VTPortal: a Scientific Community Web Portal for Reusable VisTrails Workflows

Sérgio Manuel Serra da Cruz, Alexandre Ribeiro, Marta Mattoso

PESC – COPPE/Federal University of Rio de Janeiro (UFRJ)

{serra, alexandreribeiro, marta}@cos.ufrj.br;

Abstract. *Managing large scientific experiments is a complex research challenge due to the amount of scientific resources to be managed. Science portals are a way to simplify this task by aggregating data from different sources and by providing a set of pre-designed analyses. However, such portals are often built manually, and are not flexible enough to support sharing, reusing and executing workflows enacted by SWfMS like VisTrails. In this paper we describe VTPortal, a science portal that combines a set of tools and an infrastructure for providing a collaborative environment for scientists.*

1. Introduction

Large scale scientific experiments demand high performance computing (HPC) in environments such as clusters, grids, P2P or cloud environments, spanning organizational and spatiotemporal boundaries while managing huge amounts of data. There are several research and implementation initiatives in the use of scientific workflow management systems (SWfMS) for e-Science applications. SWfMS (Taylor *et al.,* 2007) are tools that assist the scientist in conducting studies to prove a hypothesis during a scientific process. Scientific workflows are a result of collaborative team efforts which require specialist expertise that is often complex and challenging to build (Goderis *et al*., 2005). The provenance, or the origin, or source of something (Moreau *et al*., 2007), can capture a wide range of information, including, for example, who or what generated the data, history of data stewardship, manner of manufacture, place and time of manufacture.

Currently, provenance support in the existing SWfMS is designed to capture the information related to isolated workflow´s run. However, the collaborative process of scientific experiments often involves multiple executions of workflows within a single experiment. For example, different or the same workflow definition is executed through a common infrastructure by changing input data or parameters. These set of executions requires the registering of annotations which are tightly connected with provenance data. Scientists do more than produce and consume data: they comment on it and refer to it, and to the results of queries upon it. The provenance and annotations are often critical to the understandability, reusability, and reproducibility of the scientific experiment (Mattoso *et al.,* 2010).

The goal of this work is to extend the current single-workflow and single-user targeted provenance approach to a number of workflow runs within a controlled environment such as a community portal for sharing data, programs and workflows. To achieve this goal we present *VTPortal*, a scientific community portal that allows users to add annotations, to share reusable workflows, to remotely submit previously defined workflows available on VisTrails, and to select a machine for this remote execution. *VTPortal* does not require modifications on VisTrails' code. To the best of our knowledge this is the first system that supports the collection

of annotations and combines the retrospective provenance generated during distinct executions of the workflows that belong to the same scientific experiment.

2. Related Work

The advances of IT technologies are encouraging people to form large-scale and multidisciplinary e-Science research projects to solve complex scientific problems in different fields of knowledge, such as: bioinformatics, engineering, astronomy. These projects demand intensive computation and data resources and the use of workflows; they are collaborative in nature and include multiple domain scientists with domain-specific expertise located at geographically distributed organizations. Collaborative scientific projects like: BioPauá, BioWEP , WHIP, Virolab, VL-e, CrowdLabs , myExperiment and CAMERA 2.0 have something in common, they use community portals and scientific workflows.

Community scientific portals provide access to advanced tools and databases that could be shared by a community of users via web, allowing them to interact with other colleagues, processes, documents and content in a personalized and role-based fashion through the web browser. Historically, community portals environments have been evolved into two groups. The first group is represented by workflow-oriented portal concept (like BioPauá, BioWep and WHIP) that enables the interoperability of various grids during the execution of workflows, but has no concerns about data provenance. BioWeP and WHIP projects aim to bring the desktop-based workflow and the Web closer and provide rich interfaces for Web applications. The second group is represented by collaborative-oriented portal concept (like VL-e, ViroLab and CAMERA 2.0), where scientists can publish, share and find other scientific workflows, such group takes provenance into account but with distinct approaches. For instance, in the VL-e project, the focus is on the execution of workflows developed in distinct SWfMS.

3. *VTPortal*

VTPortal is a novel tool and is different from the previous discussed ones. It aims at helping to run pre-existent workflows available on VisTrails system and storing provenance from a multiple runs of the workflow that composes one *in silico* experiment. *VTPortal* captures and maintains associations across runs and allows scientists to add annotations about each individual run, while taking advantage of existing provenance gathering mechanisms of VisTrails.

3.1 *VTPortal* Features

A Portal provides a layer of abstraction over a complex system (such as clusters, grids and clouds) to make its use streamlined while efficient. The core features of *VTPortal* are: (*i*) It is agnostic with respect to the application domain; (*ii*) it provides integration of provenance metadata from different executions of the workflows of a given scientific experiment; (*iii*) it takes into account the needs of scientists for accessing remote environments like the Web; (*iv*) offers a transparent and simple access to scientific workflows, despite their physical location. The system goes beyond simple Web accesses, it provides components that can be used to connect the VisTrails workflow engine with Web servers, allowing scientists to share workflows stored in a remote repository.

3.2 *VTPortal* Architecture

We have developed a multi-tier architecture which follows standard software development guidelines. Scientists interact with the system through a standard browser. The front-end tier is the topmost level of *VTPortal*, at the Web server implements the application rules and displays

the presentation, *i.e.,* it allows the configuration of scientific experiments, the registering of the reusable workflows and users, the upload of dataset and personal annotations.

The front-end offers *VTPortal* navigation features. The Web server hosts the core system, the system is composed by several Python and PHP modules. The main modules are the *WfParser*, the *WfPersistence* and the *WfMonitor*. The *WfParser* is XML parser that converts reusable workflows into a set of tuples at the database at the back-end tier. It is responsible to parse the workflow definition and the existing prospective provenance provided by VisTrails. The *WfPersistence* module encapsulates all data access to the relational database system. The *WfMonitor* encapsulate the calls from the Web server to the workflow enactor, *i.e.,* a SWfMS running at the logic-tier, the *WfMonitor* also displays workflow status information about completed and currently executing workflows at logic-tier. The monitor reads tracking information from the database and displays current workflow state at the user session. The logic-tier is pulled out from the front-end tier; it hosts the SWfMS Server, the SWfMS not only enact reusable workflows but also offer built-in provenance gathering facilities. The provenance metadata is stored at the back-end tier.

At the back-end tier we have the Database Server, it could be any relational database system, the database stores the prospective provenance parsed from the reusable workflows and the retrospective provenance collected from the experiment and their workflow invocations. This tier keeps data neutral and independent from application servers or workflow engine. Giving data its own tier also improves scalability and performance of *VTPortal*.

4 *VTPortal* in action

Scientists need to organize their workflows into a particular experiment. An example is a research project, which contains all the VisTrails workflows executions and the dataset they use together. This allows for defining different levels of visibility: Research groups can have discussions and upload reusable workflows and dataset that are only visible to the people involved in the experiment. The ability to selectively disclose information for people outside the group is extremely important for scientists, who may work for many years before deciding to release certain types of data. Another advantage is that a project can have its own dedicated servers (application and database). This creates the possibility of having specialized servers for different types of workflows. For example, let us suppose a neural network experiment (Figure 1), where a scientist developed a set of workflows, they can be uploaded to on a remote server machine and previously registered users on *VTPortal* can execute those workflows remotely.

Ideally, the reusable workflows that belongs to *VTPortal* would be executable anywhere. One of the key requirements to make this possible is to abstract away references of parameters, paths and input data as to allow other scientists to execute those workflows. In *VTPortal* there is a simple way of making this possible, the workflow can take advance of VisTrails aliases (defined at design-time). The aliases values are defined by the scientist at the run-time and incorporated by *WfParser* module the at the workflow´s run-time. Like aliases values, annotations (defined by scientists for each workflow run) are valuable sources of retrospective provenance metadata defined at run-time.

We believe that one of the most interesting applications of a system like *VTPortal* is the impact that it can have on the provenance of the scientific experiment. In *VTPortal*, we advocate a direct linkage to the provenance information of the workflows and the experiment. We provide mechanisms to store the provenance of the different executions of workflows of a scientific experiment in a single database. The provenance metadata (input dataset, parameters, workflow, host, user account and profile) are stored in the VisTrails database, so querying experiments provenance is simple. Figure 1 presents some *VTPortal* screenshots related to a

neural network clustering data experiment, where a scientist select a workflow (Kohonen workflow), previously deployed by the experiment administrator, he specifies the input dataset and the parameters from a list and, to prepare it, he need to press the "Prepare a workflow" button (Figure 1a). Thus, *VTPortal* merges variable data with the reusable workflow and submit it to VisTrails server. In the second screen (Figure 1b) the scientist can add annotations regarded to that workflow run, he can also choose the machine that hosts the SWfMS server and the version of the workflow that will be enacted. Finally, and the scientist can run the workflow by pressing the "Execute" button. All such data are stored at the provenance database and the result dataset stored at the server file system.

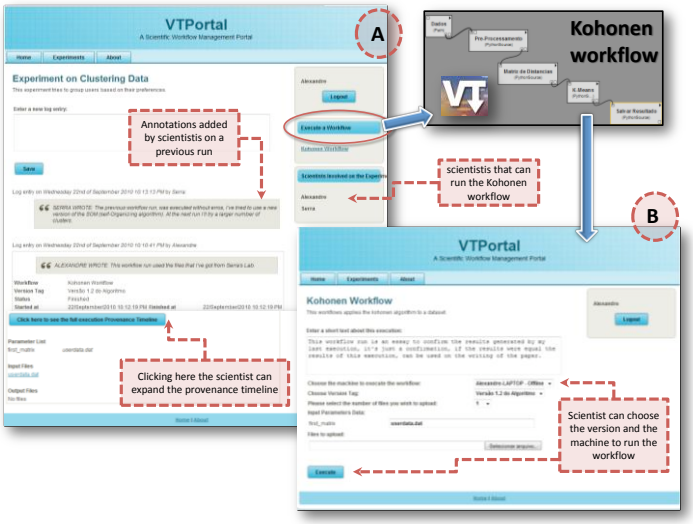

Figure 1. Configuring the workflows of the Clustering Data Experiment (1a). Setting annotation and workflows´ parameters (1b).

5. Conclusion and Future Work

Scientists involved in a scientific exploration are domain knowledge experts and should not spend their time in computational environment setups. In this paper we have introduced the *VTPortal*, a novel tool that delivers connectivity ("last mile") from remote experiments to scientists, offering a gateway that allows them to remotely browse, select, share, annotate, enact predefined reusable VisTrails workflows stored in remote machines and collect provenance from its executions. The system allows setting new parameters for pre-defined workflows at run time, querying experiments data and navigates on provenance timelines of one scientific experiment as a whole. The system was easily attached to VisTrails SWfMS and did not require modifications on it. The system suggests to the scientist the possibility to get a deeper comprehension of distinct executions of the workflows that describe the scientific protocol of an *in silico* experiment. For its most basic use, it does not require any installation of tools in the scientists´ machine, and we see this as an important advantage.

References

Goderis, A. et al., (2005) "Seven Bottlenecks to Workflow Reuse and Repurposing in The Semantic Web"In: ISWC 2005 pp. 323-337.

Mattoso, M., et al., (2010). "Towards supporting the life cycle of large scale scientific experiments" In: International Journal of Business Process Integration and Management.

Moreau, L. et al., (2007) "The Open Provenance Model", Electronics and Computer Science, University of Southampton,

Taylor, I. al., (2007) "Workflows for e-Science: Scientific Workflows for Grids". 1 ed. Springer.