# Collaborative provenance for workflow-driven science and engineering

Altıntaş, İ.

**Publication date**
2011

# Summary

## Collaborative Provenance for Workflow-Driven Science and Engineering

İlkay Altıntaş, January 2011

As the scientific knowledge grows and the number of studies that require access to knowledge from multiple scientific disciplines increases, the complexity of systematic scientific research is significantly intensified. In order to answer "grand challenge" scientific questions, scientists use computational methods that are evolving almost daily. The basic scientific method, however, remains the same, but is continuously being transformed with the advances in computer science and technology in the last few decades. These changes in computer science and technology resulted in a set of middleware tools and systems specifically developed for making the automation of scientific process more efficient and faster. These tools and systems aim at helping scientists to simplify the design and execution of subsequent and coherent computing tasks, now called "scientific workflows".

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 computing technology. Scientific workflows evolved to satisfy different scientific requirements, computational technologies and scientific approaches that have transformed the scientific method into a computationally heavy process. Moreover, as scientists learn more about efficient ways to design and execute scientific workflows, it becomes imperative to keep track of how and when specific scientific information has been obtained within the computational scientific process, i.e., to track provenance information.

Provenance tracking is a very important feature of scientific workflow systems as it helps track the origin of scientific end products, validate and repeat experimental processes that were used to derive these scientific products. 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. Currently, most provenance models are designed to capture the provenance related to a single run, and mostly executed by a single user. However, a scientific discovery is often the result of methodical execution of many

scientific workflows with many datasets produced at different times by one or more users. To promote and facilitate exchange of information between multiple workflow systems supporting provenance, the Open Provenance Model (OPM) has been proposed by the scientific workflow community. Standards like OPM open the possibility of linking provenance information for scientific workflow executions performed in different systems. This ability to link workflow executions leads to a notion of implicit collaboration between the users who designed and executed those workflows.

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 scientific workflows on how scientific 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 a new query model that captures 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 in capturing collaborative provenance scenarios, and shows 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 scientific usecase scenarios for establishing and understanding collaborative studies through interoperable workflow provenance in this thesis. Specifically, the Virolab scenario, 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 workflows, potentially also in different workflows environments.