Contextualised Workflow Execution in MyGrid

M. Nedim Alpdemir[1], Arijit Mukherjee[2], Norman W. Paton[1], Alvaro A.A. Fernandes[1], Paul Watson[2], Kevin Glover[3], Chris Greenhalgh[3], Tom Oinn[4], and Hannah Tipney[1]

[1] Department of Computer Science, University of Manchester, Oxford Road, Manchester M13 9PL, United Kingdom
[2] School of Computing Science, University of Newcastle upon Tyne, Newcastle upon Tyne NE1 7RU, United Kingdom
[3] School of Comp. Sci. and Inf. Tech., University of Nottingham, Jubilee Campus, Wollaton Road, Nottingham NG8 1BB, UK
[4] European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom

Abstract. e-Scientists stand to benefit from tools and environments that either hide, or help to manage, the inherent complexity involved in accessing and making concerted use of the diverse resources that might be used as part of an *in silico* experiment. This paper illustrates the benefits that derive from the provision of integrated access to contextual information that links the phases of a problem-solving activity, so that the steps of a solution do not happen in isolation, but rather as the components of a coherent whole. Experiences with myGrid workflow execution environment (Taverna) are presented, where an information model provides the conceptual basis for contextualisation. This information model describes key characteristics that are shared by many e-Science activities, and is used both to organise the scientist's personal data resources, and to support data sharing and capture within the myGrid environment.

1 Introduction and Related Work

Grid-based solutions to typical e-Science problems require the integration of many distributed resources, and the orchestration of diverse analysis services in a semantically rich, collaborative environment [5]. In such a context, it is important that e-Scientists are supported in their day-to-day experiments with tools and environments that allow the principal focus to be on scientific challenges, rather than on the management and organisation of computational activities.

Research into Problem Solving Environments (PSEs) has long targeted this particular challenge. Although the term *Problem Solving Environment* means different things to different people [4], and its meaning seems to have been evolving over time, a number of common concepts can be identified from the relevant research (e.g. [4, 6, 12, 8]). For example, the following features are commonly supported: problem definition; solution formulation; execution of the problem solution; provenance recording while applying the solution; result visualisation and

P.M.A. Sloot et al. (Eds.): EGC 2005, LNCS 3470, pp. 444–453, 2005.

analysis; and support for communicating results to others (i.e. collaboration). Although these are among the most common features, different PSEs add various other capabilities, such as intelligent support for problem formulation and solution selection, or highlight a particular feature, such as the use of workflow (e.g. [3, 1, 11]).

This paper emphasizes a specific aspect that has been largely overlooked, namely the provision of integrated access to *contextual information* that links the phases of a problem-solving exercise in a meaningful way. In myGrid, the following are principles underpin support for contextualisation:

Consistent Representation: The information model ensures that information required to establish the execution context conforms to a well-defined data model, and therefore is understood by the myGrid components that take part in an experiment as well as external parties.

Automatic Capture: When a workflow is executed, contextual information is preserved by the workflow enactment engine, and used to annotate both intermediate and final results.

Long-term preservation: The contextual information used to organise the persistent storage of provenance information on workflows and their results, easing interpretation and sharing.

Uniform Identification: Both contextual and experimental data are identified and linked using a standard data and metadata identification scheme, namely LSIDs [2].

The rest of the paper describes contextualisation in myGrid, indicating both how contextualisation supports users and how the myGrid architecture captures and conveys the relevant information. As such, Section 2 summarises the information model, which is at the heart of contextualisation. Section 3 provides some concrete examples of contextualisation in practice, in a bioinformatics application. Section 4 provides an architectural view of the execution environment, and finally Section 5 presents some conclusions.

2 The Information Model

The myGrid project (`http://www.mygrid.org.uk/`) is developing high-level middleware to support the e-Scientist in conducting *in silico* experiments in biology. An important part of this has been the design of an Information Model (IM) [9], which defines the basic concepts through which different aspects of an e-Science process can be represented and linked. By providing shared data abstractions that underpin important service interactions, the IM promotes synergy between myGrid components. The IM is defined in UML, and equivalent XML Schema definitions have been derived from the UML to facilitate the design of the myGrid service interfaces.

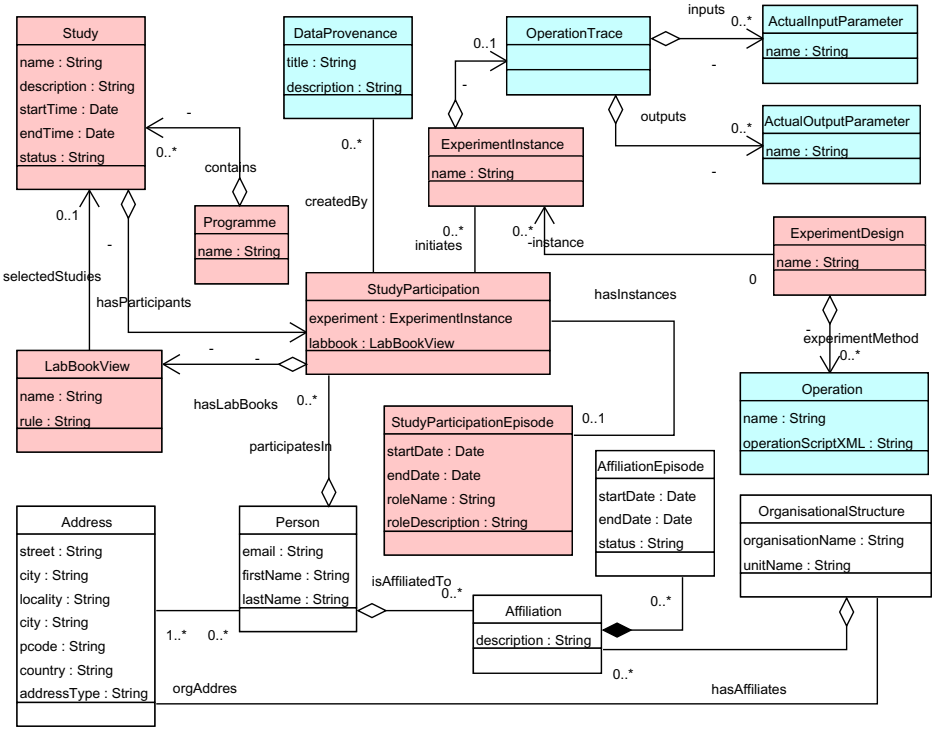

Fig. 1. A UML class diagram providing an overview of the information model

Figure 1 illustrates several of the principal classes and associations in the IM. In summary; a *Programme* is a structuring device for grouping other *Studies* and can be used to represent e.g. a project or sub-project. An *Experiment Design* represents the method to be used (typically as a workflow script) to solve a scientific problem. An *Experiment Instance* is an application of an *Experiment Design* and represents some executing or completed task. The relationship of a *Person* with an *Organizational Structure* is captured by an *Affiliation*. A *Study Participation* qualifies a person's relationship to the study by a set of study roles. An *Operation Trace* represents inputs, outputs and the intermediate results of an experiment (i.e. the *experiment provenance*), as opposed to the *Data Provenance* which primarily indicates a data item's creation method and time.

An important feature of the IM is that it does not model application-specific data, but rather treats such data as opaque, and delegates responsibility for its interpretation to users and to application-specific services. As such, concepts such as *sequence* or *gene*, although they are relevant to the Williams-Beuren Syndrome (WBS) case study described in Section 3.1, are not explicitly described in the IM. Rather, the IM captures information that is common to, and may even be shared by, many e-Science applications, such as scientists, studies and

experiments. A consequence of this design decision is that the myGrid components are less coupled to each other and to a particular domain, and so are more easily deployable in different contexts. However, interpretation and processing (e.g. content aware storage and visualisation) of the results for the end user becomes more difficult, and falls largely on the application developer's shoulders.

3 Contextualized Workflows: A User's Perspective

3.1 Case Study

Informatic studies in Williams-Beuren Syndrome (WBS) are used to illustrate the added value obtained from contextualisation. WBS is a rare disorder characterized by physical and developmental problems. The search for improved understanding of the genetic basis for WBS requires repeated application of a range of standard bioinformatics techniques. Due to the highly repetitive nature of the genome sequence flanking the Williams-Beuren syndrome critical region, sequencing of the region is incomplete, leaving documented gaps in the released genomic sequence. In order to produce a complete and accurate map of the region, researchers must constantly search for newly sequenced human DNA clones that extended into these gap regions [10].

Several requirements of the WBS application stand to benefit from integrated support for contextualisation:

- The experimenter needs to conduct several tasks repeatedly (e.g. execution of follow-on workflows), which requires the inputs, outputs and intermediate results of one step to be kept in the same experimental context, to ease comparisons of multiple runs and of alternative approaches.
- Results or experimental procedures need to be shared among researchers in the group. A contextualized environment helps scientists to migrate from ad-hoc practices for capturing the processes followed and the results obtained, to computer-supported information-rich collaboration schemes.

3.2 Contextual Data in Use

This section illustrates how integrated support for contextualisation surfaces to the user. In myGrid, workflows are developed and executed using the Taverna workbench [7], which is essentially a workflow editor and a front-end to a workflow execution environment with an extensible architecture, into which additional components can be plugged. The following myGrid plug-ins provide users with access to external information resources when designing and executing workflows:

MIR Browser: The myGrid Information Repository (MIR) is a web service that provides long-term storage of information model and associated application-specific data. The plug-in supports access to and modification of data in the MIR.

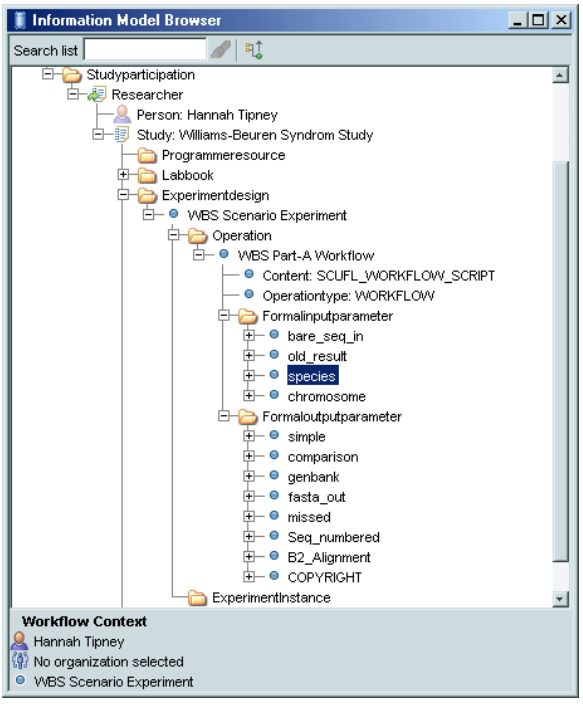

Fig. 2. MIR Browser displaying context information

Metadata Browser: The myGrid Metadata Store is a web service that supports application-specific annotations of data, including data in the MIR, using semantic-web technologies. The plug-in supports searching and browsing of information in the metadata store, as well of the addition of new annotations.

Feta Search Engine: The Feta Search Engine provides access to registry data on available services and resources.

Users are supported in providing, managing or accessing contextual data using one or more of the above plug-ins, and benefit from the automatic maintenance of contextual data in the MIR.

When developing or using workflows, the e-Scientist first launches the Taverna workbench, which provides workflow-specific user interface elements, as well as the plug-ins described above. When the MIR browser is launched, the user provides login details, and is then provided with access to their personal instance of the information model, as illustrated in Figure 2. This figure shows how access has been provided, among other things, to: (i) data from the studies in which the e-Scientist is participating – in this case a *Williams-Beuren Syndrome Study*; (ii) the experiment designs that are being used in the study – in this case a single *WBS-Scenario Experiment*; and (iii) the definitions of the workflows

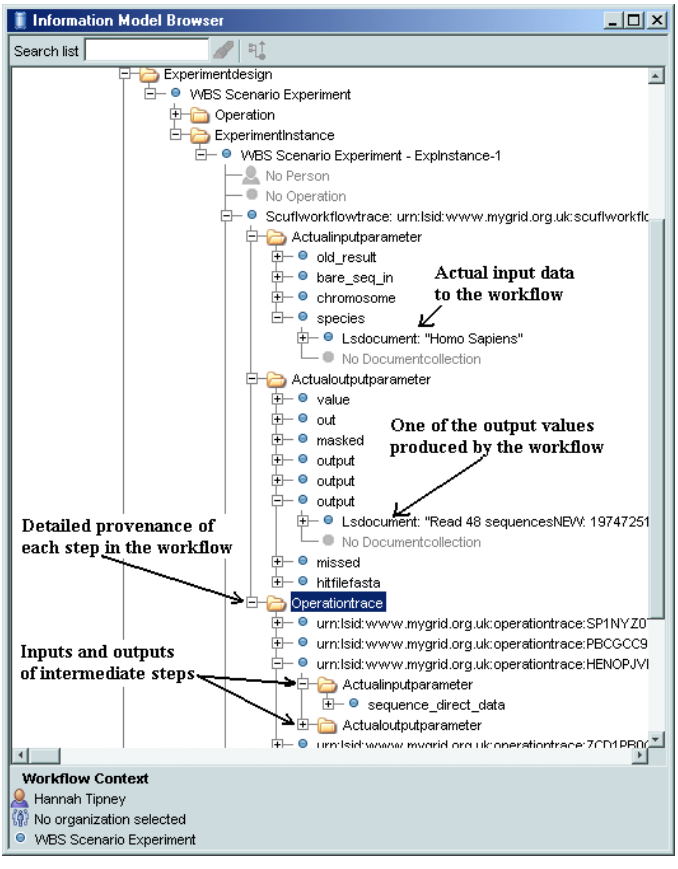

Fig. 3. Workflow execution results in the MIR Browser

that are used in the *in silico* experiments – in this case a single workflow named *WBS part-A workflow*. In this particular case, the absence of a + beside the *ExperimentInstance* indicates that the *WBS-Scenario Experiment* has not yet been executed. At this point, the e-Scientist is free either to select the existing workflow, or to search for a new workflow using Feta search engine. Either way, it is possible for the workflow to be edited, for example by adding new services discovered using Feta to try variations to an existing solution.

When a workflow is selected for execution, the e-Scientist can obtain data values for use as inputs to the workflow from previous experiment results stored in the MIR. The execution of follow-on analyses is common practice in the WBS case study.

The user's view in Figure 2 illustrates that resources, such as individual workflows, do not exist in isolation, but rather are part of a context – in this case a study into WBS. There may be many experiments and workflows that are part of that study. In addition, when a workflow from a study is executed, it is executed in the context of that study, and the results of the workflow are

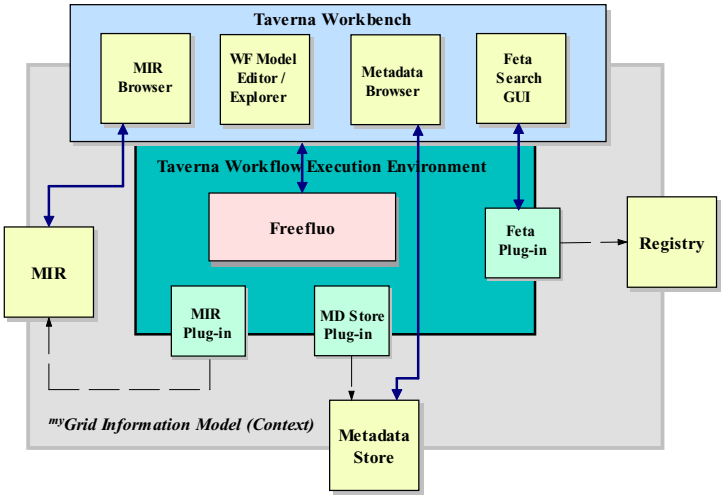

Fig. 4. A simplified architectural view

automatically considered to be among the results of the study. For example, Figure 3 illustrates the results obtained by executing *WBS part-A workflow* from Figure 2. The results of the execution are automatically recorded under the *Experiment Instance* entity, which is associated with the *Experiment Design* from Figure 2. The values obtained for each of the *Formaloutputparameter* values from Figure 2 are made available as *Actualoutputparameter* values in Figure 3. In addition, provenance information about the workflow execution has also been captured automatically. For example, the LSID [2] of each operation invoked from the workflow is made available as an *Operationtrace* value. Further browsing of the individual operation invocations indicates exactly what values were used as input and output parameters. Such provenance information can be useful in helping to explain and interpret the results of *in silico* experiments.

4 Contextualised Workflows: An Architectural Perspective

The core components that participate in the process of formulating and executing a workflow were introduced in Section 3.2. Figure 4 illustrates principal relationships between the components, where the Taverna workbench constitutes the presentation layer, and includes a number of GUI plug-ins to facilitate user interaction. The workflow enactor (Freefluo) is the central component of the execution environment, and communicates with other myGrid components via plug-ins that observe the events generated as the enactor's internal state changes. For example, when an intermediate step in the workflow completes its execution, the enactor generates an event and makes the intermediate results available to the event listeners. The MIR plug-in responds to this event by obtaining the

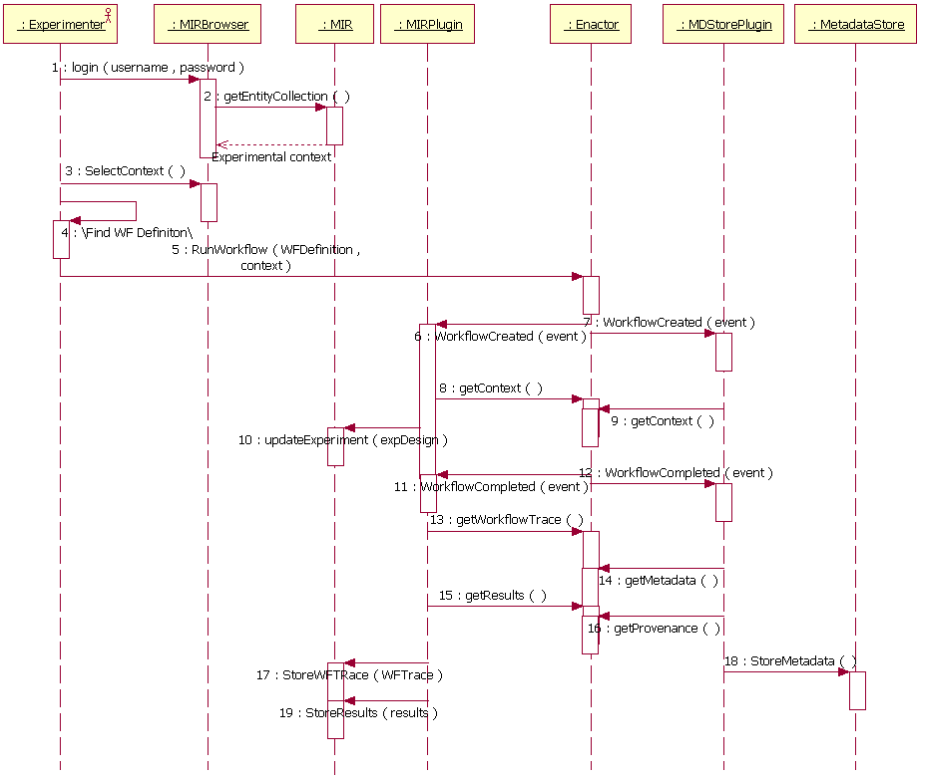

Fig. 5. Interactions between core myGrid components

intermediate results and storing them in the MIR in their appropriate context. As such, the plug-in architecture is instrumental in facilitating the automatic propagation of the experimental context across the participating components.

Figure 5 is a UML sequence diagram that illustrates a basic set of interactions between the myGrid components from in Figure 4, and provides a simplified view of the steps involved in contextualised execution of a workflow. A typical interaction for a contextualized workflow execution starts with user's login via the MIRBrowser. The next step is normally finding a workflow to execute. This could either be done by a simple load from the local file system, by a get operation from the MIR, or by a search query via the Feta GUI panel. Next, the experimenter executes the workflow. Although in the diagram this is shown as a direct interaction with the enactor for simplicity, in reality this is done via a separate GUI panel and the context is passed to the enactor implicitly. As the enactor executes the workflow, it informs event listeners (i.e. the MIR plug-in and the Metadata Store plug-in) that act as proxies on behalf of other myGrid components, at each critical event. Only two important events, namely *WorkflowCreated* and *WorkflowCompleted*, are shown in the diagram, although there

are several other intermediate events emitted by the enactor, for example for capturing provenance data on operation calls. The listeners respond to those events by extracting the context information and any other information they need from the enactor's state, thereby ensuring that the MIR and the Metadata Store receive the relevant provenance information.

An additional benefit of the automatic capturing of workflow results and provenance information is that the e-Scientist can pose complex queries against historical records. For example, a query could be expressed to *select all the workflows that were executed after date 30th March 2004, by the person 'Hannah Tipney', that had an output of type 'BLAST output'.*

5 Conclusions

This paper has described how a workflow definition and enactment environment can provide enhanced support for e-Science activities by closely associating workflows with their broader context. The particular benefits that have been obtained in myGrid derive from the principles introduced in Section 1, namely:

Consistent Representation: the paper has described how an e-Science specific, but application-independent, information model can be used not only to manage the data resources associated with a study, but also to drive interface components. In addition many myGrid components take and return values that conform to the information model, leading to more consistent interfaces and more efficient development.

Automatic Capture: the paper has described how the results of a workflow execution, plus associated provenance information, can be captured automatically, and made available throughout an e-Science infrastructure as events to which different components may subscribe. The properties of these events are generally modelled using the information model, and are used in the core myGrid services to support the updating of displays and the automatic storage of contextualised information.

Long-term Preservation: the paper has described how an information repository can be used to store the data artifacts of individual scientists in a consistent fashion, thereby supporting future interpretation, sharing and analysis of the data. Most current bioinformatics analyses are conducted in environments in which the user rather than the system has responsibility for recording precisely what tasks have taken place, and how specific derived values have been obtained.

Uniform Identification: the paper has described how a wide range of different kinds of data can provide useful context for the conducting of *in silico* experiments. Such information often has to be shared, or cross-referenced. In myGrid, LSIDs are used to identify the different kinds of data stored in the MIR, and LSIDs also enable cross-referencing between stores. For example, if an assertion is made about a workflow from an MIR in a Metadata Store, the Metadata Store will refer to the workflow by way of its LSID.

As such, the contribution of this paper has been both to demonstrate the benefits of contextualisation for workflow enactment, and also to describe the myGrid approach, both from a user and architectural perspective. The software described in this paper is available from *http://www.mygrid.org.uk*.

Acknowledgements. The work reported in this paper has been supported by the UK e-Science Programme.

References

1. S. AlSairafi et al. The design of Discovery Net: Towards Open Grid Services for Knowledge Discovery. *The International Journal of High Performance Computing Applications*, 17(3):297 – 315, Fall 2003.
2. T. Clark, S. Martin, and T. Liefeld. Globally distributed object identification for biological knowledgebases. *Briefings in Bioinformatics*, 5(1):59 – 70, 2004.
3. E. Deelman, J. Blythe, Y. Gil, C. Kesselman, G. Mehta, S. Patil, M.-H. Su, K. Vahi, and M. Livny. Pegasus : Mapping scientific workflows onto the grid. In I. Foster and C. Kesselman, editors, *2nd European Across Grids Conference*, 2004.
4. E. Gallopoulos, E. Houstis, and J. R. Rice. Computer as thinker/doer: Problem-solving environments for computational science. *IEEE Comput. Sci. Eng.*, 1(2):11–23, 1994.
5. C. Goble, C. Greenhalgh, S. Pettifer, and R. Stevens. Knowledge integration: In silico experiments in bioinformatics. In I. Foster and C. Kesselman, editors, *The Grid: Blueprint for a New Computing Infrastructure*, pages 121–134. Morgan Kaufmann, 2004.
6. E. N. Houstis and J. R. Rice. Future problem solving environments for computational science. *Math. Comput. Simul.*, 54(4-5):243–257, 2000.
7. T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Greenwood, T. Carver, M. R. Pocock, A. Wipat, and P. Li. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, page bth361, 2004.
8. K. Schuchardt, B. Didier, and G. Black. Ecce – a problem-solving environment's evolution toward grid services and a web architecture. *Concurrency and Computation: Practice and Experience*, 14(13 – 15):1221 – 1239, 2002.
9. N. Sharman, N. Alpdemir, J. Ferris, M. Greenwood, P. Li, and C. Wroe. The myGrid Information Model. In S. J. Cox, editor, *Proceedings of UK e-Science All Hands Meeting 2004*. EPSRC, September 2004.
10. R. D. Stevens, H. J. Tipney, C. J. Wroe, T. M. Oinn, M. Senger, P. W. Lord, C. A. Goble, A. Brass, and M. Tassabehji. Exploring Williams-Beuren syndrome using myGrid. *Bioinformatics*, 20(suppl_1):i303–310, 2004.
11. I. Taylor, M. Shields, and I. Wang. *Grid Resource Management*, chapter Resource Management of Triana P2P Services. Kluwer, June 2003.
12. D. W. Walker, M. Li, O. F. Rana, M. S. Shields, and Y. Huang. The software architecture of a distributed problem-solving environment. *Concurrency: Practice and Experience*, 12(15):1455–1480, 2000.