Chapter III
Data Provenance in Scientific Workflows

Khalid Belhajjame
University of Manchester, UK

Paolo Missier
University of Manchester, UK

Carole A. Goble
University of Manchester, UK

ABSTRACT

Data provenance is key to understanding and interpreting the results of scientific experiments. This chapter introduces and characterises data provenance in scientific workflows using illustrative examples taken from real-world workflows. The characterisation takes the form of a taxonomy that is used for comparing and analysing provenance capabilities supplied by existing scientific workflow systems.

INTRODUCTION

Workflow technology is increasingly adopted in modern science as a means for modelling and automating the enactment of experiments (Ludäscher et al. (006) and Taylor, I. (006). An experiment is specified as a series of analysis operations that are connected using data links that define the flow of data between the outputs and the inputs of the constituent operations, and control links that define temporal dependencies between operations’ executions. The analysis of the results obtained by the enactment of an experiment allows the scientist to verify, amongst others, the validity of a hypothesis or confirm a known fact. To draw such conclusions however, the results returned by the experiment enactment are by no mean sufficient, rather, more information that describes how and where the returned results were obtained from is needed. This information is known in the literature as data provenance or data lineage (Buneman et al. (006). Generally speaking, data provenance can be thought of as the process by which the data instances and the analysis operations that were used to derive a given data instance are identified.
While data provenance in databases, notably in datawarehouses, is a fairly understood problem (Ludäscher et al. (006), it is not comprehended yet in workflows. It was largely recognised by the researches in this area at the second provenance challenge workshop that existing workflow systems provide little support for managing and querying data provenance. A step in this direction, in this chapter we study and propose a taxonomy that thoroughly characterises data provenance in scientific workflows. It is probably worth mentioning that different from the survey on data provenance in eScience by Simmhan et al. (2005), in which the authors attempt to review data provenance in general, be it in databases, workflows or other contexts, in this chapter, we focus on data provenance in scientific workflows and identify their peculiar characteristics. The proposed taxonomy in used as an evaluation and comparison grid for existing workflow provenance systems, thereby highlighting their similarities, differences and limitations. We also discuss open issues that need further attention and investigation from the researchers in the area.

The chapter is organised as follows. We begin by motivating the need for supporting data provenance in scientific workflows using examples of queries a scientist may need an answer for and which can be automatically answered if provenance information is collected. We then formally define scientific workflows and present a real-world workflow to motivate the need for data provenance. We go on to present the taxonomy that characterises data provenance in scientific workflows. The taxonomy is composed of a set of dimensions and associated values. A dimension describes a specific aspect of data provenance and a value represents a possible implementation of a dimension. The taxonomy presented is used for characterising and comparing provenance-related functionalities provided by existing scientific workflow systems, such as Taverna (Oinn et al. (2006)) and Vistrails (Callahan et al. (2006)), highlighting their similarities, differences. It is also used to point out open issues that need further investigation from the provenance community.

BACKGROUND

In this section, we formally define what a scientific workflow is. We then go on to present example of workflow provenance queries taken from the domain of bioinformatics.

Scientific Workflow

A scientific workflow is a collection of interrelated tasks, also called activities or processors, which models a scientific experiment. Different objectives may be aimed from the modelled experiment, e.g., test the validity of a given scientific workflow; verify a known fact; compare the results obtained by a given scientist with the results obtained by other scientists worldwide. A scientific workflow can be decomposed into sub-processes, which contribute to achieve the goal of the global scientific experiment. The enactment of the tasks that compose a scientific workflow is typically constrained by dependency rules among tasks. These dependency rules can be classified into two classes: data-flow dependencies and control-flow dependencies. Data flow dependencies specify the flow of data between the outputs and inputs of the tasks, whereas control-flow dependencies specify temporal dependencies between them. For the purpose of this chapter, we define a scientific workflow by the tuple:

$$wf = \{OP, PARAM, PL, DL, CF\}$$

where $OP$ is the set of operations from which the workflow is composed, $PARAM$ is the set of inputs and outputs of the operations in $OP$, $PL$ is a set of links that associate each operation in $OP$ with its inputs and outputs, $DL$ is the set of data links connecting the operations in $OP$, and $CF$ is a set of controlflow dependencies.