Chapter IV
Provenance Tracking and End-User Oriented Query Construction

Bartosz Balis
Institute of Computer Science AGH, Poland

Marian Bubak
Institute of Computer Science AGH, Poland & University of Amsterdam, The Netherlands

Michal Pelczar
ACC CYFRONET AGH, Poland

Jakub Wach
ACC CYFRONET AGH, Poland

ABSTRACT

Provenance tracking is an indispensable element of each e-Science infrastructure for conducting in silico experiments. However, enabling end-users who are non-IT experts to query provenance and experiment data in a meaningful way is equally important. The authors propose an ontology-based provenance model which captures the execution of in silico experiments, as well as domain-specific semantics of data and computations used in those experiments. They demonstrate how ontologies can serve as inter-lingua for end-users, provenance tracking system, and query tools. Query Translation Tools (QUaTRO), enabling end-user oriented, ontology-guided visual querying over provenance records and experiment data, are also presented. In those tools, they also show how the ontology models enable semantic information integration of provenance metadata and experiment data, enabling queries capable of exploring the structure of provenance and associated experiment data. Their approach is demonstrated on a Drug Resistance application deployed in the ViroLab Project.
INTRODUCTION

The term ‘e-Science’ (Hey 2002) was coined to denote a new type of scientific research based on the collaboration within a number of scientific areas, enabled by a next generation infrastructure, wherein people, computing resources, data and instruments are brought together to bring a new quality to the everyday work of researchers. The infrastructure in question is usually identified with Grid systems which offer at least two benefits important for loosely-coupled cross-institution research and collaboration: virtualization and sharing of resources, and building of virtual organizations. Recently, the increasing importance of semantics and knowledge for a future e-Science infrastructure has been emphasized (De Roure 2005). The term Semantic Grid has been used to denote a Grid infrastructure wherein information services are enhanced with well-defined meaning enabling better cooperation between computers and people (Goble, De Roure 2004). They key enabler to achieve the vision of Semantic Grid are Semantic Web technologies, such as ontologies which are a standard, highly sharable, and machine-processable way to represent vocabularies and semantic relationships in a given domain (Goble, Corcho 2006).

Scientific experiment results are neither reliable nor reusable without their provenance, i.e. the information on the origin and history of these results (Groth 2006). Though provenance in computer science originated in the database community, the difference between database provenance and provenance in e-Science has been pointed out (Tan 2007). Both types of provenance describe the origin of a piece of data, i.e. they answer the question what other pieces of data contributed to a given result. However, while in the case of a database, the piece of data is a result of a ‘whitebox’ database query, in e-Science, it is a result of a ‘blackbox’ process. Hence different provenance models and different methods to compute provenance are needed in the two communities (Ibid.).

The importance of provenance tracking in e-Science environments has been pointed out many times and numerous provenance approaches have been proposed. However, providing an adequate end-user support for provenance querying is also a challenge. It has been recognized that the need for provenance queries goes beyond the lineage of a single data item and searching or mining over many provenance records might be useful (Moreau 2007). Nevertheless, there are technical and conceptual barriers preventing or making it difficult for end-users of e-Science environments, such as domain researchers and specialists, to construct complex queries using query languages such as XQuery, SQL or SPARQL. Therefore, there is a need for provenance query support which would enable end-users to construct powerful queries in an easy way.

The goal of this chapter is to propose an ontology-based provenance model which not only can capture the execution of in silico experiments including the domain semantics of data and computations used in an experiment, but also enable end-user oriented visual querying over provenance records. We demonstrate how ontologies are useful as a common language for: (1) end-users who use ontologies as a query language; (2) query tools using them to represent and evaluate queries, and (3) provenance repository in which the ontologies serve as the data model. Moreover, we show how the proposed ontology model enables semantic information integration – recognized as one of the key requirements for a Semantic Grid infrastructure for e-Science (De Roure 2005). Though many semantic data integration approaches have been proposed, we show how ontologies represented in OWL/RDF are used to integrate not only experiment data from various heterogeneous sources, but also data and provenance metadata, where the latter describes the process wherein the former has been used. In this approach, end users are given a unified interface to construct queries that integrate experiment data and provenance metadata.