Collaborative provenance for workflow-driven science and engineering

Altıntaş, İ.

Publication date
2011

10

Conclusions and Future Directions

> *"The future is not a result of choices among alternative paths offered by the present, but a place that is created–created first in the mind and will, created next in activity. The future is not some place we are going to, but one we are creating. The paths are not to be found, but made, and the activity of making them, changes both the maker and the destination."*
>
> – John Schaar

10.1 Summary of Contributions

The ideas in this thesis started with the following 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, a new collaborative provenance model which addresses the need for inferring dependencies across multiple workflow runs and understanding user collaborations was introduced. How the proposed collaborative provenance model allows to establish the attributes for the nature of user collaborations, the strength of collaborations and self collaborations was explained. A new collaborative provenance model which addresses the need for inferring dependencies across multiple workflow runs and understanding user collaborations was defined. A relational collaborative provenance data model to capture and query collaborative provenance was presented. The presented 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").

Through the adoption of and extensions to a high-level query language for provenance called QLP, complex collaborative provenance queries were expressed. Using the defined relational data model, the implicit user collaborations in a QLP-based query model using the workflow, run and data dependencies in an e-Science infrastructure and for generating views on top of them was investigated. The results showed how our data model can answer 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. A mapping between the presented collaborative provenance data model and OPM was established.

Finally, the feasibility of the approach on collaborative queries was demonstrated through bioinformatics usecases from from the CAMERA project, a drug ranking usecase scenario from the ViroLab project and workflows inspired by Provenance Challenge usecases for PC1 and PC3.

In addition, the interoperability requirements and challenges for collaborative provenance were summarized and prototype architectures towards solving interoperability during workflow stitching and provenance querying were presented.

This is the right time to introduce such provenance models and query languages as collaborative research projects are ever growing and Web2.0-oriented scientific sharing environments, e.g., myExperiment, are being introduced to allow for sharing and execution of workflows in different workflow systems by groups of users. Thanks to the Provenance Challenge efforts, OPM is starting to be adopted by workflow systems participating in the challenge pushing OPM as a standard for provenance data.

As it was described in Chapter 5, there are many advantages of a collaborative provenance approach since it is all about recording the e-Science activities, specifically, capturing the relationships between human users, workflow executions and data. Firstly, it builds upon existing knowledge and extends it without a re-architecturing of the components in a collaborative e-Science platform. Secondly, it allows for extensions to collaborating entities, e.g., instruments and other system modules, as long as the provenance is kept as system-wide assertions that adhere to a global data model. Thirdly, keeping a history of data and workflow provenance throughout the system increases the trustability of these artifacts and provides mechanisms for users to judge the impact of their work, and, in turn, encourages users to reuse data and workflow from other users and publish theirs. Most importantly, while minimizing the interrupts to scientists and the way they do their work, it impacts the effectivity of the collaborative research by adding value through assisting scientific work. This value comes from being able to analyze collaborations and track the footprint of data. Shortly, this approach brings together consumers and producers of data and other e-Science objects together allowing for proper tracking and acknowledgement mechanisms, leading 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.

10.2 Possible Extensions to the Model

The collaborative provenance model discussed throughout this thesis was built upon three main user actions: users publishing workflows, users publishing data and, users running workflows. The three presented collaborative relationships were identified based on these actions. An opportunity for the extension of the model comes extending these actions and relationships. As a short term extension, we plan to add co-authorship of workflows and data to the model. This will be achieved by revising the model to change the publish relationship between the user, and the workflow and data tables from one-to-many to many-to-many. Through this extension, we can infer a *co-working* relationship between user publishing data and workflows. In addition, in a scenario where users download, repurpose and publish extensions of existing workflows, a need for capturing workflow attribution chains arises. For capturing such a case, we plan to extend our model with *extends* relationships for workflows and data.

A natural extension to the model comes from scientific workflows that are built out of sub-workflows, e.g., composite actors in Kepler. We plan to apply the presented model not only to the collaborative relationships between users running multiple workflows, as described in the thesis, but also with hierarchical or compound workflows using composite actors. We plan to add extensions to Kepler's provenance framework in order to track and analyze such collaborations.

We are also working on further analysis of scaling of the model and query performance by extending the database, potentially, to all the runs in the CAMERA Provenance Database. As mentioned before, an implementation of an online collaborative provenance browser that uses the experimental database that was discussed in Chapter 8 is in progress.

In addition to working on the above-mentioned open issues in the near term as next steps, we plan to focus this work on topics towards achieving the future directions as outlined below.

10.3 Future Directions

This thesis defined a notion of collaborative provenance based on three relationships, namely, users sharing workflows, users sharing data and users running analysis by executing the workflow using the shared data. A number of future work options are considered and planned as extension to this original model. In this chapter, some of these future work plans are described.

10.3.1 Interoperable Collaborative Provenance

As described in the previous chapter, interoperability is one of the biggest challenges to achieve collaborative science. We plan to extend our current interoperability approach for linking different provenance models to go beyond Kepler and Taverna. We are also actively working on the evaluation of the developed collaborative provenance model for different

types of collaborative platforms (Altintas *et al.* 2010*f*). The extensions to the QLP engine as defined by the architecture in Figure 9.3 is under development.

10.3.2 OPM Profile for Collaborative Provenance

OPM Profiles (Moreau *et al.* 2010) are encouraged by the standard to formalize usage of the standard for specific purposes. They can be used as best practice guidelines for purposing provenance information as a specialization of OPM. We plan to create an OPM profile for collaborative provenance including a controlled vocabulary for collaborative provenance relationships, e.g., *WasDerivedFromRunResults* as a substype of *WasDerivedFrom*; *WasCopiedFrom* to map workflow data identifiers to global data identifiers.

10.3.3 Restricted User Spaces

One of the basic assumptions in this thesis was the public nature of data, where the data and workflows were shared publicly at all times and the workflow run provenance was made publicly available. However, most of the existing infrastructures provides support for data and analysis privacy letting the users share parts of the information. In such scenarios, information related to a collaborative chain might not be available or might be available partially, e.g., results are available but not the run provenance. To date, a few models for preserving privacy for standard provenance analysis (Davidson *et al.* 2010) and data mining (Aggarwal and Yu 2008) have been published. We plan to seek such scenarios in existing systems and create techniques for data anonymity (Backstrom *et al.* 2007) and including data policies of infrastructures into account.

10.3.4 Optimization of Collaborative Query Evaluation and Visualization

This thesis focused on generating a first generation data model and query language for collaborative provenance. Expressing queries over lineage relationships can be very cumbersome and difficult to understand, even for experts. We presented extensions to express the collaborative provenance queries in a high-level query language for provenance (QLP), which provides specialized constructs for formulating lineage queries (Anand *et al.* 2009*c*). Based on this data model, we plan to extend our work in (Altintas *et al.* 2010*d*) with implementation of the QLP based collaborative query engine. As a part of our future work, we intend to use the "pointer-based" efficient storage techniques to store the transitive closure in reduced form (Anand *et al.* 2009*b*). We also want to implement this model by using the query optimizations techniques developed in (Anand *et al.* 2010*b*) to execute the provenance queries faster. In addition, for visualization of collaborative provenance views, we would like to extend the Provenance Browser (Anand *et al.* 2010*a*), to display data dependency, run dependency, and user collaboration views spanning across multiple workflow runs.

10.3.5 Semantic Collaborative Provenance Analysis using RDF

Semantic provenance has been a focus area of provenance research community (Groth *et al.* 2009) bringing significant advances to data interpretation. The semantic provenance frameworks (Sahoo *et al.* 2008, Abraham *et al.* 2010, McGuinness *et al.* 2007) for e-Science infrastructures apply domain-specific provenance ontologies to management of provenance data. The Resource Description Framework (RDF[1]) is a common storage representation format for such approaches. We plan to extend our collaborative provenance framework with the ability to add provenance-aware RDF triples on top the current relational collaborative provenance model.

10.3.6 Social Network Analysis using Collaborative Provenance

At the heart of it, the user collaboration view for collaborative provenance is a 2-user interaction graph for a social network. As a part of our future work, we plan to apply quantitative social network analysis techniques (Scott 2000) to 2-user interactions. Using these techniques, we plan to get more insight into the various roles in a collaboration, e.g., connectors, mavens, leaders, bridges, isolates. These techniques can also be used to identify research clusters and who is in them, who is the most influential and which users contribute or utilize the most resources. In particular, we plan to measure:

- Influence as a metric for how a user's work affected other users, i.e., degree centrality (Stephenson and Zelen 1989).
- What expertise is key to the network as well as evaluation of rare expertise and potential areas of growth based on the expertise of users in the network, i.e., betweenness centrality (Freeman 1977)?
- Who has integrated the work from other users and who has contributed to the resource the most, i.e., closeness centrality (Freeman 1979)?
- Potentially connected research networks, i.e., network reach (Otte and Rousseau 2002).
- Who has reach to and combined different research contributions from otherwise separate cliques, i.e., boundary spanners (Yang and Tang 2003)?

10.3.7 Going Beyond Scientific Workflows and Data

The collaborative provenance notion that was defined in this thesis are based on three entities: users (i.e., people), workflows and data. It assumes that users only have one role and there are no other entities. We plan to extend this model to measure the relationships and flows between people with different roles and other system components that act as information

[1]Resource Description Framework (RDF) Model and Syntax Specification: `http://www.w3.org/TR/PR-rdf-syntax/`

or knowledge entities, e.g., instruments, services, etc. This is an important requirement for traceability of transactions in large-scale observatory systems and other networked scientific instruments, e.g., electron microscopes.