Collaborative provenance for workflow-driven science and engineering

Altıntaş, İ.

Publication date
2011

1

Background and Problem Formulation

> *"If science is an unfinished project, the next stage will be about reconnecting and integrating the rigor of scientific method with the richness of direct experience to produce a science that will serve to connect us to one another, ourselves and the world."*
>
> – Presence: An Exploration of Profound Change in People, Organizations, and Society, Pg. 212 by Peter M. Senge, C. Otto Scharmer, Joseph Jaworski, and Betty Sue Flowers

1.1 Scientific Method and The Influence of Technology

Science is Collaborative. Science is the systematic knowledge of the physical or material world gained through observation, identification, description, experimental investigation, and theoretical explanation of phenomena. Using the basic scientific method (Carey 2003, Wilson Jr. 1991), scientists still start with a set of questions then observe phenomena, gather data, develop hypotheses, perform tests, negate or modify hypotheses, reiterate the process with various data, and finally come up with a new set of questions, theories, or laws. A scientific study is a set of such activities applied to accumulate scientific knowledge on an object of inquiry or a specific scientific domain. For centuries, scientific studies have been often carried out by teams building upon the existing breadth of knowledge. Scientific articles are mostly co-written by a number of scientists working together reporting on a scientific discovery as a product of a scientific collaboration. However, the scientific method evolved in time by the influence of technological advances, especially recently (Djorgovski 2005). Today, scientists need to collaborate more than ever. Due to the increasing number and sophistication of data acquisition technologies, the amount of raw data acquired has vastly increased over the last couple of decades (Berman 2008). This explosion of scientific data and knowledge along with the increase in the number of studies that require access to knowledge from multiple scientific disciplines amplify the complexity of scientific problems, often requiring

large teams to work together. In order to answer these grand challenge scientific questions, scientists use computational, data and collaborative technologies that are evolving almost daily. The requirements for these technologies share a common goal to enable collaborative studies serving one or more scientific themes or domains. These requirements include:

- **Computational Experimental Infrastructure:** Users should be able to launch computations, e.g., scientific simulation and visualization codes.
- **Data Sharing, Publication and Preservation:** Users should be able to publish, retrieve and transform data.
- **Common User Interface:** Users should be able to interact with other users, access and publish data, use other middleware through a unified interface.

Collaborative e-Science Technologies. Community portals (Altintas *et al.* 2010*a*), virtual laboratories (Zhao *et al.* 2006*c*), and Web2.0-based social networking and sharing environments (Roure *et al.* 2007) are popular technologies and platforms that emerged as a response to the above-mentioned collaborative requirements of science. These environments establish a common infrastructure where community members may access and contribute to data, middleware and computational tools, launch computations and projects through their user spaces under generic governance rules, keep a personal account, become part of an interest group, and even blog their findings. Computations use data from external data resources and the scientific outputs are saved in data repositories, optionally along with intermediate results and information on how results were captured. Computational tools could be executed multiple times by one or more scientists, potentially from an end-user interface at times requiring repeatable aggregates of multiple tools.

Scientific Workflows. The gradual shift from manual process execution to automation of repeatable patterns and the need to use a variety of technologies resulted in the creation of scientific workflow systems (Ludäscher and Goble 2005, Taylor *et al.* 2007, Gil *et al.* 2007, Deelman *et al.* 2009). Workflow systems are useful for the way computational scientists conduct studies by making technological advances more approachable through integrative interfaces and abstractions for underlying computing and data resources. Since the initiation of the scientific workflows in the late 1990s and the initial scientific workflow applications for solving visualization challenges, there have been significant developments in technology. Scientific workflows evolved to satisfy different scientific and computational requirements, technologies and visions that transform the scientific method. Moreover, as we learn more about efficient ways to design and execute them, they matured from art to commodity, enabling and impacting scientific studies with a number of advantages including provenance support.

Provenance. Provenance tracking (Simmhan *et al.* 2005) is a very important feature of scientific workflow systems as it helps to track the origin of scientific end products, validate and repeat experimental processes that were used to derive these scientific products. Scientific workflows are repeatable patterns of computational activities, typically designed iteratively

by a user and run multiple times by one or more users. Thus, information on data collection, data usage, and, especially, the computational outcome of a scientific workflow provides a rich source for conducting similar future scientific studies (Freire *et al.* 2008, Davidson *et al.* 2007, Bowers *et al.* 2008*b*). Scientific workflow provenance collection starts with workflow design and execution and the collected information must have the ability to create and maintain associations between workflow inputs, workflow outputs, workflow definitions, and intermediate data products. The provenance of a data product contains information about how the product was derived, and is crucial for enabling scientists to easily understand, reproduce, and verify scientific results. However, this is still only a partial solution to the modern scientific process that relies on multi-disciplinary collaborative teams working on different parts of scientific studies.

1.2 The Need for Collaboration

Currently, most provenance models are designed to capture the provenance related to a single run, and mostly executed by a single user. On the other hand, the computational scientific process often involves design and execution of multiple workflows (Bowers *et al.* 2007) where different members of a team make their scientific workflows available through a common infrastructure. A scientific discovery is the result of methodical execution of many of these scientific workflows with many datasets at different times by one or more users. Through collaborative platforms, workflows could be executed multiple times by one or more scientists, potentially from an end-user interface that combines several workflows such as the interface provided by Utopia (Pettifer *et al.* 2007). In addition, the executed workflows use data from external data resources and the scientific outputs are saved in data repositories. The intermediate results can optionally be saved in data archives along with the process provenance.

A typical set of components for such an infrastructure is illustrated in Figure 1.1, where the system components are implemented behind a common user interface, e.g., a portal, and conform to a common security and governance model. The *User Spaces* are where the user actions are performed in the system. The user actions include search for data and workflows, upload of data and workflows, publication of data, workflows and workflow run products, monitoring of active runs, review of completed runs, and exploring provenance graphs for past studies. Through the *Project Spaces*, users can share their work with a set of other users and conduct joint research without making the information on the artifacts like workflows, data and runs they use publicly available. Such project spaces are generally logical to enable group-based collaboration and are governed by the system through an authentication system. Users share scientific data, metadata and provenance data through common data stores depicted by *Shared Data Store* and *Provenance Store* in Figure 1.1. Workflow definitions are shared through potentially multiple *Workflow Repositories* that are linked to the infrastructure and ran through *Workflow Engines* that the system allows. In Chapter 4, we will review some of the existing collaborative systems that are similar to this architectural model.

The requirement for workflow sharing and execution interoperability is further supported

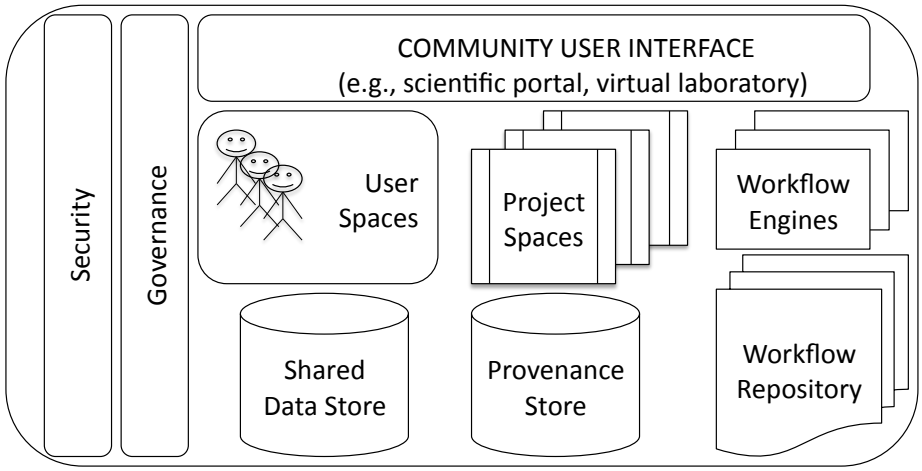

Figure 1.1: Typical components in a collaborative e-Science project (Not shown are the information flow, external data, service and computational infrastructure.)

with the introduction of environments like myExperiment (Roure *et al.* 2007) where workflows developed in many workflow management systems can be uploaded to the same Web2.0 site and the concept of Scientific Research Objects (called packs on myExperiment) (Goble *et al.* 2010) that allow users to publish a collection of references to the workflows and additional resources. The uploaded workflows become available to scientists who might be interested in running a number of existing published workflows to create complex scientific experiments, increasing the sharing and re-usability across the scientific community. To promote and facilitate exchange of information between multiple workflow systems supporting provenance, the Open Provenance Model (OPM) (Moreau *et al.* 2010) has been proposed by the scientific workflow community.

1.3 Problem Statement and Research Contributions

1.3.1 Problem Definition

This thesis addresses the following research question:

"How can the provenance information collected for execution of workflows using shared data and workflows in a collaborative e-Science environment be used to analyze the nature and strength of implicit connections between users, potentially leading to an analysis of user networks based on system observables for data publishing, workflow publishing and workflow runs?"

Starting with this question, this thesis focuses on *collaborative provenance*, collaborative views over provenance data accumulated in a collaborative environment driven by publication and execution of scientific workflows. As shown in Figure 1.1, we assume an environment that provides system components to establish user spaces, project spaces, unified access to

multiple workflow engines, shared data and provenance stores, and shared workflow repositories. We also assume that the actions performed by users get recorded as an added knowledge within the system, e.g., information on workflow sharing, workflow execution, data publish, and workflow run provenance sharing. As a result of the user actions and interactions with the system, a collaborative knowledge accumulates. This knowledge can be analyzed to infer nature and strength of *implicit* user collaborations based on system observables for data publishing, workflow publishing and workflow runs.

Based on this question, assumptions and goals, the next subsection describes the actual contributions of the thesis.

1.3.2 Contributions

In this thesis, we define a new collaborative provenance model which addresses the need for inferring dependencies across multiple workflow runs and analyzing user collaborations. Our proposed collaborative provenance model allows to establish the attributes for the nature of user collaborations, the strength of collaborations and self collaborations. We also provide some example collaborative provenance scenarios and collaborative provenance queries.

After defining collaborative provenance and understanding its potential usage, we describe a data model to capture and query collaborative provenance. This model supports collaborative provenance attributes for determining the *nature* (or type) and *strength* (or weight) of collaboration between multiple users and analysis of a researchers independent work (i.e., their "*self* collaborations"). We show how our data model is effective for answering both standard provenance queries as well as queries over the collaborative provenance attributes for determining the nature of collaborations, their strength, and for finding self relationships.

Using the defined data model, we investigate the implicit user collaborations in a Query Language for Provenance (QLP)-based (Anand *et al.* 2009*c*) query model that maps to the Open Provenance Model (OPM) (Moreau *et al.* 2010) using observables in an e-Science infrastructure (as seen in Figure 1.1) and for generating views on top of them. This approach links OPM graphs for workflow runs that have an input or output data dependency and helps to answer queries such as identifying data connections between workflow runs and contributions of users collaborating on a project based on the records of past executions. We adopt and extend a high-level query language for provenance called QLP, to express complex collaborative provenance queries. We also establish a mapping between the presented collaborative provenance model and OPM.

Furthermore, through example and real life application scenarios, we demonstrate the feasibility of how our approach will lead to development of systems that increase *interoperability* and *reusability* of workflow results by integrating provenance coming out of *different* workflow systems and, in turn, enhancing efficiency in modern collaborative research. These examples include the Provenance Challenges[1] 1 and 3 usecase scenarios, a real-world bioinformatics usecase scenario from the CAMERA (Sun *et al.* 2010) project, and a drug ranking

[1] Provenance Challenges website: http://twiki.ipaw.info/bin/view/Challenge/

usecase scenario from the ViroLab (Sloot *et al.* 2009) project.

In addition, we discuss the optimization challenges in modeling and querying interoperable collaborative provenance, describe our initial assessment of how the collaborative concepts apply to social networking analysis and provide a roadmap for future work.

1.3.3 Research Roadmap

The results of the thesis were described in the following publications:

- **Definition of Collaborative Provenance and Its Attributes:** *Collaborative Provenance for Workflow-Driven Science - A Position Paper* (Altintas *et al.* 2010*e*)
- **Description of a Relational Data Model for Collaborative Provenance:** *A Data Model for Analyzing User Collaborations in Workflow-Driven eScience* (Altintas *et al.* 2010*f*)
- **High-Level Querying Constructs for Interoperable Collaborative Provenance**: *Understanding Collaborative Studies Through Interoperable Workflow Provenance* (Altintas *et al.* 2010*d*)

Additionally, the applications of collaborative provenance in the context of CAMERA project were published in:

- *CAMERA 2.0: A Data-Centric Metagenomics Community Infrastructure Driven by Scientific Workflows* (Altintas *et al.* 2010*a*)
- *Extending the Data Model for Data-Centric Metagenomics Analysis using Scientific Workflows in CAMERA* (Altintas *et al.* 2010*c*)

This thesis also builds on the author's work[2] on scientific workflows, provenance and e-Science infrastructures. In particular, Altintas has been involved in the research and development of the Kepler scientific workflow environment since its early days and worked on challenges related scientific workflow design and execution, e.g., distributed execution. Altintas has designed and lead the development of the Kepler provenance framework, participated in the Provenance Challenges that led to the improvements of the framework for OPM, and worked with various scientific communities, e.g., fusion and metagenomics, to use the data collected by the provenance recorder. Altintas also worked as a workflow lead within a number of eScience infrastructure projects that inspired her point of view on collaborative workflow-driven science that is presented as a part of this thesis. Some of these studies are described in the following publications:

- *Accelerating the scientific exploration process with scientific workflows* (Altintas *et al.* 2006*b*)

[2] A full list of publications by the author is listed on pages 147 through 152.

- *Scientific Workflow Management and the Kepler System* (Ludäscher *et al.* 2006)
- *Kepler: An Extensible System for Design and Execution of Scientific Workflows* (Altintas *et al.* 2004*b*)
- *A Modeling and Execution Environment for Distributed Scientific Workflows* (Altintas *et al.* 2003)
- *Lifecycle of Scientific Workflows and their Provenance: A Usage Perspective* (Altintas 2008)
- *A Provenance-Based Fault Tolerance Mechanism for Scientific Workflows* (Crawl and Altintas 2008)
- *From computation models to models of provenance: the RWS approach* (Ludäscher *et al.* 2008)
- *Provenance Collection Support in the Kepler Scientific Workflow System* (Altintas *et al.* 2006*a*)
- *From Molecule to Man: Decision Support in Individualized E-Health* (Sloot *et al.* November 2006),
- *A Three Tier Architecture Applied to LiDAR Processing and Monitoring* (Jaeger-Frank *et al.* 2006*a*)
- *Linking Multiple Workflow Provenance Traces for Interoperable Collaborative Science* (Missier *et al.* 2010*a*)

We will describe the parts of the above-mentioned research throughout the rest of this book, and conclude with plans for future directions.

1.4 Overview of the Thesis

In the area of workflow-driven science and engineering, this thesis presents four main contributions for modeling and querying collaborative provenance. Firstly, it presents an overview of the effects of the scientic workflows on how scientic research is conducted with a focus on provenance tracking as a specific advantage of scientific workflows. Secondly, it provides a definition for collaborative provenance for inferring dependencies across multiple workflow runs and understanding user collaborations based on scientific workflow runs within an online community platform. Thirdly, it describes new data and query model that capture these implicit user collaborations, shows how this model maps to OPM and helps to answer collaborative queries, e.g., identifying combined workflows and contributions of users collaborating on a project based on the records of previous workflow executions. The adoption of and extensions to the high-level Query Language for Provenance (QLP) with additional

constructs allows non-expert users to express collaborative provenance queries against this model easily and concisely. Finally, the thesis provides a data model that is effective to capture collaborative provenance scenarios. We show how this data model can be used to answer collaborative provenance views queries, e.g., identifying combined data, workflow executions, and contributions of users collaborating on a project based on the records of previous workflow executions.

As a conclusion, the main collaborative provenance contributions within this thesis lead to development of systems that increase interoperability and reusability of workflow results, enhancing efficiency in modern collaborative research. This is also demonstrated through scientic usecase scenarios for establishing and understanding collaborative studies through interoperable workflow provenance. Specifically, the Provenance Challenge 1 and 3 workflows and various community workflows from the CAMERA project are adopted as collaborative and interoperable usecases, where different stages of the workflow are executed as different scientific workflows, potentially also in different scientific workflow management environments.

First chapter has been used to give a brief overview on the developments in the research areas of scientific workflows and provenance for e-Science, with the aim to formulate the contribution and problem formulation of this thesis.

Chapters 2 through 4 are used to present some of the concepts related to scientific workflows, provenance and collaborative e-Science infrastructures needed in this thesis. Based on this overview and using the concepts presented, Chapters 5 through 9 are devoted to the presentation of the main framework of this thesis along with application usecases that demonstrate the contributions. A formal definition of collaborative provenance and the contributions of the thesis on modeling and querying collaborative provenance are also explained. In particular, Chapter 5 is concerned with the definition of the notion of collaborative provenance. Subsequently, Chapter 6 discusses a data model and a querying approach for the defined collaborative provenance. Chapter 7 is reserved for the practical application of the collaborative provenance data modeling and querying approach presented. Chapter 8 explains the implementation of the collaborative provenance schema in Postgres and provides an evaluation of runtime effects of collaborative provenance queries. Chapter 9 is dedicated to the discussion of the interoperability challenges for collaborative provenance coming out of different scientific workflow environments and our approach for tackling these challenges.

Finally, Chapter 10 is used to end this thesis and contains the main conclusions and additional remarks, pointing to the possibilities of ensuing future research.