Towards a Model of Provenance and User Views in Scientific Workflows

Shirley Cohen, Sarah Cohen-Boulakia, and Susan Davidson

Department of Computer and Information Science
University of Pennsylvania, USA
{shirleyc, sarahcb, susan}@seas.upenn.edu

Abstract. Scientific experiments are becoming increasingly large and complex, with a commensurate increase in the amount and complexity of data generated. Data, both intermediate and final results, is derived by chaining and nesting together multiple database searches and analytical tools. In many cases, the means by which the data are produced is not known, making the data difficult to interpret and the experiment impossible to reproduce. Provenance in scientific workflows is thus of paramount importance.

In this paper, we provide a formal model of provenance for scientific workflows which is general (i.e. can be used with existing workflow systems, such as Kepler, myGrid and Chimera) and sufficiently expressive to answer the provenance queries we encountered in a number of case studies. Interestingly, our model not only takes into account the chained and nested structure of scientific workflows, but allows asks for provenance at different levels of abstraction (*user views*).

1 Introduction

Fueled by technologies capable of producing massive amounts of data, scientists are faced with an explosion of information which must be rapidly analyzed and combined with other data to form hypotheses and create knowledge. Scientific analyses are thus becoming increasingly large and complex, with a commensurate increase in the amount and complexity of data generated.

To address this problem, over the past several years a number of scientific workflow systems have been developed to support scientists in the analysis of their data. Such systems differ from business-oriented workflow systems in the focus on data – e.g. sequences, phylogenetic trees, proteins – and its transformation into hypotheses and knowledge [23]. Examples of scientific workflow systems include myGrid/Taverna [19], Kepler [5], Chimera [12] and DiscoveryNet [22] (see [30]). Still other interesting examples of workflow systems include MHOLline [25], HKIS-Amadea [9], and AdaptFlow [14]. Some integration solutions also include workflows to add value to warehoused data. For example, the GUS [11] system allows users to import data of interest, run bioinformatics tools over that data, and store the results obtained; pipelines are expressed using Perl.

Scientific workflows are specified using a variety of graph-based models. Nodes in the workflow specification represent *step classes* (alternatively called tasks,

U. Leser, F. Naumann, and B. Eckman (Eds.): DILS 2006, LNBI 4075, pp. 264–279, 2006.

actors, processes, boxes) and edges capture the flow of data between step classes. In many workflow systems (e.g. Kepler and myGrid), a step class may itself be a workflow. An execution of a workflow generates a partial order of *steps*, each of which has a set of *input* and *output* data objects. Each step is an instance of a step class, and the input-output flow of data and class associated with each step must conform to the workflow specification (see for example [16]).

In workflow systems, data, both intermediate and final results, is thus derived by chaining and nesting together multiple database searches and analytical tools. In many cases, the means by which the data are produced is not known, making the data difficult to interpret and the experiment impossible to reproduce. Provenance in scientific workflows is thus of paramount and increasing importance, as evidenced by recent specialized workshops [2] and surveys [23] dedicated to the subject of provenance of scientific information.

Many systems using scientific workflows provide a way to keep track of the origins of data. For example, the GUS schema contains about twenty tables dedicated to provenance information. Some scientific workflow systems, such as myGrid [28], record various kinds of metadata related to provenance. Recently, Kepler has developed a logging mechanism for tracking information and dependencies between components of the data flow [4]. Nevertheless, no formal model of provenance for workflow systems has to our knowledge been developed which precisely defines the meaning of provenance taking into account the nested structure of step classes and the data produced.

Formal models of provenance do exist within the database community (see for example [6,3,27]). However, these models reason over restricted forms of algebraic queries and give very fine-grained reasoning; for example, a tuple in a result gets its value from a particular set of tuples in the input (*where* provenance) and is there because of a (possibly bigger) set of input tuples (*why* provenance). More recently, [7] considers the problem of copying data between databases, and describes an approach in which these actions can be automatically recorded in a convenient, queryable form with acceptable overhead. However, the problem of tracking provenance in scientific workflow systems raises new challenges. First, since the operators in workflows are black boxes (step classes), fine grained reasoning cannot be performed. The most that can be assumed is that steps are *deterministic*, i.e. that given the same set of input the output will be the same. This input must include not only data but also user input (e.g. the selection of results based on visual inspection), parameter settings (e.g. the kind of matrix used in a Blast tool), and any other input used by the step (e.g. a randomize number used in a bootstrap). Second, scientific workflow systems frequently provide a notion of *user views* which determines whether or not a user can zoom into a step class to see a sub-workflow. User views therefore affect the granularity at which provenance is reasoned about.

The aim of this paper is to present a formal model of provenance in workflow systems which takes into account the chained and nested structure of scientific workflows as well as user views. The model has been formulated by interviewing numerous scientists in several domains (e.g. genomic research, and