USER PROXY SERVICE IN MYGRID

Milena Radenkovic
School of Computer Science and IT, University of Nottingham
Jubilee Campus, Wollaton Road, NG8 1BB

ABSTRACT

This paper is concerned with supporting collaboration among bioinformatics scientists across multiple organizations in Web-Services based myGrid middleware. MyGrid aims to deliver a personalized collaborative problem-solving platform for e-Scientists working in a distributed environment. It allows the users to construct long-lived in silico experiments, find and adapt others' experiments, publish their own view in shared repositories, and be better informed as to the provenance of the tools, data and users directly relevant to them. This paper focuses on asynchronous interactions that allow collaborative information retrieval and collaborative workflow orchestration. In particular, we discuss the design and deployment of the user proxy service that is a context-dependent notification service client for heterogeneous annotated services including data, metadata and workflow services.

KEYWORDS

Collaboration, Web Services, Searching and Browsing, Metadata, Web-Portal, e-Science

1. INTRODUCTION

The anticipated tremendous growth of quantity and distribution of bioinformatics resources over the Internet over the last ten years means that finding and utilising these resources effectively requires building sophisticated tools and middleware frameworks that would support scientists in conducting their research. Today many genome laboratories are typically involved in any activity, and many resources and scientists need to be tied together to accomplish a useful goal. Because they typically involve only loose cooperation between different laboratories, the need for real-time cooperative work tools is relatively small and support for collaboration should focus on retrieval and analysis of data and literature and on extending the same facilities to more informal material that is crucial to progress in science [National Research Council, 1993]. Most of the current collaborative toolkits for e-science enable two or more people to access documents, applications, resources and interfaces, as well as interlinked collection of personal and public workspaces. Workflow management and collaborative workflow is especially important part for supporting bioinformatics research. Groupware systems mostly only bundle some basic workflow capabilities that can be used to process workflows. Consequently, there is an alarming need for more sophisticated collaborative toolkits that would allow non-trivial and scalable collaboration among geographically dispersed users.

Notification of activity in a shared annotated information system is long known to support awareness and improve asynchronous collaboration. The major contribution of this paper is that it discusses integration of notifications that is tailored to suit the needs of bioinformatics users within a semantically rich middleware. Overall awareness of activities on workflow, metadata, service directory and data services and other users is crucial for collaboration among bioinformaticians. We designed and deployed an enhanced metadata driven user proxy that acts as a notification service client and can significantly increase this awareness of myGrid users. It includes more information in notification messages, support for multiple communication channels through which notifications can be received, easy customization of subscription and notification mechanisms and support for complex queries.

This paper is organized as follows. Section 2 begins by reviewing current collaboration web based technologies that are relevant to myGrid. Section 3 presents an overview of myGrid architecture. Section 4 first discusses general approach and goals for integrating collaboration support within myGrid, and then describes and demonstrates collaboration support prototype in myGrid that is based on myGrid's rich

provenance support and notification service. Section 5 gives final concludes and identifies potential future work.

2. RELATED WORK: COLLABORATION

Web based tools for collaborative information sharing have a number of distinct advantages over non-Web based facilities. Their primary goal is to achieve cross-platform interoperability to ease cross-platform and cross-organisational cooperation and in dispersed research projects. There is already a large number of existing prototype shared workspaces systems that offer various attributes and features including: containers to store and retrieve resources, administration of members of a shared workspace, generation and distribution of meta-information. This section reviews some of them that are of greatest relevance to how collaboration is supported in myGrid.

Empirical evidence suggest that systems which provide access to shared information, at any time and place, and using minimal technical infrastructure, are main requirement of groups collaborating in a decentralised working environment [Klckner, 1999]. A well known example of a system supporting this is a shared BSCW workspace can contain different kinds of information such as documents, pictures, URLs, threaded discussions [Appelt, Mambrey, 1999]. Shared annotations on digital documents are an attractive means of asynchronous collaboration. As an effective means of communication, however, annotations have a major flaw. Interaction is primarily between person and document, not person and person. As a result, communicating ideas is often slow and cumbersome. People must revisit a document to see the latest comments. In BSCW [Appelt, Mambely,1999], icons indicate recent document activity: reading, editing, or versioning. Clicking on the icon retrieves information about time and actor. The Annotator [Ovsiannikov, I., Arbib, M., and McNeill, T, 1999] and ComMentor [Röscheisen, M. Mogensen, C. and Winograd, T, 1997] annotation systems allow people to search the set of annotations made on a document. This provides information about new annotations, but requires additional work by the user.

One way to address person-to-person communication problem is to integrate a notification mechanism into a shared annotation system. When a new annotation is added, interested parties are notified and can revisit a document to read more, add a reply, or contribute new comments. Several systems [Leland, M., Fish, R., and Kraut, R , 1988], [Appelt, 1999], have used this approach with varying degrees of success. Awareness and notifications have long been recognized as important aspects of both synchronous and asynchronous document collaboration systems. A study of collaborative writing by [Baecker et al, 1993] stressed the importance of mutual awareness, knowledge of the state or actions of collaborators. [Dourish, Bellotti, 1992] discuss the importance of passive awareness, “an understanding of the activities of others, which provides a context for your own activity”.

Although notification mechanisms in shared annotations systems are common, there has been little work in integrating within the middleware infrastructures that support bioinformatics users and tailoring them to the needs of such users in order to improve their collaboration.

3. MYGRID: ARCHITECTURAL OVERVIEW

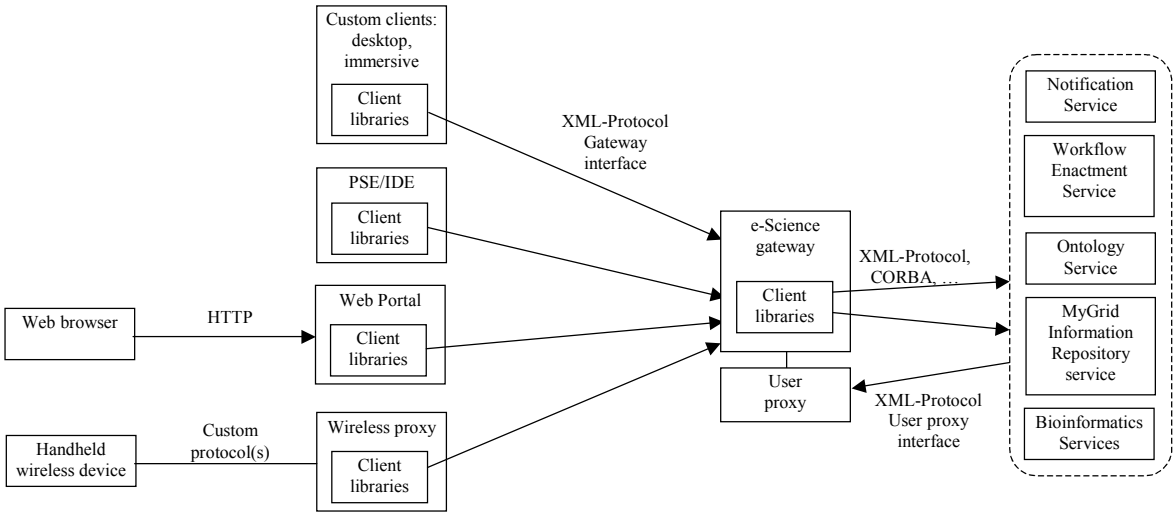
MyGrid is a Grid middleware project that aims to help bioinformaticians and biologists to perform collaborative workflow-based in-silico experiments and help them automate the management of such workflows through personalization, provenance and notification of change. In biological sciences it is not the quantity but the complexity of the data that matters [Moreau L., 2002]. Most biologists have fairly modest requirements for computer power. What is most important to them is discovering resources, discovering tools, and being able to capture when, where, and why they did a certain task as well as collaborate with the rest of the researchers. MyGrid is being developed as a semantically rich layer to sit on top of a computational Grid infrastructure, and will eventually enable scientists to find and retrieve the data and applications they want. In these environments complex heterogeneous data are under constant change. Term *data* in this context refers to: row data, provenance, annotations, versioning, results, partial results.

MyGrid adapts a service-oriented approach i.e. all resources including devices, computer facilities, file stores and algorithms are exposed as Web services. MyGrid allows large numbers of repositories and tools to

be involved in the computation process, complex queries, transparent manipulation of multiple data and metadata sources, suspending and resuming workflows, observing workflow progress, analysing their progress [Wroe et al, 2003]. Figure 1 illustrates high level overview of myGrid architecture that provides the following functionalities: database access (including access to myGrid Information Repository (MIR)), domain-specific computational services and service discovery services, workflow enactment engine, notification and metadata services as well as the user proxy and gateway services.

MIR contains domain entities, domain entity groups, conceptIDs and workflow provenance records. Each item in MIR has a unique ID and additional information about the item such as e.g. creation date, author, annotations, etc. The ontology we use in myGrid is founded in DAML-S and it can dynamically create service and data classifications. These classifications are then used to support semantic service metadata and discovery in myGRID and workflow orchestration. [Wroe at al, 2003] discusses in more detail the utility of DAML+OIL in creating dynamic classifications based on formal description. In this paper we mainly utilize myGrid's capability that the type of a workflow input/output as well as classification of a workflow can be defined with reference to the ontology.

We term gateway and user proxy isolate user from detailed operations within the core myGrid



architecture and add value in respect to collaboration and personalization. Gateway has port types that provide common and relatively abstract view of the functionality of myGrid. These port types include: *Data* that supports retrieval of a value associated with some identifier(s) e.g. domain entities, workflow definitions, *MutableData* that supports deletion and replacement (versioning) of data accessible via the Data Interface, *MetaData* that supports retrieval of metadata associated with some identifier(s) whether first party or third party, including metadata associated with data objects, services, activities, people and organization. *MutableMetadata* that supports addition and modification of metadata and *Action* that supports invocation and monitoring of individual actions or activities. These port types allow users to easily follow and contribute to other users' annotation and workflow activities.

Figure 1. myGrid deployment perspective

4. COLLABORATION IN MYGRID

myGrid aims to provide support for asynchronous collaboration among locally distributed, loosely organized bioinformatics users and provide them with information about the activities within their workspaces. Provenance (and annotations) and notification are the basis for asynchronous interactions in myGrid. User proxy/gateway is user's persistent representation of myGrid.

Rich provenance in MIR supported with rich and generic interface allows support for collaboration. Since every data in MIR has variety of additional metadata data (such as author name, date and context of work), users can browse MIR based on any of these criteria (e.g. users are allowed to list all the people working on a particular problem and view their results). Users are allowed to add their annotations to any item in the MIR

if they have permission for it. MIR keeps rich provenance data about the workflows including: input types, output types, intermediate results, operation information, services used and annotations made. This allows the users to list all the users who were constructing workflows with the same inputs, and/or the same operation type, and/or same context. Users can either explicitly query MIR for these information, or receive them via the notification service which is currently fully integrated with the workflow enactment engine. At present topics affecting collaboration over a workflow include: *PendingInputRequest* (all of the users subscribed to this topic can be asked for adding their input to a running workflow), *IntermediaryResults* and *ViewWFProvenanceData* (all of users subscribed to these topic can be notified when other members of the team working on the same workflow needs them to view the results and add their comments to them). The scenario that we describe demonstrates collaborative workflow in which there are multiple scientists working on a workflow and contributing their data and knowledge to running that workflow. In a collaborative scenario, the workflow (its WSFL definition) publishes under 3 topics: *JobStatus* (notifiying the users about the job status of the workflow e.g. *Suspended*, *Completed*, *Running*), *PendingInputRequest* and *ViewWFProvenanceData*, and all the members of the team working on this project will be subscribed to these topics. When a workflow has an intermediary result, each user may view it and discuss with the other users what they think the next step should be (currently suspend, stop or continue workflow are supported) or input the data requested by the workflow if they have the data ready.

User proxy contains four databases: *Subscriptions* that contains list of users' active subscriptions; *UserProfiles* that contains keywords, associated concepts from the ontology service, relationships to other concepts from the MIR, and possibly more complex topics that logically relate keywords and concepts (e.g. AND, OR based expressions); *TopicResults* that contains notifications, *Topics* that contains list of existing topics user can subscribe to and *WorkContext* that user proxy keeps in order to keep track of the history of user's work contexts and dynamically update the users subscription list. User proxy always matches all the published notifications to see if the same context is mentioned in *UsersProfile* or *WorkContext*, and if it does it notifies the user.

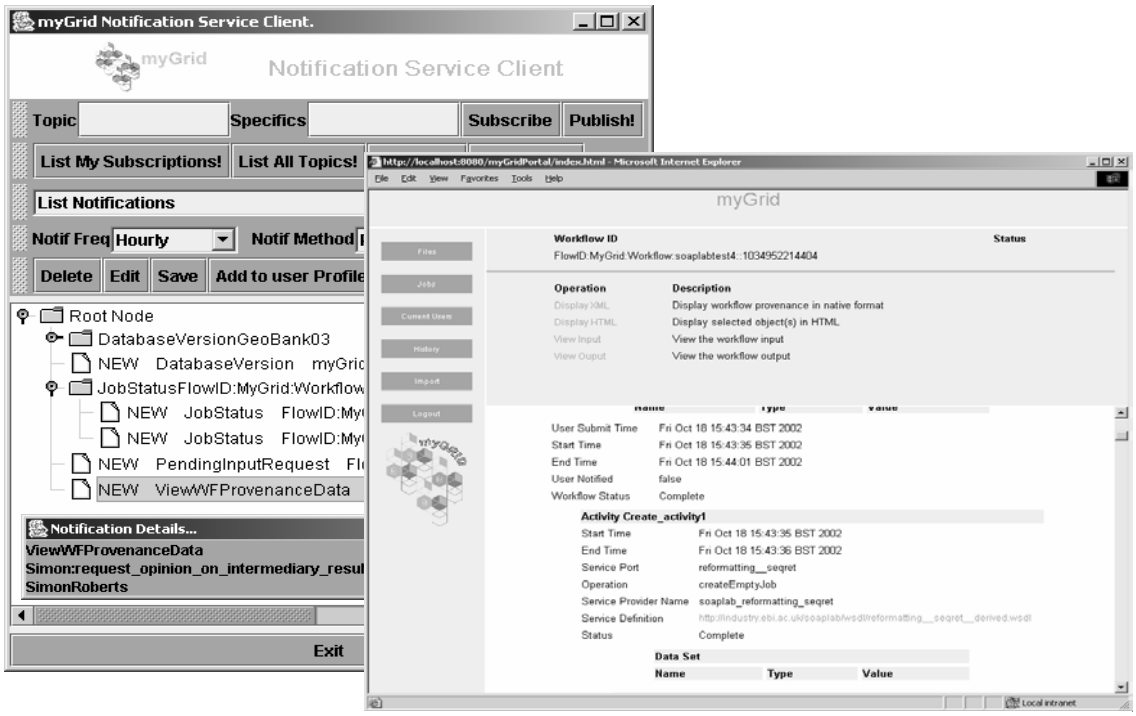


Figure 3 (a) Viewing Notifications (b) Viewing workflow provenance

Process of orchestrating a workflow is often very complex and involves multiple steps where for each step each of the users who are collaboratively working on that workflow can be notified to provide their input (i.e. either in a form of a selection, or new data, or only view data and make annotations). The places where

collaboration is required by multiple scientist are specified in the WSFL definition of a workflow which defines workflow interaction points with the notification service. At those points, the workflow publishes the message under any of topics *JobStatus*, *ViewWFProvenanceData* or *PendingInputRequest*. Figure 2a shows an example where a user who has not started the workflow has received notification in which he is required to view the data the workflow put in MIR and provide comments for the other members of the team working on that workflow. Figure 2b shows a small part of workflow provenance data captured in MIR.

5. CONCLUSION

Even though myGrid is still in its early stages, it has already been used in a variety of genome academic and industrial research projects. It provides an efficient environment for scientists to share their both formal and informal knowledge and run workflows collaboratively. Work to date mainly concentrated on designing and implementing generic myGrid services and core architecture. This paper discussed collaborative aspects of myGrid based on its rich provenance data and semantically driven notification service. We proposed the design for user proxy service that adds value to myGrid with respect to collaboration and that was already deployed in myGrid. Future work will develop more dynamic notification service topic management and further integrate notification service to service directory service and MIR as well as support for more synchronous collaboration.

REFERENCES

Appelt, W., Mambrey, 1999, WWW Based Collaboration with the BSCW System, Proc. *SOFSEM*, Springer Lecture Notes in Computer Science 1725, 66-78

Bernheim B. et al, 2002, Notification for Shared Annotation of Digital Documents, Proc. *CHI* , p 89-96

Cadiz, J. Venolia, G. Jancke, G. Gupta, 2002, A. Sideshow: Providing Peripheral Awareness of People and Information, *Microsoft Technical Report #MSR-TR-2001-83*.

National Research Council, 1993, National Collaboratories, *Applying Information Technology for Scientific Research*, National Academy Press, Washington D.C. 1993

Dourish P. and Bellotti, 1992 V. Awareness and Coordination in Shared Workspaces. Proc. CSCW , 107-114.

Gutwin, C. Roseman, M. and Greenberg, S, 1996,. A Usability Study of Awareness Widgets in a Shared Workspace Groupware System. Proceedings of CSCW 96, 258-267

Klckner, K. 1999, Computer Supported Cooperative Learning (CSCL): A state of the art', Proc. of the 19th *World Conference on Open Learning and Distance Education*, Vienna [CD-ROM], Interactive World, Hagen.

Leland, M., Fish, R., and Kraut, R, 88 Collaborative Document Production Using Quilt, *Proc. of CSCW*, p.206-215.

Luc Moreau et al, 2002, On the Use of Agents in a BioInformatics Grid. In Network Tools and Applications in Biology (NETTAB'2002) - Agents in Bioinformatics, Bologna, Italy,

Ovsiannikov, I., Arbib, M., and McNeill, T. 1999, Annotation Technology, Int. J. *Human Computer Studies*, 50, 329-362.

Röscheisen, M., Mogensen, C., and Winograd, T., 1997, *Shared Web Annotations as a Platform for Third-Party Value-Added, Information Providers: Architecture, Protocols, and Usage Examples*, *Technical Report CSDTR/DLTR,* Stanford University.

Wroe c. et al, 2003 A suite of DAML+OIL to Discrete Bioinformatics Web Services and Data, In *International Journal of Cooperative Information Systems Special Issue on Bioinormatics*, ISSN 02188430