A portal interface to myGrid workflow technology

Stefan Rennick Egglestone [a], M.Nedim Alpdemir [b], Chris Greenhalgh [a], Arijit Mukherjee [c] and Ian Roberts [d]

a. School of Computer Science and IT, University of Nottingham
sre@cs.nott.ac.uk, cmg@cs.nott.ac.uk

b. School of Computer Science, University of Manchester
alpdemim@cs.man.ac.uk

c. School of Computing Science, University of Newcastle upon Tyne
Arijit.Mukherjee@newcastle.ac.uk

d. Department of Computer Science, University of Sheffield
i.roberts@dcs.shef.ac.uk

1. Introduction

A rich selection of computational resources are available to scientists working with biological data, and it is common for these scientists to wish to perform composite analyses which use a number of such resources. To support the automation of the performance of these analyses, the myGrid project have developed the Taverna workflow workbench [2]. This is a graphical interface which allows a user to construct a workflow to represent an analysis, and to enact this workflow, thereby performing the analysis and generating results.

Taverna has been successfully used in a number of research projects, including investigations into Williams-Beuren Syndrome [4] and Grave's disease [1], and is growing in popularity in the bioinformatics user community. Workflows constructed for these projects have been used to gather substantially large volumes of data than would be possible manually. However, experience gained in these projects has revealed that the construction of workflows is a difficult process, which often requires significant bioinformatics expertise. As such, workflow construction is likely to performed by a minority of users.

Dialog with the user community has revealed that, although a minority of users are involved in workflow construction, workflows that have been constructed by a small number of expert users are often enacted by a much larger group of less-expert users. Serialized representations of these workflows are often distributed to these users by email or via shared network areas, and are then loaded into an installation of the Taverna workflow workbench for enactment, with any results generated being saved to the local file system. Files containing these results can then be shared with other users through similar mechanisms.

Although this approach does work, it also has some problems. Generic storage and communication systems cannot provide specific support for workflow data, and such systems may

lack in provision for security. As an alternative to this approach, the myGrid project have developed the myGrid Portal Interface (MPI), a simple, web-based interface which provides specific support for the storage and enactment of workflows and the archiving of data produced by these enactments. This paper describes a typical use-case for this interface, technical details of which have previously been described in [3]. Section 2 of this paper briefly introduces the workflow involved in this use-case, section 3 gives details of the use-case itself, and section 4 suggests a number of criteria against which the ease of use of this interface for this particular use-case might be evaluatated.

2. Workflow used in use-case

A previous collaboration involving the myGrid project has resulted in three workflows, labelled A, B and C, which have been constructed in Taverna and which have been used during investigations into Williams-Beuren Syndrome (WBS), a disease with a genetic basis. Experience gained during the construction of these workflows has been used to improve the design of the Taverna workflow workbench, and has also been useful during the design of the MPI. This section of the paper briefly describes the function of workflow B, which is the most complex of the workflows that have been constructed during the WBS investigation. Section 3 of the paper then describes how the MPI can be used to enact workflow B and to browse results produced by this enactment.

2.1 Details of workflow B

Workflow B has been constructed to support the characterization of a DNA sequence. This sequence must be provided as an input parameter when the workflow is enacted, and the results of enacting this workflow include large volumes of contextual information about this sequence which have been gathered from a number of distributed bioinformatics databases. This contextual information includes a prediction of which regions of the sequence may be involved in promotion and signalling activities, and a prediction of any proteins that may be transcribed from the input sequence. In earlier versions of this workflow, contextual information gathered by a workflow enactment was presented to a user as raw text. However, a more sophisticated version of workflow B has been produced which presents this information using an integrated HTML visualization. It is this later version of workflow B which is used in section 3.

3. Enacting workflow B in the MPI

The MPI provides facilities to store and enact workflows, and to browse and store results produced by these workflow enactments. This section describes the steps required to use these facilities to enact workflow B and browse the integrated HTML visualizations it produces.

Step 1: user login

To protect workflows and data stored in an MPI installation, each user must be allocated an account in the installation, with each account being protected by a username and password. To login to their account, a user must use their web browser to navigate to an initial login page, into which they can enter their username and password. Once a user has successfully authenticated themselves, they are then given access to web-pages which

expose the functionality provided by the MPI.

Step 2: workflow upload

Once a user has constructed a workflow in Taverna, it must be saved to a file on the local file system. This file can then be uploaded into an MPI installation for later enactment.

Workflows which have been uploaded into an MPI installation are grouped into collections. A logged-in user can view a list of all existing collections, as shown on figure 1 below, and has the option of creating a new, named collection.

Figure 1: Available collections of workflows

To upload a workflow into the MPI, a user must first select a collection to upload the workflow into. They are then presented with a list of workflows already available in that collection, as shown in figure 2 below.

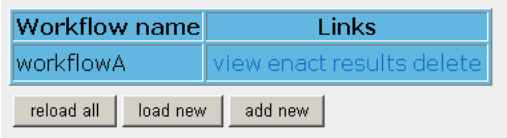

Figure 2 : Workflows that have been uploaded into a collection

On clicking the button labelled *add new*, they are provided with a form which they can use to upload a new workflow into this collection, into which they must enter a unique name for the workflow and the path of the file on their local file system from where the workflow will be uploaded. An example of such a form is shown in figure 3 below.

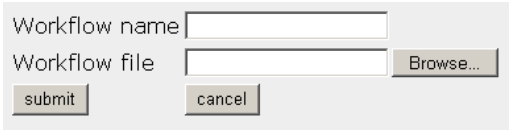

Figure 3 : Form used to upload a new workflow

Step 3 : start workflow enactment

Once a workflow has been uploaded into a collection in the MPI, a user can choose to enact it. To do so, they must provide any input values required by the workflow. In the case of workflow B, they are required to provide an item of text containing a DNA sequence, and the MPI automatically constructs a form into which this input can be entered, as shown in figure 4 below.

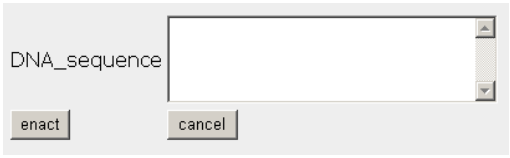

Figure 4 : Form to collect input value for workflow B

Once a user has entered an input value into this form, they can then click the button labelled *enact* and start an enactment of workflow B. Each enactment is allocated an *submission ID*, of which the submitting user is notified.

Step 4 : monitor the enactment of the workflow

The MPI provides facilities to monitor enactments which have either completed or are in progress. Figure 5 below shows information provided about one enactment of workflow B, which indicates that this enactment has completed and has produced results which are ready for browsing.

Figure 5 : Indication of status of workflow enactment

Step 5 : browse results produced by workflow enactment

Once an enactment has completed, it appears on a list of completed enactments for that workflow, identified by its submission ID and by the time and date that the enactment started and ended. An example of such a list is shown in figure 6 below.

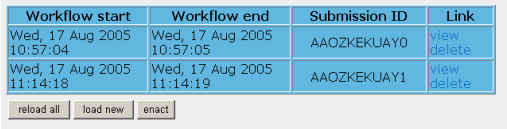

Figure 6 : A list of completed enactments of a workflow B

A user can then choose an enactment from this list for which they which they wish to browse results, and can navigate to an enactment summary page, as shown in figure 7 below.

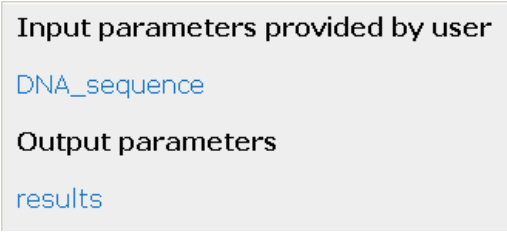

Figure 7 : Summary page for a chosen enactment

This summary page provides hyperlinks which can be used to view the input parameters used to start the enactment, and the output parameters which have been produced by the enactment. In this instance, a user can click on the hyperlink labelled *results* to view the HTML visualization produced by the enactment of workflow B, an example of which is shown in figure 8 below. This visualization displays the contextual information gathered by the workflow enactment as an image, and clicking on interesting regions of this image allows a user to navigate to further details about the contextual information represented by that region of the image.

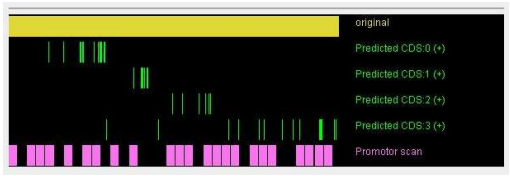

Figure 8 : Results produced by an enactment of workflow B

4. Discussion

The MPI, a typical use of which has been described in this paper, is an initial prototype of a web-based system which supports the storage and enactment of workflows, and the archiving and browsing of results produced by these enactments. Although this software is still in the prototyping stage, it is intended that it will be developed into a production-quality system that will hopefully be of benefit to the bioinformatics community, and especially the subset of this community who are already users of Taverna.

As the MPI is developed further, it is anticipated that a number of issues will have to be addressed. Two important issues which may arise are described in sections 4.1 and 4.2 below.

4.1 Data sharing model

The MPI currently supports a course-level model for data sharing, in which all users who have accounts in a particular MPI installation have shared access to all workflows and enactment data that have been stored in that installation. It is anticipated that this style of data sharing will be sufficient if a particular installation is shared by a small group of users, who produce relatively small volumes of workflows and results. However, for larger groups of users producing larger volumes of data, a more sophisticated approach may be required. It may be the case, for example, that the simple collection

structure provided by the current MPI prototype would be insufficient to organize the larger volumes of workflows that a larger group of users may potentially produce. One potential solution to this problem is to allow the user who has uploaded a particular workflow to control the visibility of the workflow to other users. Such a user may be provided with the options of marking a workflow as being private, ie only visible to themselves, as being visible to a subset of users of the MPI installation, or as being publicly visible to all users of the installation. Such an approach might successfully reduce the number of workflows visible to any particular user, but does have the disadvantage of necessitating a more complex design for the user interface of the MPI.

4.2 Assessing the usability of a web-based interface

One advantage of providing an interface to workflow enactment and storage technology using a web-based mechanism is that only a standard web-browser is required to access the interface. Since the vast majority of bioinformaticians will have such technology available pre-installed on their desktop machines, this means that these users will not be required to install additional software to use the interface. However, web-based interfaces are necessarily more limited in design than those which can be provided by a desktop application, and must therefore be carefully constructed so as to be as simple as possible to use. It is anticipated that, especially as more features are added to the MPI prototype, extensive user trialling will be required to ensure that the interface design remains acceptably easy to use.

5. Conclusion

The myGrid Portal Interface (MPI) is an initial prototype of a simple, web-based interface to myGrid workflow enactment and storage technology. This paper has described a typical use-case for this interface, and has outlined issues which will have to be taken into account in its further development. It is hoped that the MPI will eventually become production-quality system that will be of use to the bioinformatics community.

References

[1] Matthew Addis et al. *Experiences with e-science workflow specification and enactment in bioinformatics*. pages p.459–467. Proceedings of the UK e-Science All Hands Meeting 2003. ISBN - 1-904425-11-9.

[2] Tom Oinn et al. *Taverna: A tool for the composition and enactment of bioinformatics workflows*. Bioinformatics, 20(17):3045–3054, 2004.

[3] Stefan Rennick Egglestone, M. Nedim Alpdemir, Chris Greenhalgh, Arijit Mukherjee and Ian Roberts. *A portal interface to myGrid workflow technology*. Proceedings of UK e-Science All Hands Meeting 2005.

[4] R. Stevens et al. *Exploring Williams Beuren Syndrome using myGrid.* pages i303–i310. 12th International Conference on Intelligent Systems in Molecular Biology, 2004. published Bioinformatics Vol. 20 Suppl. 1.