

# Towards Privacy Aware Data Analysis Workflows for e-Science

**William K. Cheung**

Department of Computer Science  
Hong Kong Baptist University  
Kowloon Tong, Hong Kong  
william@comp.hkbu.edu.hk

**Yolanda Gil**

Information Sciences Institute  
University of Southern California  
4676 Admiralty Way, Marina del Rey, CA 90292  
gil@isi.edu

## Abstract

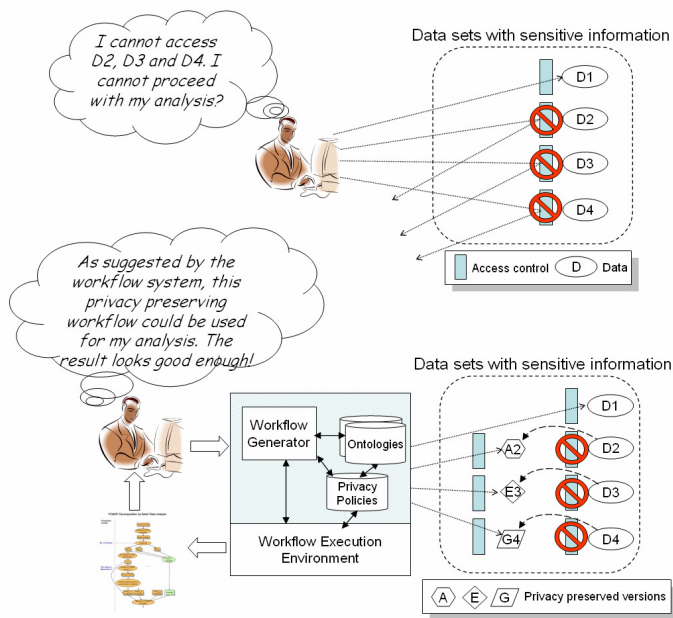
e-Science is getting more distributed and collaborative and data privacy quickly becomes a major concern, especially when the data contain sensitive information. Existing data access policies for privacy management are too restrictive for supporting the large variety of data analysis needs in e-Science. In this paper, we argue the need of a new type of policies that govern data privacy based on the type of processing done on the data. A semantic workflow approach is proposed to address the challenge. Data analysis processes are described as workflows. Ontologies for data analysis and privacy preservation describe the functionalities and the privacy attributes of the processes, as well as process-constraining privacy policies. We give some examples of related policies with their potential fields for application explained. Also, we present via a case study on distributed data clustering to illustrate how the approach could be integrated with a workflow system to make it privacy aware.

## Introduction

Data privacy is important in e-science, especially when distributed and collaborative data analysis processes are involved. It is not difficult to find scenarios where distributed data analysis and data privacy protection are both needed at the same time. For example, one can analyze individuals' clinical data like brain images by gaining access to related remote sources for disease diagnosis (Beltrame et al. 2006), where the patients' identity has to be kept strictly confidential. Other than the subjects' identity threat, the scientists themselves may have privacy concerns on their scientific findings as data sets, preliminary results and data analysis processes can now be easily and widely shared in e-Science collaborations (Deelman and Gil 2006). These privacy concerns are important even though the advantages of sharing data to facilitate collaborative scientific research are well understood (NIH 2004). Thus, the need for having privacy management support in e-Science is immediate.

Existing data access policies offer a very basic privacy protection mechanism, where a user can access a data source if his/her certificates and credentials satisfy the access policies defined for that data source. An alternative approach that is less restrictive is to apply privacy preserving techniques to the data before releasing it, where sensitive information like personal identity or medical background are properly hidden through anonymization and partitioning (Samarati 2001). In addition, one can design new data analysis algorithms that are privacy preserving, a fast-growing area for the past few years (Chris et al. 2003). These algorithms can preserve data privacy through techniques like secure multiparty computation (Lin et al 2005) and data generalization (Zhang & Cheung 2005), and yet can perform reasonable data analysis.

These existing mechanisms alone are too restrictive for many applications, especially in e-Science. Consider the case of a cancer patient signing a release form for their medical records. He or she may be not only willing but eager to allow access to for medical research purposes as long as it is anonymized. However, they suspect that if the record is released it could be used by insurance companies to design more profitable insurance rates that would raise his or her medical expenses. Given the choice to release the data or not without any say on the use of the data, the patient may decline to allow the use of its record. Clearly, a more flexible privacy mechanism would allow the patient to specify a policy on how the data will be processed, not just on the blanket release of the data per se. Consider another case, where a cancer research laboratory has collected treatment protocol data for thousands of patients over several decades. The lab is happy to share its data with medical researchers in other fields to analyze the relationship of cancer with liver transplantation or with heart failure, but would not want other competing research groups to use the data in competing cancer research quests. Existing approaches would not allow the lab to specify



**Figure 1. Data analysis using (a) traditional access control, (b) privacy-aware workflow systems.**

this kind of policy concerning the use of the data, only to specify who would have access to it.

The goal of our work is to investigate *a new kind of privacy protection policies that constrain the type of processing on the data*, rather than the access to the data. That is, instead of defining policies to specify who can access a data set and how much of it can be accessed, our goal is to define policies that specify *what can be done with the data*. This would allow a more flexible approach to privacy that covers data processing in addition to the existing data access techniques.

The key idea in our approach to privacy protection is the use of *workflows* to describe the type of processing done to a dataset, and to express policies that can be used to control the creation and execution of workflows. Workflows have recently emerged as a useful paradigm to represent and manage complex computations in many scientific applications (Deelman and Gil 2006). *We propose to extend workflow systems to be privacy-aware*, so that they can be given privacy policies defined in terms of types of analysis and data handling performed by the workflow system. Figure 1 illustrates a traditional access control approach (a) and contrasts it with a privacy-aware workflow system (b). A data access control policy would either enable or disable a user's access solely based on their credentials and certificates. In this example, the user is only able to access D1 but not D2, D3, or D4. In contrast, a privacy-aware workflow system would enable the expression of additional kinds of privacy policies that would enable access based on the type of analysis done as

expressed in the workflow. The workflow system could assist the user in modifying the analysis in order to satisfy the privacy policies stated. In this framework we can selectively specify privacy policies for the same data set that would allow it to be accessible for certain types of workflows (analyses) and not for others.

This paper describes our work to date in defining a new class of privacy policies that apply at the workflow (data processing) level. To define these workflow-level policies, one needs to address the following major issues:

- how to properly represent policies in terms of data analysis processes, data privacy concepts and related workflow constructs (ontological issue),
- how to automatically enforce those policies in data analysis process management within the workflow system (policy enforcement issue), and
- how to provide provenance regarding the privacy of the data as well as their data analysis history so that the system can justify its use of the data (provenance issue).

In this paper, we describe how a semantic approach can be adopted to address those issues in the context of workflow. In particular, we show how semantic web technology can be used to describe data analysis workflows as well as their data privacy requirements. Also, we explain how the privacy related ontology could be used together with some policy framework for representing privacy policies for controlling data analysis process creation and execution. Examples of possible privacy policies enabled by the introduced privacy awareness are provided. In addition, a case study is presented to illustrate how the proposed approach can be used to govern a particular distributed data mining process.

## Motivation

Managing privacy in data analysis processes in e-Science has two different aspects of concern, namely *data* privacy and *process* privacy. The former one concerns the privacy of *data sources* as well as *data products* created during data analysis processes. The latter one concerns the privacy of knowledge captured in *data analysis processes*. In this paper, we mainly focus on protection of data privacy in the context of workflow.

## Our Goal: Privacy-Aware Data Analysis

Instead of specifying policies to control data access, we set policies on types of data analysis that can be applied to the data. The following are some higher level expressions of the policies that we are targeting where notions like purposes of analysis, types of analysis, characteristics related to data privacy and analysis accuracy are involved to govern data analysis processes.

Example 1 *Patient medical images* should not be released for analysis except for the *purpose of supporting a particular medical image analysis project* and the images have to be *encrypted* if they are transmitted via untrusted networks.

Example 2 Given the *purpose of medical diagnosis*, any *classification* performed on *clinical data* must provide the *confidence level* for each data item and have its *overall accuracy* reaching a particular *level of quality standard*.

Example 3 Data containing *drug dosage information* should not be released for *any analysis* except for the *purpose of public health care study*, and the data should not contain *any personal identification attribute* and have to be *properly anonymized* before they can be used.

For the three examples provided, terms in italics reveal the need of a vocabulary to describe workflow-relevant concepts about data privacy and data analysis, and the remaining non-italic portions correspond to the constructs for describing policies in existing data access control frameworks. With a similar analogy, our proposed approach (to be shown in the later sections) also involves two parts, namely ontological description of privacy and data analysis, and the adoption of a policy framework with the ontological description integrated.

### **Related Research on Privacy Preserving Data Analysis**

While restricting access to the data could be found to restrict to support various kind of data analysis, one could adopt the approach of restricting information in the data so that they are (a) free of *identifiers* that would permit linkages to any target individual and (b) free of *content* that would create unacceptably high risks of individual identification. For example, one may allow a set of data to be released and analyzed as far as fields related to personal information are anonymized.

In the literature, techniques for releasing data without disclosing sensitive information have been proposed for various applications. For example, cryptography-based techniques have been found useful in private data communication in untrusted networks (Stalling 2005). Techniques like anonymization (Samarati 2001) and microaggregation (Domingo-Ferrer & Mateo-Sanz 2002) have been found useful in applications like statistical disclosure control. Also, there has been recent research interest in developing data mining algorithms which are privacy preserving with underlying techniques including secure multiparty computation (Lin, Clifton & Zhu 2005), random data perturbation (Kargupta et al. 2005) and data generalization (Cheung et al. 2006).

Before we proceed, it is worth mentioning that our concern is not only limited to the identity threat. In fact, one could generalize the target to be hidden from individuals' identity to some important data attributes or experiment runs which will depend on the particular application and situation at hand.

### **Related Research on Policy Governed Data Analysis**

As an alternative approach for privacy protection in data analysis, policies of data usage can be adopted for governing data analysis processes (Weitzner et al. 2006). For example, a research lab wants some of their on-line data and analysis tools to be only used for the purpose of demonstrating the system's analysis capability and thus posts a related data usage policy. In case the data set is later on found to be used (say together some other data sources) for re-identification disclosure of the subjects who provide the data, the one doing that will be accountable for the consequence. In addition, it will be even more appealing if such policy-violating data analysis processes can be caught early on and be stopped before they are actually executed.

While there has been work found in the literature for representing and reasoning about privacy policies (Bradshaw et al. 2003, Kagal, Finin & Joshi 2003), only conventional security concepts like authentication, authorization and encryption have been considered. We envision that privacy preserving data analysis techniques will soon get more mature and widely accepted. The family of privacy policies will need to be further enriched with the additional privacy related semantics being properly represented and reasoned.

### **Approach**

Our approach to develop a privacy-aware data analysis framework is to extend existing workflow systems to incorporate privacy policies that control the type of data analysis done on the data. Modeling data analysis processes as workflows, also called scientific workflows, is common in e-Science (Deelman & Gil 2006, Ludascher et al. 2006, Oinn et al. 2006, Wassermann et al. 2006). However, in the literature, workflow systems possessing data privacy awareness are still lacking. Conventional workflow systems were designed with the primary objectives of providing component abstraction, interconnectivity and reliable execution in mind. In e-Science, data oriented and user (scientist) oriented perspectives have been stressed. Examples include the use of visual programming environments for constructing the data analysis processes, e.g., Taverna (Oinn et al. 2006), Kepler (Ludascher et al. 2006) and Sedna (Wassermann et











