Collaborative provenance for workflow-driven science and engineering

Altıntaş, İ.

Publication date
2011

4

Scientific Research and Collaboration Environments

> *"Innovation is fostered by information gathered from new connections; from insights gained by journeys into other disciplines or places; from active, collegial networks and fluid, open boundaries. Innovation arises from ongoing circles of exchange, where information is not just accumulated or stored, but created. Knowledge is generated anew from connections that werent there before."*
>
> – Margaret J. Wheatley, in her book Leadership and the New Science.

E-Science relies on collaborative efforts of inter-disciplinary research groups requiring substantial distributed infrastructure including access to networked computational resources, tools and digital data archives. We often see projects that facilitate construction and utilization of scientific infrastructure resulting in collaboratories where scientists conduct and participate in research and communicate with each other. The tasks scientist carry out through such collaboratories include: (i) publish their digital artifacts including data and analysis artifacts; (ii) build, run, share and repeat online scientific experiments; (iii) browse for related data and tools; and (iv) participate in focused collaborative activities.

Over the last decade, a number of e-Science platforms were built with similar goals. While some of these platforms focus on a particular scientific community or domain, e.g., Virolab (Malawski *et al.* 2010), VL-e (Zhao *et al.* 2006*c*) and CAMERA (Altintas *et al.* 2010*a*), others focus on enabling communication of scientists (Goble *et al.* 2010) and dissemination of e-Science tools to the general scientific community.

This chapter focuses on three different types of collaborative e-Science environments, namely, virtual laboratories, scientific portals, scientific social networking environments. Virtual laboratories and scientific portals that are explained as examples to single-community oriented platforms, where as the social networking environments have a more broad user base. These categories were chosen as examples of collaborative platforms where sharing and executing workflows, tracking their provenance and analyzing user collaborations through the

collected provenance information would add value. We review some of these projects as examples and a few recent infrastructures for sharing scientific workflows.

4.1 Virtual Laboratories

4.1.1 Virolab

In the ViroLab[1] virtual laboratory (Sloot *et al.* 2009) supports virologists, epidemiologists and clinicians investigating the HIV virus and the possibilities of treating HIV-positive patients. Although the ViroLab Virtual Laboratory is being built specifically for this domain of science, the conceptual solutions and the technology developed could be reused for other domains. The system is built as a set of integrated components that, used together, form a distributed and collaborative space for science. Multiple, geographically-dispersed laboratories and institutes use the virtual laboratory to plan, and perform experiments as well as share their results.

In Virolab, scientific applications are executed as scripts, which invoke distributed services wrapped as so-called Grid Objects. Provenance is recorded by collecting events emitted by the GridSpace engine that executes the experiment scripts (Balis *et al.* 2008). Events are next aggregated, translated to an ontology-based provenance model, and stored in a repository. Ontologies enable capturing of semantics of data and computations used in an experiment, and visual querying over provenance records (Balis *et al.* 2009). Provenance of combined scripts can be captured by explicit reuse of results from previous experiments. For this purpose, the virtual laboratory assigns unique identifiers to experiment results and maintains a repository in which they are saved. The user can browse the result repository and reuse a previous result as input for a next experiment. In this way, unique identifiers of data products are preserved between independent executions, which is regarded as a data dependency.

4.1.2 The Virtual Laboratory for e-Science

The Virtual Laboratory for e-Science (VL-e)[2] project was built to decrease the gap between the technology push of the high performance networking and the Grid and the application pull of a wide range of scientific experimental applications. It provided functionalities for e-Science application environments with as experimental infrastructure for the evaluation of the ideas.

In the context of the VL-e project, the WFBus (Zhao *et al.* 2006*c*) focuses on the execution of workflows developed in various workflow management systems. A number of issues in workflow management aiming at covering the entire lifecycle of collaborative e-Science experiments have been addressed: experiment design, mapping, execution, and sharing results. In e-Science experiments, workflows encode the logic of the experiment processes and

[1] Virolab website: `http://virolab.org/`
[2] VL-e website: `http://www.vl-e.nl/`

become an important resource to promote knowledge transfer among scientists. The aim of this work was not only on the modeling and composition aspects of workflows, but also on the validation and reproducibility of results. In VL-e, meta-workflows (VL-e WFBus) (Zhao *et al.* 2006*c*) and execution environments for scheduling and runtime control (WS-VLAM) (Korkhov *et al.* 2007) have been developed. A bus-like architecture was designed to aggregate workflows from different systems. The VL-e WFBus provides interfaces to wrap and integrate legacy scientific workflows. It also provides tools to recognize different workflow descriptions stored in the system, and describe the Meta information according to the predefined schema. While the VL-e WFBus is concentrating on the execution of workflows developed in various workflow management systems, the notion of the collaborative provenance covers both the execution and provenance aspects of an aggregate workflow.

4.2 Scientific Portals

In the context of CAMERA[3] and GEON[4] projects, scientific workflows run through a GridSphere (Novotny *et al.* 2004) portal environment. The scientific products used and produced by the workflow are stored in data repositories that are also accessible through the project portal or the user's local computer. The provenance of workflow execution is stored in a data archive through the workflow execution portlet.

4.2.1 Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis

Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA)(Sun *et al.* 2010) is a Cyberinfrastructure project to enable the microbial ecology community to manage the challenges of metagenomics (Committee on Metagenomics: Challenges and Functional Applications 2007) analysis. Now with a new component-based infrastructure, CAMERA version 2.0 (CAMERA 2.0) (Altintas *et al.* 2010*a*) is available to support semantically-aware data acquisition and access. In CAMERA , scientific workflows enable the use of various community tools in daily science by a user community of more than 3000 metagenomics researchers in more than 75 countries. CAMERA also provides scientific workflows that allow users to invoke integrated annotation of genomic datasets and use a wide range of analysis tools, as well as search and sort information based on metadata (data about data).

The layered and modular software architecture of CAMERA is illustrated in Figure 4.1. This architecture is designed to serve two purposes: (1) to provide an adequate separation of concerns (SoC) for different system components including data interfaces; and (2) to allow external scientific developers to create workflows that can fully utilize software and hardware resources of CAMERA.

[3]CAMERA website: `http://camera.calit2.net/`
[4]GEON website: `http://geongrid.org/`

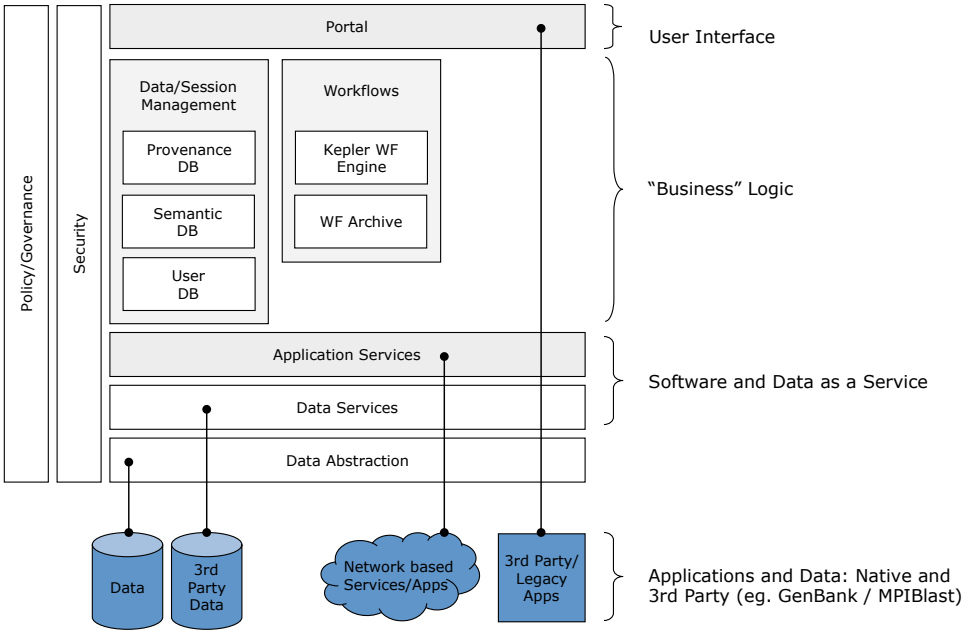
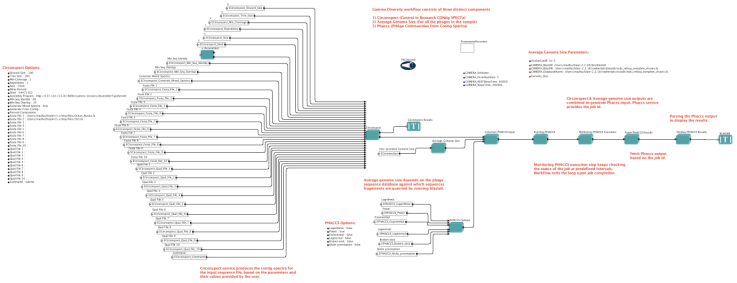


Figure 4.1: CAMERA 2.0 Architecture provides an adequate separation of concerns and allows external scientific developers to create workflows.

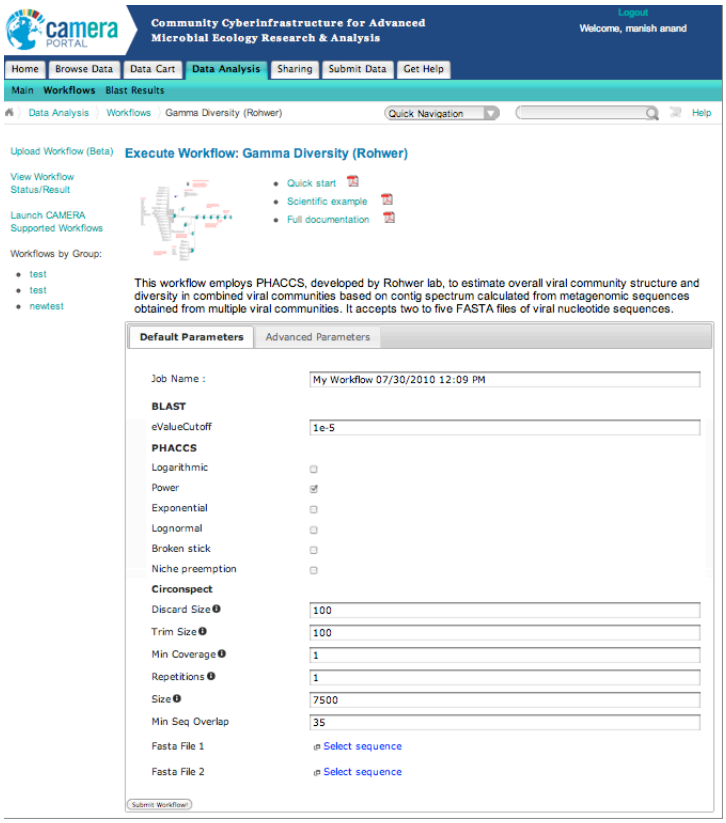
The elements and techniques readily incorporated into the architecture of CAMERA include a user interface that facilitates and enhances the process of collaborative scientific discovery for domain scientists. This end-user interface blends both web-based and traditional desktop application environments. Primary user interaction is provided via a centralized open-source Web Portal interface based on GridSphere standards compliant portal framework. Under the portal layer are the data and workflow management components to assist with assembly of components into useful and more complex scientific discovery tools.

A typical CAMERA workflow for Gamma diversity analysis is shown in Figure 4.2(a). This workflow predicts overall biodiversity of two or more communities by building models of possible community structure using a modified Lander-Waterman algorithm to predict the underlying contig spectrum, shotgun (or 454) DNA sequences obtained from environmental samples. The input to the workflow is one or more viral metagenomic datasets in FASTA format. The Gamma diversity workflow integrates Circonspect (calculating contig spectrum), Average Genome Size (BLAST-based calculation) and PHACCS (Angly *et al.* 2005) (modeling structure of diversity) tools. Users can modify default parameters for it through the portal interface shown in Figure 4.2(b).

The current infrastructure of CAMERA employs the Kepler scientific workflow system (Ludäscher *et al.* 2006) and is also built for extensions to accept workflows from other workflow systems. Kepler assists scientists to edit, manage and execute scientific workflows



(a)



(b)

Figure 4.2: (a) The CAMERA Gamma Diversity Workflow; and (b) The portal interface to run the CAMERA Gamma Diversity workflow.

through an graphical user interface and an execution engine. The ability to separate Keplers execution engine from its user interface (Vergil) enables the execution of existing work-

flows in batch mode. Actors (computational units or processors) are dragged and dropped onto Kepler canvas, where they can be customized, linked and executed. Customized actors, workflows and provenance of their runs can be easily exported through built-in wizard tools. The sharing happens locally or publicly through Kepler actor repository, and facilitates reuse.

In this new release of CAMERA, Kepler supports the interaction of automated computational tools and human inspection and interaction. Kepler provides capabilities to record the processing history of workflow execution, i.e., provenance. CAMERA enables users to create and retrieve the provenance of workflows specific to their own experiments. The enhanced query capabilities allow users to access data via processing tools hosted by CAMERA. A researcher may thus combine local data and information from outside databases with CAMERA-hosted services and processing tools supported by other groups. With these significant enhancements, the new CAMERA Cyberinfrastructure is more useful, flexible, scalable and sustainable. In addition, CAMERA utilizes project-dedicated, area-dedicated and very large, multi-community shared resources. These resources span computing, storage and visualization tasks.

4.2.2 The Geosciences Network

The Geosciences Network (GEON)[5] was funded by NSF to facilitate collaborative, interdisciplinary science efforts. GEON developed an infrastructure that supports advanced semantic-based discovery and integration of data and tools via portals (the GEON portal (Youn *et al.* 2007)), to provide a unified authenticated access to a wide range of resources to conduct comprehensive analysis using emerging web and grid-based technologies to enable the next generation of science and education.

One of the challenging problems GEON focused on is distribution, interpolation and analysis of LiDAR (Light Distance And Ranging) (Carter *et al.* 2001) point cloud datasets. The high point density of LiDAR datasets pushes the computational limits of typical data distribution and processing systems and makes grid interpolation difficult for most geoscience users who lack computing and software resources necessary to handle these massive data volumes. The geoinformatics approach to LiDAR data processing requires access to distributed heterogeneous resources for data partitioning, analyzing and visualizing all through a single interactive environment. A three tier architecture (Jaeger-Frank *et al.* 2006*a*) was developed in GEON that utilizes the GEON portal as a front end user interface, the Kepler workflow system as a comprehensive environment for coordinating distributed resources using emerging Grid technologies, and the Grid infrastructure, to provide efficient and reliable LiDAR data analysis. This framework was one of the first portal-based infrastructures to utilize a scientific workflow engine as a middleware behind a portal environment for coordinating distributed Grid resources and their provenance.

[5] GEON website: `http://geongrid.org/`

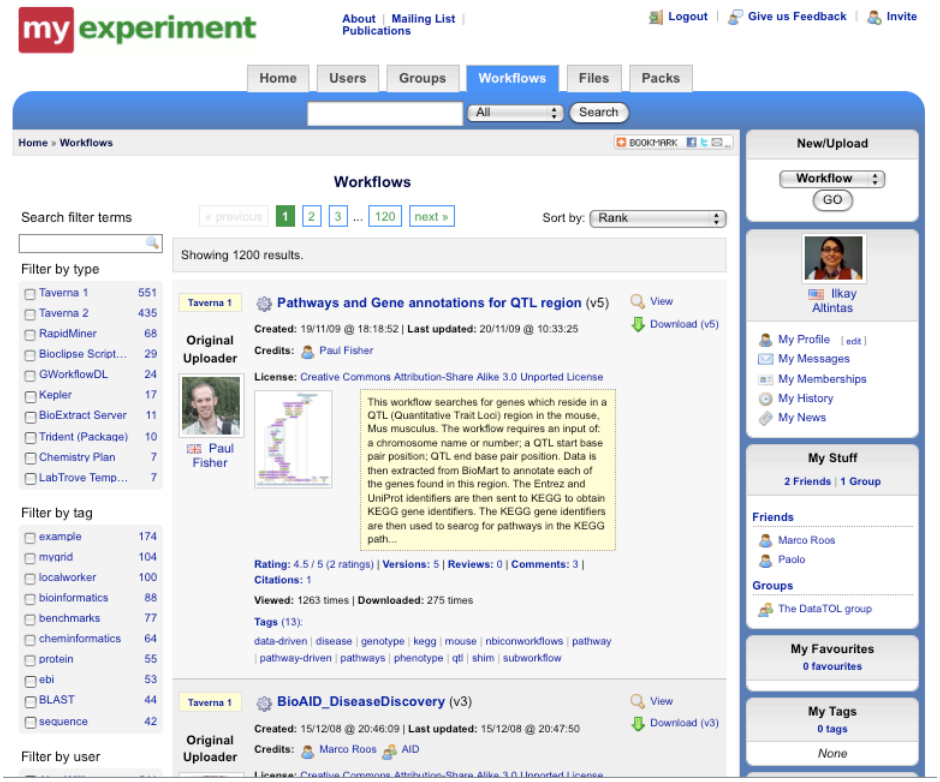


Figure 4.3: myExperiment enables researchers build social networks and share workflows.

4.3 Social Networking and Sharing Environments

4.3.1 myExperiment

myExperiment[6] (Goble *et al.* 2010) is an online research environment that supports the social sharing of scientific workflows. Since its release in 2007, myExperiment currently has over 3500 registered users and contains more than 1000 workflows, demonstrating the support from the scientific community for the notion of such collaborations. The myExperiment repository provides public and private user spaces, allowing for discovery, reuse and repurposing of scientific workflows. Through this repository, scientific workflow developers share their workflows in a secure manner and participate in virtual communities bound together by a common interest or research project. The workflows are rated by these communities through a feedback and credit mechanism. Figure 4.3 shows a typical user space in myExperiment showing available workflows when workflow tab is enabled.

An interesting opportunity for collaborative provenance management arises from the support for Scientific Research Objects (SROs) by myExperiment. Currently, myExperiment im-

[6] myExperiment: `http://www.myexperiment.org`

plements this concept through packs in which workflow references can be combined, but also any URL referring to data that is associated with the experiment, including other packs. For the collection of so-called Scientific Discourse data, myExperiment follows RDF-encoded models emerging from the Semantic Web community, such as those explored by the Semantic Web Applications in Neuromedicine (SWAN) project (Gao *et al.* 2006). An RDF interface allows users to access experiment-associated metadata, including the provenance of combined workflows. Applications such as portal environments can deploy the content of SROs in new ways. myExperiment already provides a SPARQL endpoint to its current content (http://rdf.myexperiment.org). An obvious application would be where the aforementioned workflow bus makes use of the workflows combined in SROs on myExperiment.

4.3.2 crowdLabs

crowdLabs[7] is a social visualization repository that is built on the social Web paradigm. It is primarily built for sharing visualization pipelines built using the Vistrails (Callahan *et al.* 2006) scientific workflow management system. It provides an infrastructure to for scientists to collaboratively analyze and visualize data.

crowdLabs is built to support the needs of computational scientists, including the ability to access high-performance computers and manipulate large volumes of data, while fostering collaborations. In addition, it allows publishing data and workflows and use of analysis pipelines, providing a potentially ideal platform for collaborative provenance usecases described throughout this thesis.

Summary

Examples to e-Science frameworks that utilize scientific workflows and their provenance were presented. These infrastructures are examples to where collaborative provenance notion presented in the paper could be used as a plug in to analyze user collaborations, impact of the developed infrastructure and its links to scientific results. Next, we will describe collaborative provenance along with a data model that can be deployed on top of these infrastructures.

[7] crowdLabs: `http://www.crowdlabs.org/`