment uses a modularized ontology set to provide its formalisation (http://rdf.myexperiment.org/ontologies/). The myExperiment ontology is modularised to promote reuse with each module designed for a specific sub-domain, e.g. types of annotation/contribution, attribution and creditation, packs, experiments, etc. As well as promoting reuse, the myExperiment ontology reuses parts of other ontologies/schemas: FOAF and SIOC for representing the social network, Creative Commons for contribution licenses, Dublin Core for common metadata properties and OAI-ORE for representing packs.

4.3 Ontology Modules Architecture

Through reuse it is possible to make some sense of myExperiment data outside its domain, allowing data from different sources to be collated. By making the myExperiment ontology reusable its saves reinvention and allows similar projects to map their data in the same way. Significant effort is being given to representing experiments and the data they produce in such a way that their insights can be shared across multiple fields. The Scientific Discourse subgroup of the W3C's Health Care and Life Sciences (http://esw.w3.org/topic/HCLSIG/SWANSIOC) has been considering how to reconcile a number of ontologies that treat experiments as first class objects.

A SPARQL endpoint is a way of providing this server-side linking-up. By collating all myExperiment's RDF data in a single data structure, known as a triplestore, it is possible to provide a very flexible querying interface using the Semantic Web querying language SPARQL. myExperiment's SPARQL endpoint (http://rdf.myexperiment.org/sparql) allows the execution of simple queries that could be performed by a RESTful API call with the appropriate parameters, such as returning all the workflows uploaded by a specific user. However, if you wanted to further restrict that to those you had also commented on, the RESTful API would not be able to do this without additional functionality being added.

SPARQL queries are essentially trying to map networks where one or more of the nodes or links are unknown. myExperiment's RDF provides a listing of components (sources, sinks processors and links) for Taverna workflows; SPARQL provides a facitlity for searching for workflows where these components link up in a specific user-defined way. This makes it possible to find a workflow that is much more tailored to a searcher's requirements, which becomes ever more important as the number of workflows grow.

Returning SPARQL results is an appropriate format if this data is to be used within a Semantic Web application, however this is often not the case. Results may need to be ported in a more generic way or understood by a real person. The SPARQL endpoint allows results to be exported as comma-separated values (CSV) or visualized as an HTML table. Other more specific use cases have also arisen, in particular a capability to represent SPARQL queries that return mappings (e.g. between users who are friends) as a matrix that can be exported as CSV. It is important to understand that although RDF and SPARQL are Semantic Web technologies this should not prevent them from working seamlessly with applications that are not.

5. DISCUSSION AND FUTURE WORK

myExperiment is an important component in the revolution in creating, sharing and publishing scientific results, and has already established itself as a valuable and unique repository with a growing international presence. It demonstrates the success, and exposes the challenges, of blending modern social curation methods with the demands of researchers sharing hard-won intellectual assets and research works within a scholarly communication lifecycle.

The results of our study myExperiment users reflect a community in formation, whose members' attitudes, intentions and expectations are diverse and evolving. As such, significant changes in attitudes, intentions and expectations are likely to occur as the activities of community members filter through the networks of relationships, reinforcing successful innovations and defining good practice.

As we have developed the functionality of myExperiment, and used myExperiment within a variety of other projects, we have begun to identify the reusable components and resources of myExperiment itself. For example, the Biocatalogue website provides a registry of web services in the life sciences and borrows directly form myExperiment's curation models. The Biocatalogue services can be seeded from myExperiment's workflow collection; in turn, myExperiment can import functionality from Biocatalogue.

Hence we envisage the *e-laboratory* (or *e-lab*), a set of integrated components that, used together, form a distributed and collaborative space for research, facilitating the planning and execution of *in silico* experiments. An e-lab brings together people, materials and methods in order to support scientific investigation. ROs play a role both in driving the components and capabilities within an e-lab. For example, the internal structure of an RO can be used within a workbench to determine appropriate visualisation methods for the contents of the RO. ROs are not, however, simply internal to a particular e-lab platform – they will also play a role in sharing/communicating not just between services and components within an e-lab, but also with other e-labs (or e-labs services). See Figure 5.

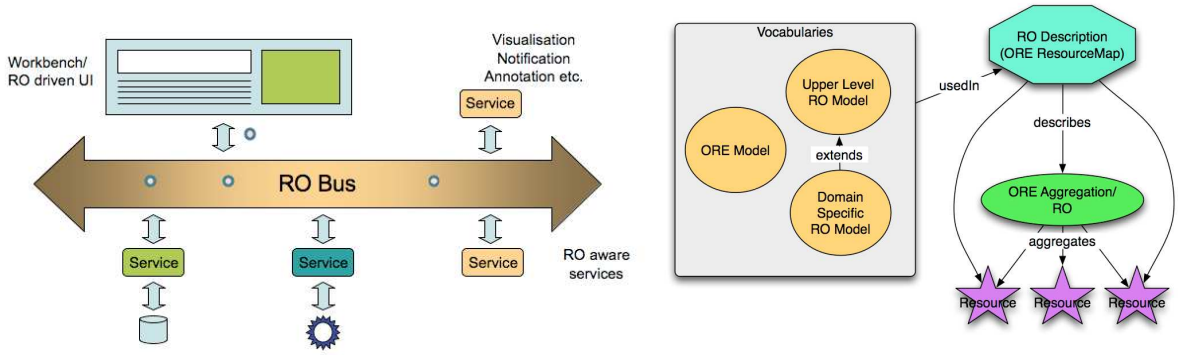

Figure 5: The Research Object Bus

We have argued for the sharing of methods and the combining of methods with results, in pursuit of open science in which the facility of exchange leads to enhanced scientific outcomes. We have introduced the notion of the Research Object, which encapsulates methods and results and thereby enables repeatable, replayable and repurposeable research. myExperiment illustrates an early repository of ROs which majors on the social dimension, and we have demonstrated that an online community and workflow collection has been established.

Our future plans involve more extensive discovery capabilities through autotagging, controlled vocabularies and recommendation. We are pursuing further repository integration, greater componentization, and support for interworking with Research Objects, as we progress towards our vision of the e-Laboratory.

ACKNOWLEDGEMENTS

The design of myExperiment and Research Objects has been a collaborative exercise involving a large group of people including Mark Borkum, Les Carr, Simon Coles, Phil Couch, Catherine De Roure, Tom Everleigh, Paul Fisher, Jeremy Frey, Antoon Goderis, Matt Lee, Cameron Neylon, Savas Parastatidis, Meik Poschen, Marco Roos, Robert Stevens, Shoaib Sufi, Franck Tanoh, David Withers, Katy Wolstencroft. myExperiment is funded by JISC, Microsoft Technical Computing Initiative and EPSRC.

REFERENCES

Will be managed in endnote – just drop text inline or here for now so I know what they are!

[Barend Mons] B. Mons. Which gene did you mean? BMC Bioinformatics 6(), p.142 2005 DOI: 10.1186/1471-2105-6-142

[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
[D12] http://dx.doi.org/10.1186/1471-2105-7-260
[D13] http://dx.doi.org/10.1371/journal.pcbi.1000136
[D14] http://pubmed.gov/18428093
[D15] http://dx.doi.org/10.1016/j.drudis.2008.03.015
Axelrod, R. (1997). Complexity of Co-operation: agent based models of competition and collaboration. Princeton. NJ. Princeton University Press.
Fleck, J (1993). Innofusion: feedback in the innovation process. In: Stowell, S.A., West, D. and Howell, J.G. (eds.) Systems Science: addressing global issues, Kluwer Academic / Plenum Publishers, pp. 169-174.
Fogg, B. J. and Tseng, H. (1999). The elements of computer credibility. In Proceedings of CHI 99, New York, NY: ACM, pp. 80-87.
Kipnis, D. (1996). Trust and Technology. In R. M. Kramer and T. R. Tyler (eds.): Trust in Organizations: Frontiers of Theory and Research, London: Sage, pp. 39-50.
Luhmann, N (1990). Familiarity, Confidence, Trust: Problems and Alternatives. In Gambetta, D, (ed.): Trust: Making and Breaking Cooperative Relations, Oxford. Basil Blackwell.
Piwowar, H., Day, R. and Fridsma, D. (2007). Sharing Detailed Research Data Is Associated with Increased Citation Rate. Nature Precedings : doi:10.1038/npre.2007.361.1
Williams, R., Stewart, J, Slack, R. (2004). Social Learning in Technological Innovation, Cheltenham, Edward Elgar.