CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE

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

Towards Open Science: The myExperiment approach

David De Roure[1,*], Carole Goble[2], Sergejs Aleksejevs[2], Sean Bechhofer[2], Jiten Bhagat[2], Don Cruickshank[1], Paul Fisher[2], Duncan Hull[3], Danius Michaelides[1], David Newman[1], Rob Procter[4], Yuwei Lin[4], Meik Poschen[4]

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

[2] *School of Computer Science, The University of Manchester, Manchester M13 9PL, U.K.*

[3] *The Manchester Interdisciplinary Biocentre, The University of Manchester, Manchester M13 9PL, U.K.*

[4] *National Centre for e-Social Science, The University of Manchester, Manchester M13 9PL, U.K.*

SUMMARY

By making research content more reusable, and providing a social infrastructure which facilitates sharing, the human aspects of the scholarly knowledge cycle may be accelerated and 'time-to-discovery' reduced. We propose that the key to this is the sharing of methods and processes. We present myExperiment, a social web site for discovering, sharing and curating Scientific Workflows and experiment plans, and describe how myExperiment facilitates the management and sharing of research workflows, supports a social model for content curation tailored to the researcher and community, and supports Open Science by exposing content and functionality to the users' tools and applications. Based on this 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 research that is more reusable and repeatable and hence rapid and robust.

KEY WORDS: *Scientific Workflow, Web 2.0, Data Curation, Research Object, Semantic Web, 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 social processes in which scientists share hypotheses, insights and results, and the data and methods that support these. Traditionally, scholarly discourse and dissemination have focused on peer reviewed journal articles, mediated by the scholarly publishing process and gatherings such as conferences where researchers exchange knowledge in more informal ways. The Web is now widely used as a distributed platform for the dissemination of an increasingly diverse range of digital research materials: we are witnessing evolving practice in scholarly publishing [1] and communities supported by research portals and repositories. Significantly, there are also now tens of thousands of publicly available web services across business and science [2]. In this evolving landscape we observe an expansion in the kinds of scientific commodities being published, for example:

- *Primary and secondary data sets*, along with standard metadata sufficient to support their interpretation and re-use, although tying together published results with the "supplementary data" upon which they are based has unsolved issues to do with persistence [3].
- *Algorithms, software tools, scripts and procedures*, through community services like OpenWetWare [4], which provides an exchange for techniques in biological sciences, and the nanoHUB gateway [5] which hosts user-contributed resources in the nanotechnology domain.

* Corresponding author. Email dder@ecs.soton.ac.uk

This latter point is the focus of our work. Researchers need to share (and find) not just the digital materials of research but also the *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, life cycles and hence curation needs [6]; as with data and articles, they are subject to their own forms of authorship, credit and reuse criteria. We propose that:

- *By pooling and sharing methods* we have the potential to accelerate science through exchanging know-how and best practice, avoiding reinvention and hence reducing time-to-experiment. Moreover, participating researchers 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 research [7] and maintain the robustness of the accelerated process. 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.

1.2 Workflows

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 digital research methods, and one that has gained popularity and adoption in a short time [8].

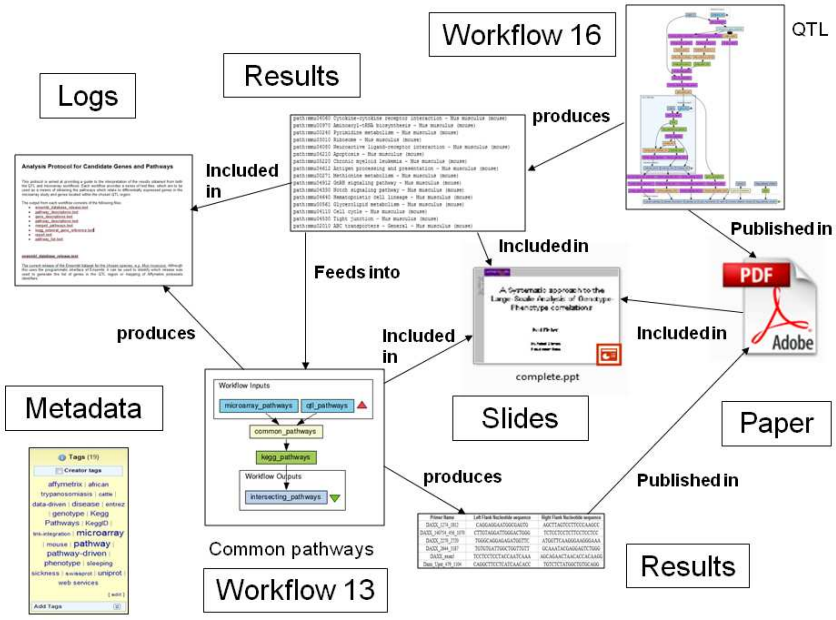

Figure 1: Workflows and associated items used in the production of a research article

Workflows can require specialist expertise that is hard-won and may be outside the skill-set of the author, and they are often complex and challenging to build [9]. Figure 1 illustrates a piece of research which involves two workflows developed for a particular bioinformatics investigation (investigating the Trypanosomiasis resistance phenotype in the mouse model) which led to publication of an article in Nucleic Acids Research [10]. The suite of scientific workflows in this work took a bioinformatics expert six months and over 40 versions to develop; however, once developed they were immediately reusable by other, perhaps less experienced, researchers – in turn accelerating their research.

In addition to the workflows and the pdf we see all the supplementary information relating to the published paper, including all workflow outputs, Word documents on result interpretation, spreadsheets detailing the re-sequencing of one candidate gene and a table from the paper itself, a PowerPoint presentation outlining the project's background, and descriptions of the work carried out so that the provenance of the results can be established. In combination these items enable the research to be repeated, the research outcomes to be properly interpreted and trusted, and the components to be better repurposed.

1.3 Social Infrastructure

The benefits which we have outlined will not be achievable without a *social infrastructure* to facilitate the pooling, sharing and combining of methods and resources. A data and method deluge demands new techniques, especially in the context of Open Science, where primary research data are posted that can be added to/interpreted by anybody who has the necessary expertise and who can therefore join the collaborative effort. The Open Science movement [11, 12], though currently niche, vocally advocates the large scale, open distributed collaboration that 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 as workflows become obsolete or data outdated can be addressed by community curation.

Hence, there is great potential in providing social tools to support the research process and the sourcing, sharing and continued curation of research resources [13]. This is possible because (a) increasingly the various research resources are born or available digitally and (b) a new generation of researchers are digitally native. Researchers 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 [14]. We propose that:

- *By adopting social content sharing tools* for repositories of research materials and methods we can harness a social infrastructure that enables social networking around research items 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 items from previously deposited ones. We can take advantage of popular and familiar user interfaces of social content sharing sites such as Flickr (www.flickr.com), YouTube (www.youtube.com) and Slideshare (www.slideshare.net). These sites provide excellent native functionality for particular content types.
- *By adopting an open, extensible and participative development environment* 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 researcher 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 [15] which, in turn, is essential to build a community and catalyse community network effects.

We have put this thinking into practice in the creation of myExperiment [16], a socially-sourced content repository that supports the sharing and curating of methods-based objects used by researchers, 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 three key aspects:

- *Facilitates the management and sharing of research workflows.* The public repository (www.myexperiment.org) has established a significant collection of scientific workflows, spanning multiple disciplines (biology, chemistry, social science, music, astronomy) and multiple workflow systems (12 workflow types), which has been accessed by over 24,000 users worldwide. At the time of writing the public site has over 660 different workflows (with a further 190 versions), drawn from multiple workflow management systems including Taverna [17], Kepler [18], Triana [19] and Trident [20]. There are 1680 registered users. In section 2 we introduce myExperiment, briefly present our development methodology and compare our work with other method repositories. myExperiment provides an open, extensible environment to permit ease of integration with other software, tools and services, and benefits from participative contribution of software. We show how, by exposing the myExperiment functionality, new interfaces have been built and existing interfaces have incorporated myExperiment functionality.
- *Supports a social model for content curation tailored to the scientist and community.* Producers of workflows 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 environment for studying such issues [21]. In section 3 we describe the social model that myExperiment implements and discuss some of the issues 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 what we characterise as a market and a toolbox; and that sharing is desirable and possible but anonymous reuse is challenging.
- *Establishes Research Objects and the e-Laboratory.* We conclude in section 4 by discussing myExperiment's role as a first step towards the realisation *Research Objects*, which are a more general

concept of a method-based digital research item, and a greater vision of interoperable e-Laboratories. 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. We envisage that the scholarly publishing process will evolve to support this more general notion of 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 [16, 22] for more on our rationale and [15] 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 researchers – for example, careful attention to issues of attribution, credit and licensing, and fine control over sharing amongst groups and friends. The website is illustrated in Figure 2, which shows a workflow, and the associated metadata relating to licensing and credits, framed by a tabbed interface above and a dashboard of the user's items and online community to the right.

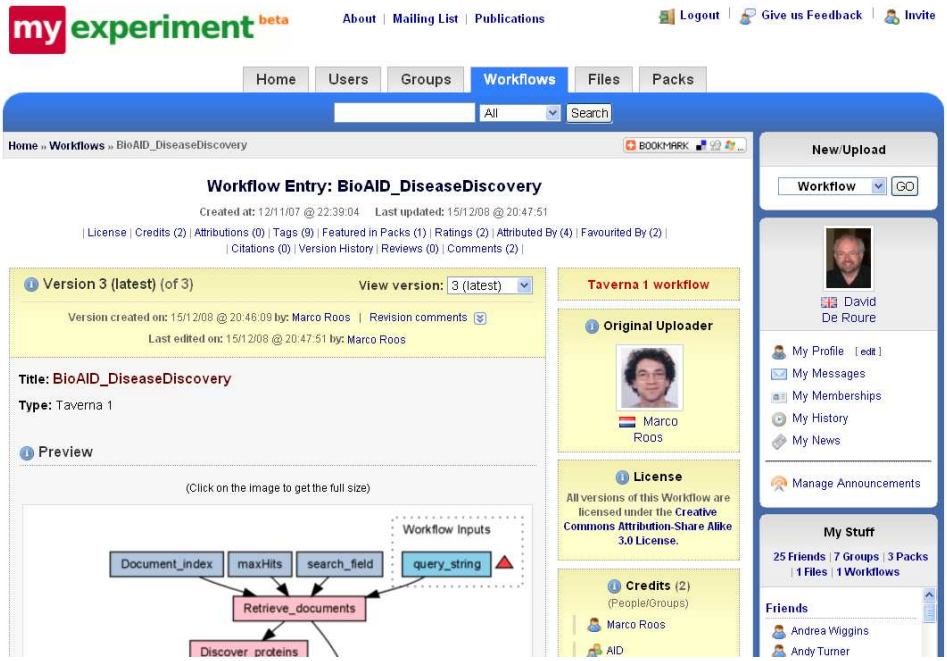

Figure 2: The myExperiment social website

The system provides a distinctive combination of several facets which are demonstrated by other systems:

- *A repository for digital research items.* Our public web site is one instance of myExperiment; other instances are being customised and instantiated for the Astronomy and Numerical Algorithms communities. The architecture and adoption of persistent URLs, standard protocols and RESTful APIs support 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 (accelrys.org) [23]. Distinctively, myExperiment implements mechanisms for creating *Packs* of resources, which are collections of research objects to form aggregate entities exactly as depicted in Figure 1 (which is available as pack 55, on http://www.myexperiment.org/packs/55). In addition to objects on the current server, packs can also contain links to objects on other servers.
- *An open Virtual Research Environment (VRE) for social curation of research items*. myExperiment is not intended to be a general social networking environment for scientists like Twine (www.twine.com), SciSpace (www.scispace.net), BioMedExperts (www.biomedexperts.com) or Nature Networking (network.naturecom). The focus is on social networking around shared artifacts. In this way it is more like the social bookmarking systems like CiteULike (www.citeulike.org) and Connotea (www.connotea.org),

but with a much wider and richer remit than published articles, or social content systems like YouTube (www.youtube.com), SlideShare (www.slideshare.net) and Flickr (www.flickr.com). It effectively creates a social network of people and the items that they share.

- *An execution platform for workflows*. In the same way that *Kepler's Hydrant* supports workflow execution, so myExperiment provides a platform for executing workflows. It offers a rich API and remote execution. This recognises that workflow authors and those who run them may be entirely different groups of users with entirely different interfaces. myExperiment is designed to provide services to a portal and also to be used as a Web 2.0 'skin' over existing services.

2.1 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. In line with our open and componentised approach, 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. We have multiple domain-specific myExperiment instances: ultimately scalability will also be achieved by federating multiple instances of myExperiment.

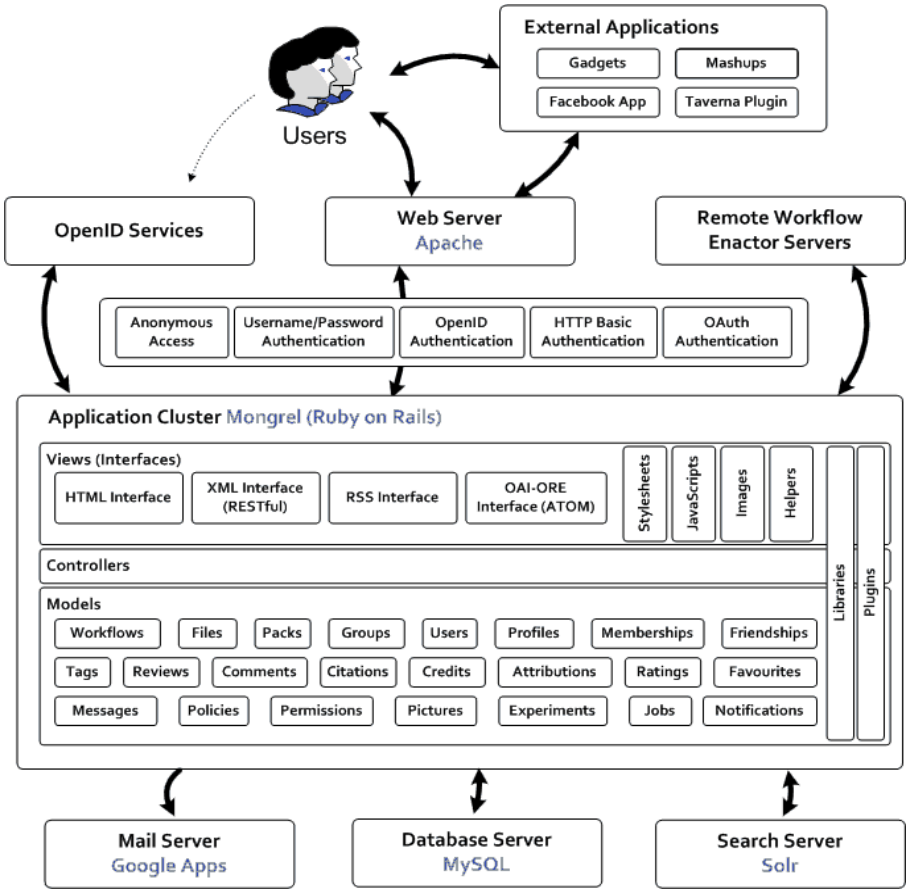

Figure 3: Architecture of a myExperiment instance

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 (http://openid.net/) or the internal username/password mechanism. The agile 'perpetual beta' development process [24] requires frequent updates to be rolled out to the main myExperiment.org service and extensive user testing, 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.

2.2 Programmatic interface

As well as bringing our social repository and VRE capabilities 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 directly within their familiar work environment. 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 (oauth.net), whose purpose is not just to authenticate that a user has given a service consumer access to a service provider; it is a specific key that may have certain privileges assigned to it. With OAuth, a user can create several keys which could be used with one service, and each of those keys may have a different set of privileges.

A developer community is growing up around the API, developing new user interfaces and bringing myExperiment through into existing interfaces. Four of these interfaces are illustrated in figure 4.

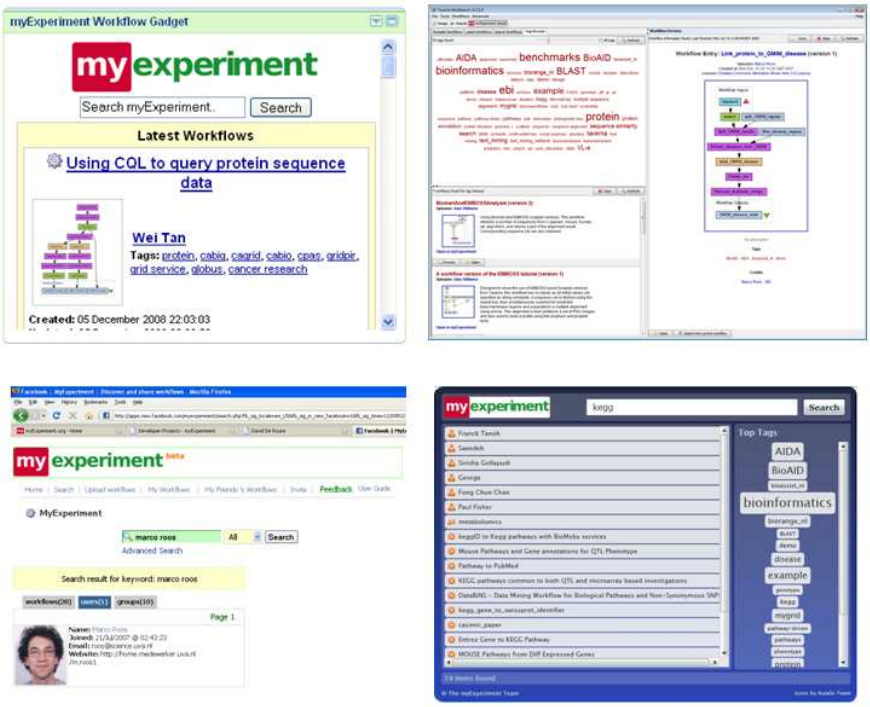

Figure 4: Interfaces to myExperiment that use the API – a Google Gadget, the Taverna plugin, the Facebook application and a Silverlight mashup.

- *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 functionality mashups, using Silverlight, an extension to the browser in which rich content and functionality can be provided to users, 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 and PubMed. 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, so that Taverna users can access the myExperiment capabilities from within the Taverna environment. Other developers have integrated myExperiment as an application inside Facebook. We are currently integrating with Microsoft's Trident Scientific Workflow Workbench [20], and for this we have developed 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' [11, 25].

Our development community is supported by the myExperiment wiki (http://wiki.myexperiment.org). A test server containing a recent snapshot of the public data from the live site is provided to developers writing applications that make use of the myExperiment API, and the public content is also available via our RDF server (see section 4). The software is released under the BSD open source licence.

3. SUPPORTING COMMONS-BASED PRODUCTION

myExperiment relies on self-deposit 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[†]. In a one month period[‡] the site received 13681 page views in 3492 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.

In partnership with the UK National Centre for e-Social Science, we have conducted an ongoing investigation of our users' sharing and re-use practices, their motivations and their concerns. A series of interviews have been conducted 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, 34 interviews have been conducted 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.

3.1. 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 [26]. 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 support creative commons licensing.

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 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

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

[‡] Feb 16th 2009-March 18th 2009

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, perhaps with a path for later publication to the public instance.

At the time of writing, of the 661 workflows, 531 are publicly visible whereas 502 are publicly downloadable. 3% 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, and 69 workflows (over 10%) have been shared, with the owner granting edit permissions to specific users and groups. In addition there are 52 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 workflow has 1566 views; 108 workflows have never been downloaded. There are 50 packs, ranging from tutorial examples to bundles of materials relating to specific experiments as in figure 1.

3.2. Collaborative Curation

Unless they are annotated with metadata, workflows and other items 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 items 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; contributors are 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 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 [27]. *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 standardised and tag terms need to be harmonised or 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 (e.g. web services) 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 gather information (usage, co-usage patterns, etc) automatically. The more we gather incidentally the better. The rewards and fears discussed in section 3.1 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 criticise publicly, poor metadata leading to inability to rate effectively, 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). Finally, we have 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.

3.3. Reuse and Re-Sharing workflows

At this stage in the evolution of myExperiment content we observe that 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 [28, 29].

Although *anonymous reuse* (i.e. the author was not contacted by the user) is observed 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 items are all evidence. The myExperiment approach is to pay attention to "the social mechanisms which generate trust" [8] and provide a range of 'trust affordances' which users may turn to when trust becomes a practical issue.

4. OPEN SCIENTIFIC PLATFORMS

In section 1 we argued that by pooling and sharing methods we have the potential to accelerate science through exchanging know-how, and by combining methods with results we can accelerate discovery by enabling transparent, comparable and reproducible science. This has been illustrated by the sharing of workflows and the use of packs in myExperiment. Here we extend this to the more general notion of *Research Objects* (ROs) – the work objects that are built, transformed and published in the course of scientific experiments – and explain how we are supporting these in myExperiment through the use of Semantic Web technologies.

4.1 Research Objects

ROs are compound objects that group together resources used in an investigation, experiment, question or process – an aggregation of datasets, analysis methods, workflows, results, electronic records and the corresponding metadata in order to capture the narrative of the investigation. For example, all the items in Figure 1 constitute a RO when we add metadata (1) attached to individual items (e.g. "Common Pathways"), (2) describing the relationships between them (e.g. "produces", "published in") and (3) associated with the RO itself (e.g. tags). 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.

Based on our experience in myExperiment and analysis of requirements across several projects, we have identified five key characteristics of Research Objects:

1 **Composite**. They contain typed interrelationships and dependencies between resources and are in turn labeled and identifiable as an individual resource. These are depicted by the arrows in figure 1.

2 **Distributed**. They are structured collections of references to locally managed and externally located resources; for example, some myExperiment packs contain sample data and references to large datasets elsewhere. This 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 (co-referencing, co-searching etc). This is the 'social metadata' depicted in figure 2.

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) [30]. Figure 1 is an example of the range of resources needed to achieve this.

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.

Research Objects are machine-processable and support automation, increasing the robustness of the research. In [31] 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. 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 are not lost during that publication process.

4.2 Supporting Research Objects in myExperiment

myExperiment has explored the requirements of Research Objects through workflows and packs. 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 (www.sysmo.net) 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 party 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 externally 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 collection of example 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 provide a partial implementation of Research Objects, though without an interoperable representation outside myExperiment. 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.

In order to support a more general and open model for Research Objects we have created a live replica of the myExperiment content in our Research Objects data model, which has the benefit of (1) being independent of evolution of the data model in the codebase supporting the user interface, and (2) demonstrating how Research Objects can be supported as an adjunct to a separate application. In line with our open approach we deliver Research Objects as Web content that can be handled by standard tools. Our Research Objects are *descriptions* of the aggregations of resources and associated metadata rather than the actual data: from a description it is possible to capture the actual data in an appropriate archive format if this is required. Our descriptions are in the Resource Description Framework (RDF) [32], which provides a simple subject-predicate-object (triple) structure that can be processed by established tooling.

myExperiment publishes all its public data as RDF at http://rdf.myexperiment.org/. To make sense of this data, myExperiment also provides a meta-structure in the form of an OWL ontology to formalise relationships within this data. The myExperiment ontology (http://rdf.myexperiment.org/ontologies/) reuses properties from more generic ontologies/schemas, in particular: 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 and experiments. Through this reuse it is possible to make some sense of myExperiment data outside its domain, allowing data from different sources to be collated.

Instead of being written as a single monolith, the myExperiment ontology is a built as a set of modules that can be assembled to provide a comprehensive representation. There is an initial base module to define and reconcile basic terms for content management, object annotation and social networking. On top of this there are a number of modules that relate to specific aspects of the ontology, (types of contribution, types of annotation, credit and attribution, usage statistics, packs, experiments and workflow components). A final module performs the assembly using the OWL's import property and adds the most specific terms.

Modularising the myExperiment ontology makes it less restrictive and more suitable for reuse, allowing analogous projects (see section 5) to map their data in a very similar way. Significant effort is currently focused on how to represent 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 group (http://esw.w3.org/topic/HCLSIG/SWANSIOC) has been considering how to reconcile a number of ontologies, including the myExperiment ontology, that treat experiments as first class objects.

myExperiment's SPARQL endpoint (http://rdf.myexperiment.org/sparql) allows queries to be performed across all its RDF data. SPARQL's flexible nature (it essentially just maps networks where one or more of the nodes or links are unknown) allows anything from simple queries, comparable to REST API calls, to much more complex bespoke queries; e.g. myExperiment's RDF provides a listing of components (sources, sinks, processors and links) of Taverna 1 workflows. It is possible to construct a SPARQL query to represent the interlinking of these components in a specific user-defined way, allowing workflows to be found that are tailored to a particular person's requirements.

5. DISCUSSION AND FUTURE WORK

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. myExperiment is the first repository of methods which majors on the social dimension, and we have demonstrated that an online community and workflow collection has been established and is now growing around it. As such, we believe that myExperiment represents an important step towards the realisation of a radical new vision for the creation, sharing and publishing of 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 trajectories of innovations are notoriously hard to predict and to direct. In myExperiment the innovation is as much (if not more) social as it is technical, so the outcome will depend upon how the community responds to its potential. This is subject to being renegotiated by the community as it explores and discovers uses for myExperiment, some of which may not have been anticipated [29]. Our study of the myExperiment community shows that its members' aims and expectations are diverse and evolving. We expect that significant changes will occur as community members learn from one another – as participants in "the experiment that is myexperiment" – what can be achieved through using myExperiment and as examples of good practice propagate through the

community [33]. For the myExperiment team it follows that it is crucial that it has in place a design and development methodology that enables it to build on what is successful, diagnose and respond quickly to problems and thereby ensure that myExperiment continues to evolve along with its community [15].

As we have developed the functionality of myExperiment, and used myExperiment within a variety of other projects, we have begun to identify patterns of use and identify the reusable components and resources of myExperiment itself. A common pattern across projects is discovery and acquisition of digital resources, then conducting research in a private group, followed by disseminating results publicly; the resources then need curation. This reuse of services is a step towards our vision of the *e-laboratory* (or *e-lab*), an assembly of 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 investigation. ROs play a role in driving the components and the 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

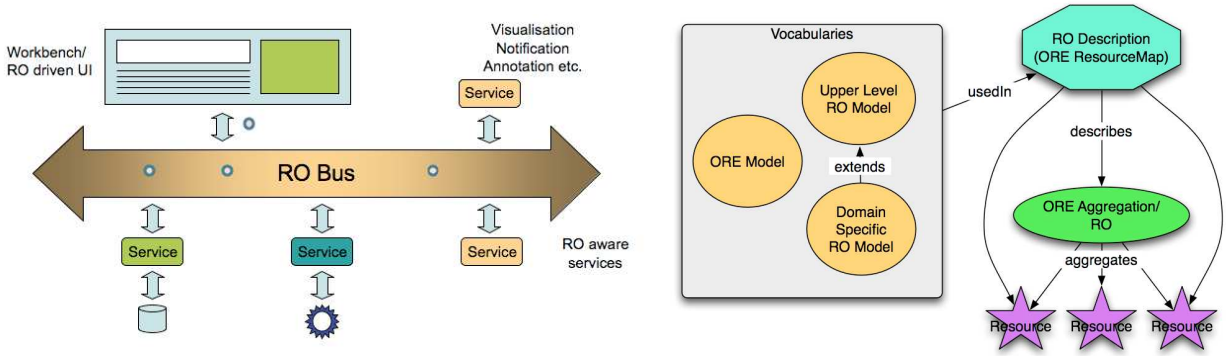

Figure 5. The Research Object bus underpins the e-Laboratory

will also play a role in sharing/communicating not just between services and components within an e-lab, but also with other e-labs. Figure 5 illustrates the "Research Object bus" which couples together e-Lab services, and the Research Object data model.

The core myExperiment service is evolving to greater repository integration and, with this, a federation model for content. In line with our Open Science approach we will publish our Research Objects in RDF and comply with Linked Data guidelines (http://linkeddata.org/). We are addressing the associated challenges in shared names (http://sharedname.org/) and co-reference resolution [34]. We are committed to the RESTful, resource-oriented world because it empowers our users; we also recognise the essential automation provided by the service-oriented world. These come together in the Biocatalogue project (www.biocatalogue.org) which provides a registry of web services in the life sciences and demonstrates a particular symbiosis with myExperiment, providing service information for workflow users, learning from service usage within the workflow collection, and borrowing directly from myExperiment's curation models. As the myExperiment collection grows we will work to facilitate discovery through recommendation and incentivise contribution and curation.

There are many exciting developments as myExperiment is applied in a range of disciplines, from Obesity e-Labs (https://www.nibhi.org.uk/obesityelab/) and shared genomics to music information retrieval (http://nema.lis.uiuc.edu/) and e-Books in social statistics. All these efforts will bring new capabilities to the researcher, and enable the ideas of myExperiment to evolve to underpin the e-Laboratory and our vision of research within and across disciplines which is more reusable and repeatable and hence more rapid and robust.

ACKNOWLEDGEMENTS

The design of myExperiment and Research Objects has been a collaborative exercise involving a large group of people including Mark Borkum, Iain Buchan, Les Carr, Simon Coles, Phil Couch, Catherine De Roure, Tom Eveleigh, Jeremy Frey, Antoon Goderis, Matt Lee, Cameron Neylon, Stuart Owen, Savas Parastatidis, Marco Roos, Robert Stevens, Shoaib Sufi, Franck Tanoh, David Withers and Katy Wolstencroft. Thanks also to our external developers, Tony Linde, Paul Groth, Hong Nguyen and Wim De Smet, and to colleagues in the W3C Semantic Web Health Care and Life Sciences (HCLS) Interest Group. The myExperiment project is funded by the UK Joint Information Systems Committee, the UK Engineering and Physical Sciences Research Council and the Microsoft Technical Computing Initiative.

REFERENCES

1. Dirks, L. & Hey, T. (2007) The Coming Revolution in Scholarly Communications & Cyberinfrastructure, *CTWatch Quarterly. 3*.
2. Seekda http://seekda.com/, visited 8 April 2009
3. Anderson, N., Tarczy-Hornoch, P. & Bumgarner, R. (2006) On the persistence of supplementary resources in biomedical publications, *BMC Bioinformatics. 7*, 260.
4. Pearson, H. (2006) Online methods share insider tricks, *Nature. 441*, 678-678.
5. Klimeck, G., McLennan, M., Brophy, S. P., Adams, G. B. & Lundstrom, M. S. (2008) nanoHUB.org: Advancing Education and Research in Nanotechnology, *Computing in Science & Engineering. 10*, 17-23.
6. Goble, C. & De Roure, D. (2008) Curating Scientific Web Services and Workflow, *EDUCAUSE Review. 43*.
7. Buckheit, J. B. & Donoho, D. L. (1995) WaveLab and Reproducible Research in *Tech. Rep.*, Dept. of Statistics, Stanford University,
8. Gil, Y., Deelman, E., Ellisman, M., Fahringer, T., Fox, G., Gannon, D., Goble, C., Livny, M., Moreau, L. & Myers, J. (2007) Examining the Challenges of Scientific Workflows, *Computer. 40*, 24-32.
9. Goderis, A., Sattler, U., Lord, P. & Goble, C. (2005) Seven Bottlenecks to Workflow Reuse and Repurposing in *The Semantic Web – ISWC 2005* pp. 323-337.
10. Fisher, P., Hedeler, C., Wolstencroft, K., Hulme, H., Noyes, H., Kemp, S., Stevens, R. & Brass, A. (2007) A systematic strategy for large-scale analysis of genotype phenotype correlations: identification of candidate genes involved in African trypanosomiasis, *Nucleic Acids Res. 35*, 5625--5633.
11. Neylon, C. Science in the Open http://blog.openwetware.org/scienceintheopen/, visited April 8 2009
12. Bradley, J.-C. UsefulChem http://usefulchem.blogspot.com/, visited April 8 2009
13. Tapscott, D. & Williams, A. D. (2008) *Wikinomics: How Mass Collaboration Changes Everything*, Penguin Group (USA) Incorporated.
14. Shneiderman, B. (2008) Science 2.0, *Science. 319*, 1349-1350.
15. De Roure, D. & Goble, C. (2009) Software Design for Empowering Scientists, *Software, IEEE. 26*, 88-95.
16. De Roure, D., Goble, C. & Stevens, R. (2009) The design and realisation of the Virtual Research Environment for social sharing of workflows, *Future Generation Computer Systems. 25*, 561-567.
17. Oinn, T., Greenwood, M., Addis, M., Alpdemir, N., Ferris, J., Glover, K., Goble, C., Goderis, A., Hull, D., Marvin, D., Li, P., Lord, P., Pocock, M., Senger, M., Stevens, R., Wipat, A. & Wroe, C. (2006) Taverna: lessons in creating a workflow environment for the life sciences: Research Articles, *Concurr. Comput. : Pract. Exper. 18*, 1067--1100.
18. McPhillips, T., Bowers, S., Zinn, D. & Ludäscher, B. (2009) Scientific workflow design for mere mortals, *Future Generation Computer Systems. 25*, 541-551.
19. Taylor, I., Shields, M., Wang, I. & Harrison, A. (2007) The Triana Workflow Environment: Architecture and Applications in *Workflows for e-Science* pp. 320-339.
20. Trident: Scientific Workflow Workbench for Oceanography http://www.microsoft.com/mscorp/tc/trident.mspx, visited April 8 2009
21. Lin, Y., Poschen, M., Procter, R., Voss, A., Goble, C., Bhagat, J., De Roure, D., Cruickshank, D. & Rouncefield, M. (2008) Lessons from Developing a Social Networking Site for Scientists in *e-Social Science '08* Manchester, UK.
22. De Roure, D., Goble, C., Bhagat, J., Cruickshank, D., Goderis, A., Michaelides, D. & Newman, D. (2008). myExperiment: Defining the Social Virtual Research Environment. Paper presented at the *IEEE Fourth International Conference on eScience (eScience '08)*.
23. Williams, A. J. (2008) Internet-based tools for communication and collaboration in chemistry, *Drug Discovery Today. 13*, 502-506.
24. Lin, Y., Poschen, M., Procter, R., Voss, A., Goble, C., Bhagat, J., De Roure, D., Cruickshank, D. & Rouncefield, M. (2008) Agile Management: Strategies for Developing a Social Networking Site for Scientists in *4th International Conference on e-Social Science* Manchester, UK.
25. Bradley, J.-C. (2007) Open Notebook Science Using Blogs and Wikis in *Nature Precedings*
26. Piwowar, H. A. & Chapman, W. W. (2008) A review of journal policies for sharing research data in *Nature Precedings*
27. Nielsen, J. (2006) Participation Inequality: Encouraging More Users to Contribute in
28. Groth, P. T. & Gil, Y. (2009) A scientific workflow construction command line in *Proceedings of the 13th international conference on Intelligent user interfaces*, ACM, Sanibel Island, Florida, USA.
29. Biton, O., Davidson, S. B., Khanna, S. & Roy, S. (2009) Optimizing user views for workflows in *Proceedings of the 12th International Conference on Database Theory*, ACM, St. Petersburg, Russia.
30. Goderis, A., Fisher, P., Gibson, A., Tanoh, F., Wolstencroft, K., De Roure, D. & Goble, C. (2009) Benchmarking Workflow Discovery: A Case Study From Bioinformatics, *Concurrency and Computation: Practice and Experience. Accepted for publication.*
31. Mons, B. (2005) Which gene did you mean? *BMC Bioinformatics. 6*, 142.
32. Miller, E. (1998) An Introduction to the Resource Description Framework, *D-Lib Magazine*.
33. Williams, R., Stewart, J. & Slack, R. (2005) *Social Learning In Technological Innovation: Experimenting With Information And Communication Technologies*, Edward Elgar Publishing, Incorporated.
34. Glaser, H., Millard, I., Jaffri, A., Lewy, T. & Dowling, B. (2008) On Coreference and The Semantic Web in *7th International Semantic Web Conference*, (submitted), Karlsruhe, Germany.