

myExperiment: a repository and social network for the sharing of bioinformatics workflows

Carole A. Goble¹, Jiten Bhagat¹, Sergejs Aleksejevs¹, Don Cruickshank², Danius Michaelides², David Newman², Mark Borkum², Sean Bechhofer¹, Marco Roos^{3,4}, Peter Li^{1*} & David De Roure²

¹*School of Computer Science, The University of Manchester, Manchester M13 9PL, UK.*

²*School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK.*

³*BioSemantics group, Human Genetics Department, Leiden University Medical Centre, Albinusdreef 2, 2333 ZA Leiden, The Netherlands.*

⁴*Adaptive Information Disclosure group, Informatics Institute, University of Amsterdam, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands.*

*To whom correspondence should be addressed.

ABSTRACT

myExperiment (<http://www.myexperiment.org>) is an online research environment that supports the social sharing of bioinformatics workflows. These workflows are procedures consisting of a series of computational tasks using web services performed on data from its retrieval, integration and analysis, to the visualisation of the results. As a public repository of workflows, myExperiment allows anybody to discover those that are relevant to their research which can then be reused and repurposed to their specific requirements. Conversely, developers can submit their workflows to myExperiment and enable them to be shared in a secure manner. Since its release in 2007, myExperiment currently has over 3500 registered users and contains more than 900 workflows. The social aspect to the sharing of these workflows is facilitated by registered users forming virtual communities bound together by a common interest or research project. Contributors of workflows can build their reputation within these communities by receiving feedback and credit from individuals who reuse their work. Further documentation about myExperiment including its REST web service is available from <http://wiki.myexperiment.org>. Feedback and requests for support can be sent to bugs@myexperiment.org.

INTRODUCTION

The deployment of data and tools as web services has gained increasing popularity over recent years (1). Major data providers such as the European Bioinformatics Institute (2), National Center for Biological Information (<http://eutils.ncbi.nlm.nih.gov>) and the DNA Database of Japan (3) as well as specialist research groups (4-7) have utilized the standardized set of web services protocols to provide much needed programmatic access to their computational resources. This has enabled information to be served directly to applications for performing common bioinformatics analyses such as the derivation of summaries, testing of hypotheses and the search for patterns in data.

Such processes can be represented as workflows that define the sequential flow of data through bioinformatics databases and analytical tools involved in a pipeline. A wide variety of workflow languages have been created to describe data processing pipelines enabling them to be stored for repeated use (8). These processes can be

constructed using workflow management software such as Taverna (9,10), Pipeline Pilot (11) and Kepler (12) to combine web services with local tools in a graphical fashion for querying, integrating and analysing data. A plethora of workflows have now been written by developers to form a critical mass of knowledge. For example, bioinformatics workflows have been designed for the analysis of microarray data (10), integration of gene expression levels with systems biology models (13), the extraction and structuring of knowledge from text (14) and the identification of genes associated with diseases (15). This has led to a need for workflows to be shared with colleagues and other interested parties. Sharing supports the reuse and repurposing of workflows for other applications and facilitates the building of workflows. The use of workflows also encourages reproducible research which is an issue that is becoming important across different scientific domains (16).

Motivated by the needs of scientists and inspired by social network websites such as Facebook (<http://www.facebook.com>) and MySpace (<http://www.myspace.com>), the myExperiment (<http://www.myexperiment.org>) project has developed an open source Web 2.0 infrastructure that enables scientific artifacts including workflows to be shared within the life sciences community (17). This infrastructure is comprised of a repository of workflows supported by a social networking environment that facilitates the sharing of workflows. In this paper, we show how users can interact with myExperiment to discover workflows for reuse in their research, and also to submit workflows for sharing in manner controlled by the uploader. In addition, we will address the relationship between myExperiment, workflow management systems and web service registries, such as BioCatalogue (18). These components perform crucial functions at different stages in the life cycle of workflows that could assist in the reproducibility of data-driven experiments when those experiments are reported in the scientific literature (19).

USERS AND COMMUNITIES IN MYEXPERIMENT

Users are the heart of myExperiment. They may be developers interested in contributing their workflows into the repository for subsequent sharing with the scientific community. Users may also be scientists wishing to discover workflows to be reused in their own research. myExperiment has built a social community around its repository of workflows which facilitates the sharing of workflows between

developers and interested parties. In this respect, myExperiment also acts as a training resource providing exemplar workflows and access to expertise to guide users in the orchestration of web services.

The creation of a social community has necessitated the registration of users in order for individuals to be identifiable in myExperiment. Whilst public content can be freely browsed and downloaded by anonymous users, a richer user experience is available upon registration. This is a simple process in myExperiment, requiring the entry of a username and password, or an openID URL (<http://openid.net>). An email address is also requested to confirm the registration of a user. Registration leads to a user profile which can be edited with further information about contact details and research interests. Furthermore, each profile provides listings of friends, workflows and other digital objects in myExperiment belonging to or valued by the user.

Users in myExperiment can benefit from two mechanisms to form communities. Firstly, a user may request friendship from other registered people. Friendship links lead to the building of a network of trusted individuals. Users can then opt to restrict the sharing of particular workflows and other associated documents to this trusted network. If data security is of the utmost importance for an organization then system administrators can deploy their own instance of myExperiment within an intranet. Secondly, communities in myExperiment can be formed by the creation of groups. A registered user can set up a group for which they become the administrator and they can invite other users to join. Other users can also request to join the group. This type of community is designed to allow people who, for example, want to work on the same project, are at the same institution or have the same research interests to share and manage data in their collaboration.

WORKFLOW DISCOVERY

New users of workflows will come to myExperiment wanting to query its repository to look for pre-existing workflows which they can make use of in their own research and data analyses. The workflows home page provides the starting entry point for the discovery of workflows (Fig. 1). myExperiment is open to any workflow system, supporting the sharing of workflows written in a range of workflow languages.

Whilst, at present, the majority are Taverna workflows, myExperiment contains workflows written in 24 other languages (Fig. 1). The workflows web page also categorizes the latest, last updated, most viewed, most downloaded and most favoured workflows.

The discovery of workflows in myExperiment can be performed in two ways. Firstly, a set of workflows can be selected for browsing based on popular tags that have been used to describe them (Fig. 1). Secondly, workflows can be discovered using a keyword search (Fig. 1). For example, a search using ‘BLAST’ (20) currently leads to 102 workflows being found by the keyword. Each workflow in myExperiment has a dedicated web page showing descriptive information about its inputs, outputs and the operations it makes on data, as well as a graphical representation of the workflow where possible (Fig. 2). Feedback can be provided on a workflow. This can be in the form of a rating, review and comment, or simply the marking of a workflow as a favourite by a user. Feedback helps a contributor to build up their reputation. In order for feedback to be given, a closer review of a workflow is required and this depends on it being downloadable for inspection and execution so that a user can understand the operations it makes on data. All workflows in myExperiment have a hyperlink which can be used to download it (Fig. 2) and then opened within its native workflow system for further editing or enactment.

In order to bring myExperiment to potential users, external applications can access its content by making use of its programmatic interface. Designed for ease of reuse and for community development, applications such as wikis and blogs or even mashups can access content in myExperiment or be augmented with its functionality through a RESTful application programming interface. This has, for example, been used by the myExperiment plugin for Taverna which allows its workbench to access and download workflows directly from myExperiment (21).

SHARING WORKFLOWS

Authors of workflows share the efforts of their work through the social infrastructure provided by myExperiment around its workflow repository. Whilst workflows have mainly been contributed for philanthropic purposes, they have also been submitted as

part of the publication process of papers (10,14,15) or for demonstrating how web services developed by organizations can be used in bioinformatics applications (2). The process of making a workflow shareable begins by selecting workflow on the New/Upload panel (Fig. 1). This leads to a workflow submission web page which requests selection of the workflow file to be uploaded as well as a title and a description. The submission of a workflow also provides the opportunity to tag it with keywords to support discovery. If the workflow is based on previous work present in myExperiment then this can also be acknowledged by crediting the relevant person or group in myExperiment. It is also possible to credit other users if the uploaded workflow was a collaborative effort. Furthermore, other workflows or digital documents in myExperiment can be attributed if they were reused in the creation of the uploaded workflow.

The privacy of data in the scientific community can be a serious concern (22). myExperiment provides a flexible authorisation model and allows any uploaded content to be made available with varying levels of sharing permissions. The types of people who can view, download and update a given workflow can be configured based on their relationship with the uploader. For example, maximum security can be placed on a workflow by setting it as private so that it is only accessible by the uploader, whereas the most open option is to allow a workflow to be viewed and downloadable by anyone. The rights by which a user can use a downloaded workflow are governed by the licensing assigned to it in myExperiment of which there are several choices such as Creative Commons and a GNU General Public Licence.

Since workflows may be the subject of a scientific paper (10) or may have been used in the analysis of published data (15), it is possible to associate workflows with citations. A complementary approach to tying workflows with publications and other types of digital documents is to use a myExperiment pack (Fig. 3). Packs are collections of items such as example enactment input data, Powerpoint slides and PDF files of scientific papers that have been uploaded into myExperiment as well as URL links to data on the Web which can be bound with a workflow (Fig. 3). Since packs can be the subject of sharing, tagging and discovery, they extend the application of myExperiment beyond workflows to any type of digital object associated with a scientific experiment involving computation.

THE WORKFLOW LIFE CYCLE AND REPRODUCIBLE SCIENCE

The life of a workflow extends beyond its initial construction and execution followed by its deposition in a repository. Its reuse also involves the discovery of existing and relevant designs, editing the workflow to repurpose it by the addition or removal of services, trying out the workflow, and then re-registration of the workflow as a new version (23) (Fig. 4). A workflow repository and a workflow construction environment such as myExperiment and Taverna, respectively, represent two components in the workflow life cycle. Existing work can be searched by making use of myExperiment which can then be downloaded and edited in its native workflow system. If the repurposing of downloaded workflows requires the addition of other web services then their discovery can be aided by using service directories such as BioCatalogue (18) and the EMBRACE registry (24) (Fig. 4). Updated workflows can be deposited into myExperiment with a link back to the original so that the evolution of workflows can be traced. This does not have to be performed by the uploader of the original workflow since other members on myExperiment can contribute new versions depending on the access permissions of the initial workflow they have reused.

A workflow repository and construction tool provide two components targeted towards improving the reproducibility of data-driven research involving a combination of software packages that is now conducted in contemporary science (19). Such analyses are often repeated several times with modification of the parameters until the final results are produced. Whilst these results are reported in scientific papers, the actual process of computation is often neglected and makes replication of the computational analysis by an independent scientist difficult if not impossible. Mesirov, (2010) proposes the use of a Reproducible Research System (RRS) to enable reproducible science. This RRS is comprised of a Reproducible Research Environment (RRE) to perform the computational analysis and a Reproducible Research Publisher (RRP) that is responsible for the preparation of a document describing the results of the computation.

The infrastructure provided by myExperiment and Taverna, together with the BioCatalogue registry of web services, can offer some of the functionality required of

a RRS to replicate analyses of data. The analysis of data is described in a step-by-step manner as a workflow that can be constructed and enacted using Taverna, which is also responsible for recording the execution provenance in a separate repository (25). The published workflow can be deposited in myExperiment and the web services it uses are described in BioCatalogue. Whilst a document preparation system to complete the proposal by Mesirov, (2010) is not yet offered directly, this type of component could be provided in the future, perhaps by making use of myExperiment packs for packaging workflows with provenance, input data and final results for redistribution with published papers.

FUTURE WORK

Packs in myExperiment can, at least to some extent, encapsulate digital objects such as input data, final results and provenance that are associated with workflows for distribution with published papers. However, these collections of data do not contain enough information about each object nor their relationships with one another to adequately describe an *in silico* experiment on data and make it reproducible. Future work in myExperiment will evolve workflow packs into linked research objects whose properties are self-describing (26). To this end, a prototype service is currently being developed to deliver myExperiment content in RDF format based on a modularized ontology drawing on concepts from the Dublin Core, FOAF and OAI Object Reuse and Exchange vocabularies. RDF content in myExperiment is queryable from a SPARQL endpoint available at <http://rdf.myexperiment.org/sparql>.

CONCLUSIONS

Since its introduction in November 2007, myExperiment has 3572 registered users and contains 919 workflows. In this period, 27,840 visits have been made by returning visitors coming from 168 countries. By showing how potential users can interact with myExperiment, we hope this will provoke further interest from bioinformaticians for myExperiment in the hope that they will share their knowledge with the wider community or be able to support them in the reuse of workflows required for their research. The myExperiment software can be downloaded as open

source at <http://rubyforge.org/projects/myexperiment>. Extensive documentation and help pages is available at <http://wiki.myexperiment.org>, and requests for support can be sent to bugs@myexperiment.org.

ACKNOWLEDGEMENTS

This work was supported by the Engineering and Physical Sciences Research Council; Joint Information Systems Committee and the Microsoft Technical Computing Initiative. The authors would also like to thank Andrea Wiggins for her help in the development of myExperiment.

REFERENCES

1. Stockinger, H., Attwood, T., Chohan, S., Côté, R., Cudré-Mauroux, P., Falquet, L., Fernandes, P., Finn, R., Hupponen, T., Korpelainen, E. *et al.* (2008) Experience using web services for biological sequence analysis. *Briefings in bioinformatics*, **9**, 493-505.
2. McWilliam, H., Valentin, F., Goujon, M., Li, W., Narayanasamy, M., Martin, J., Miyar, T. and Lopez, R. (2009) Web services at the European Bioinformatics Institute-2009. *Nucleic acids research*, **37**, W6-W10.
3. Kwon, Y., Shigemoto, Y., Kuwana, Y. and Sugawara, H. (2009) Web API for biology with a workflow navigation system. *Nucleic acids research*, **37**, W11-W16.
4. Wang, J. and Mu, Q. (2003) Soap-HT-BLAST: high throughput BLAST based on Web services. *Bioinformatics*, **19**, 1863-1864.
5. Jacobsen, A., Krogh, A., Kauppinen, S. and Lindow, M. (2010) miRMaid: a unified programming interface for microRNA data resources. *BMC Bioinformatics*, **11**, 29.
6. Wittig, U., Golebiewski, M., Kania, R., Krebs, O., Mir, S., Weidemann, A., Anstein, S., Saric, J. and Rojas, I. (2006), SABIO-RK: Integration and Curation of Reaction Kinetics Data. *Lecture Notes in Bioinformatics*, Vol. 4075, pp. 94-103.

7. Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K., Itoh, M., Kawashima, S., Katayama, T., Araki, M. and Hirakawa, M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic acids research*, **34**, D354-D357.
8. van der Aalst, W. (2003) Don't go with the Flow: Web Services Composition Standards Exposed. *IEEE Intelligent Systems*, **Jan/Feb**, 72-76.
9. Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M.R., Li, P. and Oinn, T. (2006) Taverna: a tool for building and running workflows of services. *Nucleic acids research*, **34**, W729-W732.
10. Li, P., Castrillo, J., Velarde, G., Wassink, I., Reyes, S., Owen, S., Withers, D., Oinn, T., Pocock, M., Goble, C. *et al.* (2008) Performing statistical analyses on quantitative data in Taverna workflows: an example using R and maxdBrowse to identify differentially-expressed genes from microarray data. *BMC Bioinformatics*, **9**, 334.
11. Kappler, M. (2008) Software for rapid prototyping in the pharmaceutical and bioechnology industries. *Current Opinion in Drug Discovery and Development*, **11**, 389-392.
12. Altintas, I., Berkley, C., Jaeger, E., Jones, M., Ludascher, B. and Mock, S. (Year) Kepler: an extensible system for design and execution of scientific workflows. 16th International Conference on Scientific and Statistical Database Management,
13. Li, P., Oinn, T., Soiland, S. and Kell, D. (2008) Automated manipulation of systems biology models using libSBML within Taverna workflows. *Bioinformatics*, **24**, 287-289.
14. Roos, M., Marshall, M., Gibson, A., Schuemie, M., Meij, E., Katrenko, S., van Hage, W., Krommydas, K. and Adriaans, P. (2009) Structuring and extracting knowledge for the support of hypothesis generation in molecular biology. *BMC Bioinformatics*, **10 (Suppl 10)**, S9.
15. Fisher, P., Hedeler, C., Wolstencroft, C., Hulme, H., Noyes, H., Kemp, S., Stevens, R. and Brass, A. (2007) A systematic strategy for large-scale analysis of genotype–phenotype correlations: identification of candidate genes involved in African trypanosomiasis. *Nucleic acids research*, **35**, 5625-5633.
16. Editorial. (2010) Supporting data. *Nature Medicine*, **16**, 131.

17. De Roure, D., Goble, C. and Stevens, R. (2009) The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows. *Future Generation Computer Systems*, **25**, 561-567.
18. Goble, C., Belhajjame, K., Tanoh, F., Bhagat, J., Wolstencroft, K., Stevens, R., Nzuobontane, E., McWilliam, H., Laurent, T. and Lopez, R. (2008) Biocatalogue: A Curated Web Service Registry for the Life Science Community. 2008 Microsoft eScience Workshop, Indianapolis, IN.
19. Mesirov, J. (2010) Accessible Reproducible Research. *Science*, **327**, 415-416.
20. Altschul, S., Gish, W., Miller, W., Myers, E. and Lipman, D. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403-410.
21. De Roure, D., Goble, C., Aleksejevs, S., Bechhofer, S., Bhagat, J., Cruickshank, D., Fisher, P., Hull, D., Michaelides, D., Newman, D. *et al.* (2009) Towards Open Science: The myExperiment approach. *Concurrency and Computation: Practice and Experience*, **Submitted**.
22. Waldrop, M. (2008) Big data: Wikiomics. *Nature*, **455**, 22-25.
23. Wroe, C., Goble, C., Goderis, A., Lord, P., Miles, S., Papay, J., Alper, P. and Moreau, L. (2007) Recycling workflows and services through discovery and reuse. *Concurrency and Computation: Practice and Experience*, **19**, 181-194.
24. Pettifer, S., Thorne, D., McDermott, P., Attwood, T., Baran, J., Bryne, J., Hupponen, T., Mowbray, D. and Vriend, G. (2009) An active registry for bioinformatics web services. *Bioinformatics*, **25**, 2090-2091.
25. Missier, P., Paton, N. and Belhajjame, K. (2010) Fine-grained and efficient lineage querying of collection-based workflow provenance. 13th International Conference on Extending Database Technology, Lausanne, Switzerland.
26. De Roure, D. and Goble, C. (2009) Lessons from myExperiment: Research Objects for Data Intensive Research. 2009 Microsoft eScience Workshop, Pittsburgh, PA.

FIGURES

Figure 1. A screenshot of the workflows web page showing (A) the number of workflows written in different languages, (B) a cloud of popular tags used to describe workflows, and (C) the latest workflows submitted to myExperiment. Panels on the right hand side of web pages in myExperiment are used to (D) upload new workflows and digital documents, create new groups, and (E) access content associated with the registered user.

Workflow Entry: Entrez Gene to KEGG Pathway
 Created at: 03/10/07 @ 18:36:00 Last updated: 04/12/09 @ 16:04:39
 | License | Credits (1) | Attributions (0) | Tags (8) | Featured in Packs (0) | Ratings (3) | Attributed By (0) | Favourites By (1) |
 | Citations (0) | Version History | Reviews (0) | Comments (2) |

Version 4 (latest) (of 4) View version: **4 (latest)**

Version created on: 04/12/09 @ 16:04:38 by: Paul Fisher | Revision comments

Title: Entrez Gene to KEGG Pathway
Type: Taverna 1

Preview

(Click on the image to get the full size)

[Download Scalable Diagram \(SVG\)](#)

Description

This workflow takes in Entrez gene ids then adds the string "ncbi-geneid:" to the start of each gene id. These gene ids are then cross-referenced to KEGG gene ids. Each KEGG gene id is then sent to the KEGG pathway database and its relevant pathways returned.

Download

[Download Workflow File/Package \(SCUFL\)](#)

Taverna 1 workflow

Original Uploader

Paul Fisher

License
 All versions of this Workflow are licensed under:

Credits (1)
 (People/Groups)
 Paul Fisher

Attributions (0)
 (Workflows/Files)
 None

Tags (8)

Original Uploader tags

entrez | genotype | kegg | pathway | pathway-driven | pathways | phenotype | shim

[Add Tags](#)

Shared with Groups (0)
 None

Featured In Packs (0)
 None

[Add to your Pack](#)

Ratings (3)
 Hover and click to rate

 Current:
 4.7 / 5
 (3 ratings)
 You haven't rated yet

Figure 2. A screenshot of a workflow web page in myExperiment. A diagram and a description of the workflow is provided together with information about the user who uploaded it onto myExperiment.

Home » Packs » Microarray data analysis using R BOOKMARK

[Manage Pack](#) [Delete Pack](#)

Pack: Microarray data analysis using R

Created at: 02/07/08 @ 11:30:06

[Add an Item](#) | [Sharing](#) | [Tags \(5\)](#) | [Featured in Packs \(0\)](#) | [Favourited By \(0\)](#) | [Comments \(0\)](#)

Title: Microarray data analysis using R

Description

Not set

Items (3)

A **File:** Taverna microarray BMC Bioinformatics paper (Peter Li) ✎ ✕

[Add a comment here]

Added by Peter Li ... less than a minute ago (04/02/10 @ 11:41:11) more ▾

B **Workflow:** Identification of differential genes using the LIMMA Bioconductor package within R (Peter Li) ✎ ✕

[Add a comment here]

Added by Peter Li ... more than 1 year ago (02/07/08 @ 11:31:33) more ▾

C **Workflow:** Identification of differential genes using t-tests by R (Peter Li) ✎ ✕

[Add a comment here]

Added by Peter Li ... more than 1 year ago (02/07/08 @ 11:31:14) more ▾

Download

[Download Pack Items \(ZIP archive\)](#)

Add an Item


Quick add: (a link) Add

eg: "http://www.myexperiment.org/workflows/1" or "http://www.example.com/something-nice"

Quick add: (from your stuff) Add

Advanced add: [Click here](#)

Creator



Peter Li

3 items in this pack

Tags (5)

Creator tags

bioconductor | limma | maxd | microarray | r [edit]

[Add Tags](#) ▾

Tags from Items (8)

beanshell | bioconductor | limma | maxd | microarray | pdf | r | user_interaction

Shared with Groups (0)

None

Featured In Packs (0)

None

[Add to your Pack](#) ▾

Favourited By (0)

No one

[Add to your Favourites](#)

Statistics

491 viewings
81 downloads
[\[see breakdown \]](#)

[More](#) ▾

Figure 3. A screenshot of a myExperiment pack for analysing microarray data containing (A) a PDF file of a scientific paper and (B and C) two workflows described in the publication.

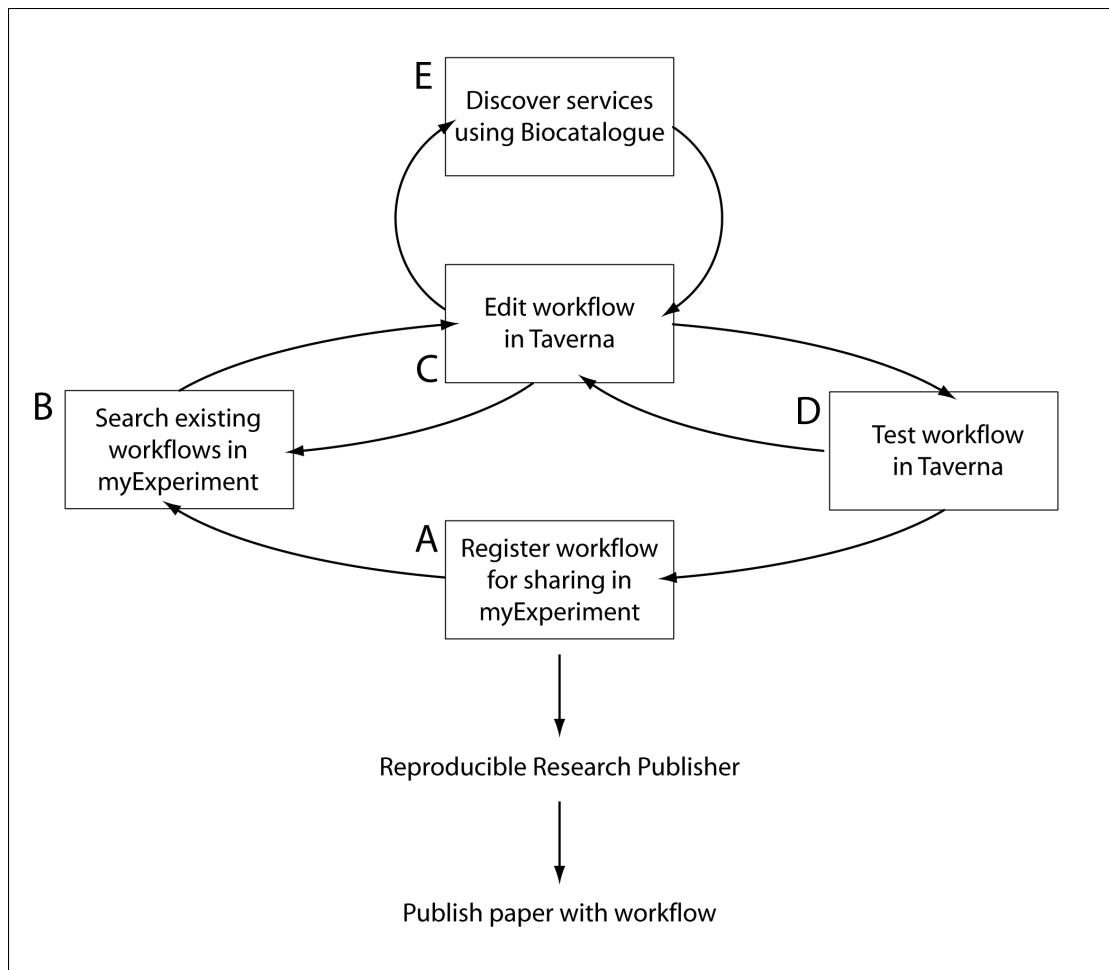


Figure 4. A schematic diagram showing the workflow life cycle. Workflows can be stored in myExperiment (A) enabling them to be shared and queried (B). Downloaded workflows are edited (C) and their enactment tested using applications such as Taverna (D). The editing of workflows may involve the addition of new web services which can be discovered using registries such as that provided by BioCatalogue (E). In the future, a Reproducible Research Publisher component can help with the replication of the data analysis implemented by workflows for scientific publication.