Designing the ^{my}Experiment Virtual Research Environment for the Social Sharing of Workflows

David De Roure
University of Southampton
dder@ecs.soton.ac.uk

Carole Goble
University of Manchester
carole.goble@manchester.ac.uk

Robert Stevens
University of Manchester
robert.stevens@manchester.ac.uk

Abstract

Many scientific workflow systems have been developed and are serving to benefit science. In this paper we look outside the workflow to consider the use of workflows within scientific practice, and we argue that the tremendous scientific potential of workflows will be achieved through mechanisms for sharing and collaboration – empowering the scientist to spread their experimental protocols and to benefit from the protocols of others. We discuss issues in workflow sharing, propose a set of design principles for collaborative e-Science software, and illustrate these principles in action through the design of the ^{my}Experiment Virtual Research Environment for collaboration and sharing of experiments.

1. Introduction

Scientific workflows are attracting considerable attention in the community, as demonstrated by workshops, conferences and journal special issues and books, e.g. [1]. Increasingly they support scientists in advancing research through *in silico* experimentation, while the systems themselves provide challenges for the community that designs and develops workflow software.

The National Science Foundation Workshop on the Challenges of Scientific Workflows [2] identified the potential for scientific advance as workflow systems address more sophisticated requirements and as workflows are created “through collaborative design processes involving many scientists across disciplines”. Understanding the whole lifecycle of the workflow – design, management, publication and discovery – is fundamental to developing systems that support the scientists’ work and not just the workflow’s execution. Supporting that lifecycle can be *the* factor that means a workflow approach is adopted or not. Workflow design is challenging and labour-intensive. Reusing a body of prior established designs through registries or catalogues is highly desirable [3-5].

Reuse is a particular challenge when scientists are outside a predefined Virtual Organisation or enterprise. These are individuals or small groups, decoupled and acting independently, who are seeking workflows that cover processes outside their expertise. This latter point arises when workflows are shared across discipline boundaries and when inexperienced scientists need to leverage the expertise of the great and the good.

Rather than looking at the machinery of workflow systems, it is the dimension of collaboration and sharing that is the focus of this paper, as we consider the lifecycle of workflows in the context of scientific research.

Our contribution is the identification of a set of design principles for the collaborative software that will enable scientists to exchange workflows. These principles have emerged from our experiences on a range of e-Science projects. We apply these principles in the design of the ^{my}Experiment Virtual Research Environment for collaboration and sharing of experiments [6], which aims to provide a “workflow bazaar” for any workflow management system, and for many other kinds of scientific assets.

In the next section we discuss the use of workflows for science, the power of workflows as first class citizens, their lifecycle and hence the requirements for sharing. This is followed in Section 3 by our design approach based on our experience in e-Science projects, which we then apply to the design of ^{my}Experiment, followed by an instructive review of this design against the O’Reilly Web 2.0 design patterns [7]. We close in Section 4 with discussion.

2. Scientific Workflows

There are many workflow systems available – we found over 75 after conducting an informal search. These systems vary in many respects: e.g. who uses them; what resources they operate over; whether the systems are open or closed; how workflows are expressed (e.g. how control flow is handled); how interactive they are; how tasks are allocated to resources and how exceptions are handled.

Our focus is on workflows near the application level rather than those further down in the infrastructure; i.e. we are interested in composing scientific applications and components using workflows, over a service oriented infrastructure (which may include Grid services). These are the workflows which are close to the scientist, or indeed the researcher in any domain. There is a distinction between workflow templates and workflow instances: the former describes the steps and order of the process without identifying particular end points of services (or codes), while the workflow instance binds in the concrete executions [3]. Both may have sample data or real data associated with them. Here we use "workflow" to mean both templates and instances unless otherwise stated.

2.1 Workflow Systems

The ^{my}Grid project (http://www.mygrid.org.uk) has developed the Taverna workflow workbench [8], used extensively across a range of Life Science problems: gene and protein annotation; proteomics, phylogeny and phenotypical studies; microarray data analysis and medical image analysis; high throughput screening of chemical compounds and clinical statistical analysis. Taverna is now part of the Open Middleware Infrastructure Institute UK (http://www.omii.ac.uk) portfolio of supported software development, so that scientists can rely upon it as part of their regular collection of tools.

Importantly, Taverna has been designed to operate in the "open wild world" of bioinformatics. Rather than large scale, closed collaborations which own resources, Taverna is used to enable individual scientists to access the many open resources available "in the cloud", i.e. out on the Web, not necessarily within their enterprise.

The services are expected to be owned by parties other than those using them in a workflow, so they are volatile, weakly described and there is no contract in place to ensure quality of service; they have not been designed to work together, and they adhere to no common type system. Consequently, they are highly heterogeneous. By compensating for these demands, Taverna has made, at the time of writing, over 3500 bioinformatics orientated operations available to its users. This has been a major incentive to adoption. This openness also means that Taverna is not tied exclusively to the bioinformatics domain—any services can be incorporated into its workflows.

To compare this with another point in the rich space of workflow systems, the Pegasus system [3,9] has more of a computational and Grid emphasis. Pegasus maps from workflow instances to executable workflows, automatically identifying physical locations for workflow components and data and finding appropriate resources to execute the components. It reuses existing data products where applicable. The lifecycle of Pegasus workflows is described in [3].

Pegasus is used within large scale collaborations and big projects. It is perhaps more typical of e-Science and grid activities, while Taverna gives an interesting insight into another part of the scientific workflow "ecosystem". We note that it is being used by many scientists on their personal projects – they constitute a distributed, disconnected community of users who are also the developers of the workflows[1]. While e-Science has often focused on specialist early-adopter scientists and large scale collaborative projects, Taverna is used by the "long tail" of researchers doing everyday science.

2.2. The workflow as a first class citizen

A feature of workflows leading to their uptake is the easing of the burden of repetitive manual work. However, we suggest that the key feature for scientific advancement is reuse. Workflow descriptions are not simply digital data objects like many other assets of e-Science, but rather they actually capture pieces of scientific process – they are valuable knowledge assets in their own right, capturing valuable know-how that is otherwise often tacit [4]. Reuse is effective at multiple levels:

- The scientist reuses a workflow with different parameters and data, and may modify the workflow, as part of the routine of their daily scientific work;
- Workflows can be shared with other scientists conducting similar work, so they provide a means of codifying, sharing and thus spreading the workflow designer's practice;
- Workflows, components of workflows and workflow patterns can be reused to support science outside their initial application.

The latter point illustrates the tremendous potential for new scientific advance. An example of this is a workflow used to help identify genes involved in tolerance to Trypanosomiasis in east African cattle [10]. The same workflow was reused without change over a new dataset to identify the biological pathways involved in sex dependence in a separate mouse model

[1] Taverna has ranked in the top 200 on Sourceforge and in July 2007 crossed 36,000 downloads.

for the whipworm parasite tolerance. This reuse was made easier by the explicit, high-level nature of the workflow that describes the analytical protocol.

Workflows bring challenges too. They can be difficult and expensive to develop – realistic workflows require skill to produce. Consequently, workflow developers need development assistance, and prefer not to start from scratch. Furthermore it is easy for the reuse of a workflow to be confined to the project in which it was conceived. In the Trypanosomiasis example, the barrier to this reuse was how the knowledge about the workflow could be spread to the scientists with the potential need. In this case it was word of mouth within one institution; this barrier needs to be overcome. So, we have a situation of workflows as reusable knowledge commodities, but with potential barriers to the exchange and propagation of those scientific ideas that are captured as workflow [11].

Significantly, there is more to a workflow than the declaration of a process. An individual workflow description may take the form of an XML file, but these do not sit in isolation. We identify a range of properties that are factors in guiding workflow reuse, including: descriptions of its *function and purpose*; documentation about the services with which it has been used, with example input and output data, and design explanations; *provenance*, including its version history and origins; *reputation and use* within the community; *ownership and permissions* constraints; *quality*, whether it is reviewed and still works; and *dependencies* on other workflows, components and data types. By binding workflows with this kind of information, we provide a basis for workflows to be trusted, interpreted unambiguously and reused accurately.

Workflows enable us to record the provenance of the data resulting from workflow enactment, and the log of the execution run. By binding outcomes with a package of their workflow instance and data, we provide a basis for the outcomes to be trusted, interpreted unambiguously and reused accurately. Like the workflows themselves, this provenance information is currently often confined to the system from which it originated and thus is not used as a useful commodity in its own right.

2.3. Thinking outside the workflow

It is apparent then that we can view workflows as potential commodities, as valuable first class assets in their own right, to be pooled and shared, traded and reused, within communities and across communities. Workflows themselves can be the subject of peer review. We can conceive of packs of workflows for certain topics, and of workflow pattern books – new structures above the level of the individual workflow. We call this perspective of the interacting data, services, workflow and their metadata within a scientific environment the *workflow ecosystem* and we believe that by understanding and enabling this we can unlock the broader scientific potential of workflow systems.

We start by looking at emerging practice. Taverna gives us a useful evidence base for such an investigation. In a small study (conducted in March 2007) we found around 400 Taverna workflows publicly available on the Web. In addition to these we are aware of workflows developed within specific projects and restricted to the project partners. Workflows are being placed on websites, and in particular they are being placed on Wikis as these are increasingly used by scientists to record and share their work. We have even seen a workflow made available through Flickr.

Thus we predict that workflows will become part of the scholarly knowledge cycle – the processes by which we publish scientific outcomes into the community and reuse these results in moving science forward. It will take time for workflows to be fully integrated in the cycle, as we are only now seeing scientific data being incorporated. However, we must plan to enable this activity. It will, for example, permit peer review of workflows. In fact we anticipate a more profound effect – that workflows themselves will feature in the process, for example in being used to reproduce results as part of the review process.

2.4 Social Sharing

If we accept that the key to the value of workflows is reuse, we need to understand how scientists will share and work collaboratively with them – the social and technical challenges.

Supporting the lifecycle of workflows is not about the workflows themselves but about the infrastructure for finding, using and sharing those workflows. This is about workflow metadata [4]. Consider how we might find a workflow: perhaps it is accurately catalogued and described so that we can identify it by its function. This kind of metadata is hard to both generate and maintain, as we know from first hand experience because we are obliged to provide this quality of data in service descriptions to support their use, both automatically and via human driven discovery mechanisms.

The key to ease workflow discovery lies in their use by a *community of scientists*. This acknowledges a central fact, sometimes neglected, that the lifecycle of the workflows is coupled with the process of science – that the human system of workflow use is coupled to the digital system of workflows. The more workflows, the more users and the more invocations then the more evidence there is to assist in selecting a workflow. The rise of harnessing the "Collective Intelligence" of the Web, the so-called "Socio-Web" and now the "Social Grid" [12], has dramatically reminded us that it is *people* who generate and share knowledge and resources, and people who create network effects in communities. Blogs and wikis, shared tagging services, instant messaging, social networks, semantic descriptions of data relationships, etc. are flourishing. Within the Scientific community we have examples: OpenWetWare, Connotea, PLoS on Facebook etc.

By mining the sharing behaviour between users within such a community we can provide recommendations of use. By using the structure and interactions between users and workflow tools we can identify what is considered to be of greater value to users. Provenance information helps track down workflows through their use in content syndication and aggregation.

2.5 Issues Summary

With workflow capture of an analytical protocol as a concrete object, we can see a nascent workflow ecosystem with its own environments, providers and consumers. If such an ecosystem with an overall goal of promoting workflow reuse is to be fully realised, several issues touched upon above need to be addressed in a design for software support:

Issue 1: Support for aspects of the scholarly cycle over and above *de novo* workflow construction.

Issue 2: *Attribution* of scholarly work—if scientists are to share intellectual property then the commodity needs to carry appropriate attribution. This is the means by which reputation is propagated through the community.

Issue 3: *Recommendation* of workflow, services, etc. is a vital part of enabling sharing through discovery by other scientists. It is also a part of communicating know-how.

Issue 4: The ability to *communicate* know-how about running or using an experiment; dissemination of best practice.

Issue 5: The ability to *review* and comment is an inherent part of recommendation and communication.

Issue 6: A scientist must be allowed entry at any point in the experimental or scholarly lifecycle.

Intuitively, such an idea is about scientists giving away their know-how. Why would a scientist release such valuable commodities to the wider community? Why would scientists share? However, this is the nature of the established scholarly knowledge cycle. The efficient unfolding of new knowledge in science rests on a set of idealised institutional norms, one of which is the sharing of knowledge among scientists [13]. The citing of published material is a form of reuse. Citing a scientist's paper is almost as valuable as the publication itself. By sharing or publishing a workflow, with the appropriate attribution, a scientist can allow their work to be reused with the concomitant spread of their scientific reputation—their workflow is, in effect, being cited.

3. The design of ^{my}Experiment

To explore these issues, and to support a growing user base of distributed and isolated workflow developers, we are developing collaborative software – a *Virtual Research Environment* – to support scientists using workflows and let them concentrate on being scientists and not programmers. We call this ^{my}Experiment [6]. Inspiration is drawn from the Web 2.0 community – Facebook, MySpace, Amazon, Digg etc – rather than the one of conventional scientific portals. ^{my}Experiment aims to make it easier for workflow workers to gossip about and exchange workflows, regardless of the workflow system – Taverna, Kepler, Triana, ActiveBPEL etc. We envisage: *a gossip shop* to share and discuss workflows and their related scientific objects; *a bazaar* for sharing, re-using and repurposing workflows; *a gateway* to other established environments, for example: depositing into data repositories and journals; and *a platform* to launch workflows, whatever their system. We hope that our scientists will use whatever workflow is appropriate for their applications– a kind of "workflow mashing".

3.1. Designing e-Science software

The design principles for ^{my}Experiment are based upon reflection and synthesis on our experiences in the first phases of the ^{my}Grid and CombeChem (http://www.combechem.org) e-Science pilot projects and their associated activities. These projects set out to empower the scientists and support them in developing new ways of working, and both have worked with 3rd party resources "in the cloud". myGrid focused on workflows, whereas CombeChem focused on publishing.

3.1.1 myGrid. Adopting a "come as you are" approach, myGrid makes no demands on 3rd party services to adapt to Taverna, which itself is delivered as a client-side application. Support within the enterprise was not needed for the desktop application, nor the remote services. This made it easy to get started, lowering the barrier of engagement and tapping into one of the reward incentives of the stakeholders, which we characterise as "Jam Today and more Jam Tomorrow".

Taverna is designed to enable users to add value through building and sharing workflows or developing plug-ins for the Taverna system. By solving a specific problem we again get early rewards for users and greater buy-in. This contrasts with building a generic solution in the abstract. Customisation also gives a sense of "specialness" and relevance to the scientist. "Act local, think Global" works because specific solutions developed for a representative user usually have broader applicability.

3.1.2 CombeChem. Like many e-Science projects, CombeChem addresses the data deluge brought about by new experimental techniques, in this case combinatorial chemistry. In contrast to others, it has taken a holistic view of the scholarly knowledge lifecycle, and its simply-stated objective is to provide a complete chain of knowledge from the laboratory bench through to scholarly publication.

CombeChem can be viewed as a Semantic DataGrid [14] in that it links up decoupled data "in the cloud", and it is just as much about publishing data and metadata as it is about using them. For example, CombeChem data outputs and provenance information are easily available through Web pages, and rather than hoarding metadata in a central store, CombeChem makes it available across multiple simple web server interfaces. The Open Archives Initiative (OAI) protocols (http://www.openarchives.org) are used to support federated stores, enabling CombeChem to integrate with distributed repositories of data and publications.

3.2. Design Principles

Based on these experiences we propose six principles for designing e-Science software that empowers the scientist and maximises reuse:

1. **Supporting everyday science**. Aim at the larger number of scientists conducting research on an everyday basis, and benefit from the network effects.
2. **Come as you are**. The user should not have to make any changes in order to use the system – its functionality should sit comfortably with their existing environment and practice. Offer a low activation energy: bring the functionality to the user rather than the user to functionality.
3. **Jam today, more jam tomorrow**. Understand the incentive models. There should be immediate benefit, with the potential for greater benefit in time. Remember that scientists wish to do science, not IT.
4. **Cooperate and get users to add value**. Enable users and developers to enhance the system through creating content, integrating services and developing software.
5. **Think about publishing data as much as using it**. The data is where the value lies. All data and metadata to be made available easily, with minimum restrictions and maximum ease of re-use.
6. **Metadata matters**. Metadata is data too and has its own lifecycle and need for infrastructural machinery. This is often neglected, because the rewards are for others.

We have also compared our approach with the established literature in group working and Computer Supported Collaborative Work – this combination of disciplines was addressed in a recent workshop (see reference [15]). We have found some principles to be consistent, but in many ways we are in a very different place to earlier work since a distributed application platform (the Web) and collaborative tools (Wikis) are already widely deployed.

3.3. The myExperiment design space

We held three workshops leading to the initial design of myExperiment: a "portal party" (http://www.mygrid.org.uk/wiki/Portal) with end-users to establish requirements, followed by two design scoping workshops coupled with presentations from specific end-user groups – we are starting with the life sciences and then extending to chemistry, astronomy and social sciences.

The portal party identified 26 requirements grouped under workflow repository, workflow run time environment, community development and cross-cutting requirements (confidentiality etc). Workflow enactment is particularly significant because it distinguishes ^{my}Experiment from existing repository and social networking solutions and captures the distinctive feature of our work, i.e. *in silico* science.

The scoping workshops took the issues presented in Section 2 and the design principles presented above, and explored the following design dimensions:

1. *Federation*. A centralised “workflow warehouse” versus a federation of workflow repositories.
2. *Interface*. Is it a custom application, a website, or is it integrated with existing tools on the desktop, in the enterprise or on the Web (e.g. Wikis, Google gadgets).
3. *Granularity*. Do we work with individual workflows or “experiments”, how do we group constituent objects together, and how do we identify them?
4. *Guardianship*. Open versus managed content – scientific data has issues of quality, reliability, validation, safety, intellectual property, ownership and confidentiality.
5. *Execution*. Where are workflows enacted – on the client, on remote servers or in the enterprise?
6. *Social networks*. How are social networks defined, how are tags created and shared – is visibility confined to groups and what is the balance with the network effects?
7. *Control*. An organically growing social space (like Wikis, such as OpenWetWare) versus a highly organised workflow “shop”.
8. *Software*. Open source, open standards, open interfaces. How is community development supported – plugins versus mashups.
9. *Automation*. Workflow creation wizards, automatic tagging, autonomic curation.

The outcomes of our discussion on key dimensions are presented in [15] and the resulting architecture is illustrated in Figure 1. Below we summarise the first three dimensions to illustrate the application of the design principles.

3.3.1. Federation – Warehouse or Distributed Repositories? In one model, ^{my}Experiment could be a Web site with its own workflow repository, either constructed as a completely new site or by tailoring existing solutions such as Media Wiki. Alternatively, the various objects (workflows, data, provenance records) could be maintained in a number of distributed repositories. A ^{my}Experiment Web site is then just one of many possible interfaces to this content.

We have chosen to build a Web site which can store workflows, thus providing a standalone solution, and which can also participate in a federated repository model (“come as you are”). This is achieved through metadata harvesting and repository interoperability

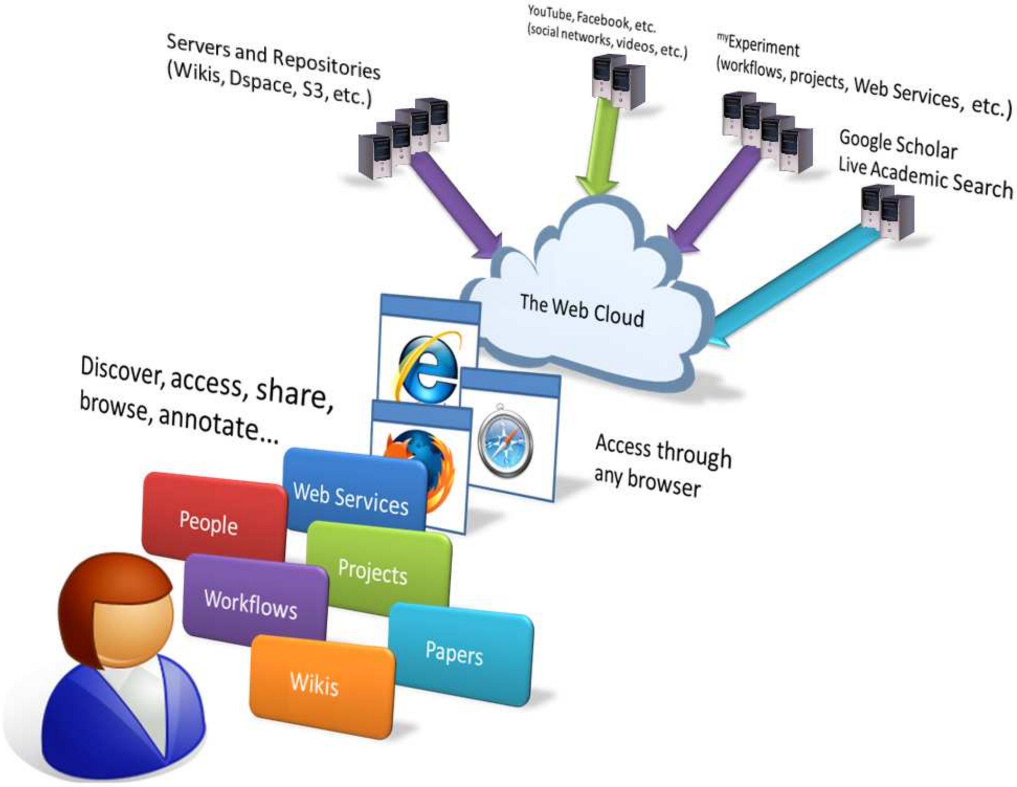

Figure 1: ^{my}Experiment “exploded”

protocols such as OAI, and builds on the experience of the publishing ethos of CombeChem and using OAI with scientific data in eBank-UK [16] (metadata and publishing).

By providing a public site we support everyday science and provide community focus, and others can use our services ("cooperate and let users add value"). Scientists gain immediate benefit from available workflows ("jam today"). By using our own site rather than augmenting sites such as Facebook we can address the granularity and guardianship issues for scientific data. Individuals and labs are free to install their own ^{my}Experiment instances and link them up into the federation model as they wish.

3.3.2. Interface – Bringing ^{my}Experiment to the user. As well as "exploding" the back end from one site to many, we explode the interface so that ^{my}Experiment functionality can be brought to the user through existing interfaces (again, "come as you are"). For example, people who are already using Wikis can access ^{my}Experiment functions through plugins – a workflow on a Wiki page can be executed and new pages generated to record this. In a more extreme example (which we call the "bioinformatics sweatshop") the interface might not make the workflows visible at all. We also envisage ^{my}Experiment add-ons for sites such as Facebook.

3.3.3. Granularity – Towards experiment objects. We have focused on workflows, partly because we have an established user community with an immediate need, but as discussed in section 2 workflows alone are insufficient, and the ^{my}Experiment concept is also about sharing other digital objects – in general, to share experiments, which includes data, results, provenance information, tags, associated documentation etc.

To address this we are designing a simple way of composing dispersed items into a *Encapsulated ^{my}Experiment Object* (EMO) (the notion of encapsulation captures versioning) using a model which is consistent with scientific practice and with the linked data model of the Web. The ^{my}Experiment environment then becomes a way of working with EMOs and of providing experiment object services for others to use. This is an example of "metadata matters".

3.4. Web 2.0 Design Patterns

Having proposed and exercised our design principles, it is instructive to compare ^{my}Experiment with current social Web site practice by reviewing it against the Web 2.0 design patterns [7]. We review these briefly below.

The Long Tail – Our target users are not just the specialist e-Scientists using computing resources to tackle major scientific breakthroughs, but also the large number of scientists conducting the routine processes of science on a daily basis. Through sharing we have the potential to enable smart scientists to be smarter and propagate their smartness, in turn enabling other scientists to become better and conduct better science.

Data is the Next "Intel Inside" – ^{my}Experiment understands that scientists are focused on data, and that scientific workflows are components of customised data-oriented applications. Furthermore, workflows themselves are the data of ^{my}Experiment and provide its unique value.

Users Add Value – ^{my}Experiment makes it easy to find workflows and is designed to make it useful and straightforward to share workflows and add workflows to the pool.

Network Effects by Default – ^{my}Experiment aggregates data as a side-effect of using the VRE, for example the numbers of times workflows and services are used.

Some Rights Reserved – ^{my}Experiment users require protection as well as sharing, but the environment is designed for maximum ease of sharing to achieve collective benefits – workflows are "hackable" and "remixable". Initiatives such as Science Commons provide a useful context for this.

The Perpetual Beta – ^{my}Experiment is an online service – indeed a collection of online services – and is continually evolving in response to its users. To support this, the project commenced with developers being embedded in the user community. Through day-to-day contact between designers and researchers, design is both inspired and validated.

Cooperate, Don't Control – ^{my}Experiment is a network of cooperating data services with simple interfaces which make it easy to work with content. It both provides services and reuses the service of others. It aims to support lightweight programming models so that it can easily be part of loosely coupled systems.

Software Above the Level of a Single Device – The current model of Taverna running on the scientist's desktop PC or laptop is evolving into ^{my}Experiment being available through a variety of interfaces and supporting workflow execution.

4. Discussion

We have made the case for a mechanism for sharing workflows in order to realise their scientific potential, and have identified the issues in achieving this. Enabling incentive models for sharing within a "community of practice" and supporting an emergent model of sharing, is a challenge. The Virtual Organisations of Grid computing often attempt to achieve a similar objective, although they are typically centred on a common technically defined problem and do not focus on social aspects that might involve different incentive structures. To rise to this challenge we have proposed design principles for collaborative e-Science software and demonstrated their application in ^{my}Experiment through design workshops.

We have also shown that our resulting design is consistent with the Web 2.0 design patterns. We note that the collective benefits of participation which characterise Web 2.0 arise not only from the users but also from the developers – ease of use and ease of development. Fundamentally it is the simplicity of Web 2.0 which is attractive. Not only is ^{my}Experiment something that can be built using the Web 2.0 approach but it can be used this way too, and it sits comfortably in a Web 2.0 context for reuse. e-Science is difficult – workflows and Web 2.0 make it easier.

Development of the ^{my}Experiment web site commenced in March 2007, and user trials of the beta service have been conducted with bioinformatics users since July; the next user group will be in chemistry. The site will evolve to support other types of object and other workflow systems. We will report on the evaluation in a future paper.

^{my}Experiment is one case study in one set of communities. However we believe that many of the principles we have discussed in this paper are relevant to anyone developing software at the interface the infrastructure and the users – other Virtual Research Environments. We hope our design principles and the ^{my}Experiment design exercise process will help others who are developing collaborative software.

Acknowledgements

Thanks to the ^{my}Experiment team, all who have worked on ^{my}Grid and Taverna, and all our users; in particular we thank and Andy Brass for his support. Figure 1 is due to Savas Parastatidis. This work is funded by JISC VRE2 (04/06) with additional support through EPSRC grants EP/D044324/1, EP/C536444/1 and EP/C008863/1.

References

[1] Taylor I, Deelman E, Gannon D, and Shield M (Eds.) Workflows for e-Science, Springer 2006

[2] Deelman, E. and Gil, Y. (eds) NSF Workshop on the Workshop on the Challenges of Scientific Workflows, May 2006. See http://www.isi.edu/nsf-workflows06

[3] Deelman E, Gil Y. (2006)) Managing Large-Scale Scientific Workflows in Distributed Environments: Experiences and Challenges, Workflows in e-Science, e-Science 2006, Amsterdam, Dec 2006

[4] Wroe C, Goble CA, Goderis A et al *Recycling workflows and services through discovery and reuse* Concurrency and Computation: Practice and Experience 19(2) pp: 181-194, February 2007

[5] The Kepler Repository http://kepler-project.org/ (2007)

[6] Goble, C.A. and De Roure, D. myExperiment: social networking for workflow-using e-scientists, WORKS '07: Proceedings of the 2nd workshop on Workflows in support of large-scale science (2007) pp:1-2

[7] O'Reilly, T. What Is Web 2.0 – Design Patterns and Business Models for the Next Generation of Software. http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

[8] Oinn, T et al. (2006). Taverna: lessons in creating a workflow environment for the life sciences. Concurrency and Computation: Practice & Experience 18, 1067-1100.

[9] Deelman, E., Singh, G. et al. Pegasus: a Framework for Mapping Complex Scientific Workflows onto Distributed Systems, Scientific Programming Journal, Vol 13(3), 2005, Pages 219-237

[10] Fisher, P. et al. Nucleic Acids Research, accepted (2007)

[11] Goderis, A., Sattler, U., Lord, P. and Goble, C.A.. Seven bottlenecks to workflow reuse and repurposing. In Fourth International Semantic Web Conference (ISWC 2005), volume 3792, pages 323-337, Galway, Ireland, 2005.

[12] Hey, A.J.G. The Social Grid, Presentation at Open Grid forum 20, Manchester UK, May 2007.

[13] Ziman, J.M. (1968) Public knowledge: An essay concerning the social dimensions of science. (London: Cambridge University Press).

[14] Taylor, K., Gledhill, R., Essex, J. W., Frey, J. G., Harris, S. W. and De Roure, D. (2005) A Semantic Datagrid for Combinatorial Chemistry. In Proceedings of The 6th IEEE/ACM International Workshop on Grid Computing, pp. 148-155, Seattle.

[15] De Roure, D. and Goble, C.A. myExperiment - A Web 2.0 Virtual Research Environment. In International Workshop on Virtual Research Environments and Collaborative Work Environments, NeSC, Edinburgh, UK. May 2007.

[16] Duke, M. Day, M., Heery, R., Carr, L.A. and Coles, S.J.. Enhancing access to research data: the challenge of crystallography, JCDL 2005 Digital Libraries: Cyberinfrastructure for Research and Education, Denver, Colorado, USA June 7-11, 2005.