PAPER PROPOSAL FOR OPEN REPOSITORIES 2010

http://or2010.fecyt.es

Dataset Lifecycle Management:

an integrative approach from scientific workflow composition and execution through project collaboration to sharing and archiving

Alex D. Wade[1], Dean Guo[1],
Simon Mercer[1], Oscar Naim[1], Michael Zyskowski[1]

Microsoft Research
awade@microsoft.com
diguo@microsoft.com
smercer@microsoft.com
onaim@microsoft.com
michaelz@microsoft.com

ABSTRACT

With a growing demand for transparency and openness around scientific research and an emphasis on the sharing of scientific workflows and datasets, there is a similarly increasing number in the variety of client and web-based tools required to manage each stage in the lifecycle of individual datasets. Datasets are produced from a variety of instruments and computations; are analyzed and manipulated; are stored and referenced within the context of a research project; and, ideally, are archived, stored, and shared with the rest of the world. Each of these efforts, however, requires a number of user actions involving a growing number of systems and interfaces. In an effort to preserve the flexibility and autonomy of the researchers, but also to minimize the logistical effort involved, we present in this paper a partial solution approach to this problem through the integration of workflow execution, project collaboration, project-based dataset management and versioning, and long-term archiving and dissemination. This example demonstrates the orchestration of a number of existing Microsoft Research projects;

however, the interaction between each uses existing web interoperability protocols and can easily support the replacement of individual architectural components with related services.

BODY

The Research Information Centre (RIC) is a virtual research environment framework being jointly developed by Microsoft External Research and the British Library (1). Research collaboration sites can be defined, deployed, and managed using the RIC Framework, with a goal of providing a unified web-based interface for research project team, integrating domain specific tools and services, real-time collaboration, and collaborative task, calendar, data, and document management.

Project Trident is a scientific workflow workbench, which provides the ability to visually compose workflows using a catalog of existing activities and complete workflows (2). The workflow workbench also offers a tiered library that hides the complexity of different workflow activities and services for ease of use. Project Trident is a rich client application built upon Windows Workflow Foundation, and authored workflows can easily be discovered by and shared with other researchers via direct publication to social networking sites such as myExperiment (3). Workflows are also published to a Project Trident server store, and workflows may execute on the Project Trident server, on a High Performance Computing (HPC) cluster, over cloud computing resources, and/or distributed across any number of remote web-services, depending on the computational, storage, and service requirements of the individual workflow, as determined by the workflow author.

Zentity (4) is a research-output repository platform that provides a suite of building blocks, tools, and services to create and maintain an organization's digital library ecosystem. Zentity offers a flexible data model that supports a wide variety of entity-types, formats, and metadata models.

Our solution provides a web-based view into composed scientific workflows within the context of a RIC project site. Any member of a research team, then, may use the general project collaboration space to browse and navigate the available workflows to which she has access. Further, that team member may also schedule and/or execute individual workflows from the context of the project collaboration space. If a specific workflow requires execution variables or other inputs, these may be inserted directly

from the web user-interface, and if local data files are required, these can easily be navigated and selected from the web UI.

Workflows are then executed on the Trident Server, on an HPC cluster, or in the cloud, and/or across multiple web services as determined by the workflow composer. Knowledge of this complexity however need not be surfaced to the individual researcher who is executing the workflow.

Upon workflow completion, the resulting output(s) such as datasets, images, etc., may be written back to the project collaboration space allowing the entire researcher team to easily view, download, or visualize the results from within the context of the project collaboration space. RSS or email notifications may also be sent to the researcher when workflows have completed. All workflow execution, provenance and performance information is maintained within the Project Trident database, and may be viewed from within the context of the project collaboration site.

Finally, any member of the research team may flag a designated dataset (or other workflow output) for longer-term archiving. This flagged status can then be used to facilitate harvesting by the intuitional repository or related services for duplication, management, and sharing.

WORKS CITED

1. **Microsoft Research.** Research Information Centre. *Microsoft Research.* [Online] 2009. [Cited: 02 28, 2010.] http://research.microsoft.com/ric/.

2. —. Project Trident: A Scientific Workflow Workbench. *Microsoft Research.* [Online] 2009. [Cited: 02 28, 2010.] http://research.microsoft.com/en-us/collaboration/tools/trident.aspx.

3. *myExperiment: social networking for workflow-using e-scientists.* **Goble, Carol Anne and De Roare, David Charles.** Monterey, CA : ACM, 2007. Proceedings of the 2nd workshop on Workflows in support of large-scale science . pp. 1-2.

4. **Microsoft Research.** Zentity: a research-output repository platform. *Microsoft Research.* [Online] 2009. [Cited: 02 28, 2010.] http://research.microsoft.com/zentity/.