Letter from the Special Issue Editor

The importance of data provenance has been increasingly recognized by both users and publishers of data. For users of data, the scientific basis of their analysis relies largely on the credibility and trusthworthiness of their input data. For publishers of data, the provision of provenance as part of their published data is important for scholarship and reproducibility. In today's Internet era, where complex ecosystems of data are even more prevalent, it is no wonder that data provenance has become a major research topic in many conferences and workshops. Data provenance was already the topic of the December 2007 issue of IEEE Data Engineering Bulletin. This issue attempts to complement the December 2007 issue by focusing on how provenance has been captured and exploited by systems that either have been developed or are in the process of being developed in the industry and the academia, and by reporting on new interesting research directions.

The first three articles describe how provenance has been applied in systems developed in the industry and academia. *Provenance in ORCHESTRA*, by Green *et al.*, describes a collaborative data sharing system ORCHESTRA, whose architecture and provenance model are largely motivated from the needs of the life sciences community. As part of ORCHESTRA, the authors describe an interesting application of provenance to the specification of trust policies and update exchange. In the article titled *Refining Information Extraction Rules using Data Provenance*, Liu *et al.* describe how provenance can be used to aid the process of developing information extraction rules. Here, provenance is exploited to understand the changes that need to be made to the information extraction rules so that the result is what the user desires. Gonzalez *et al.* describe how Google Fusion Tables can be used to faciliate collaboration around data sets in their article *Socializing Data with Google Fusion Tables.* Attribution and provenance are anticipated to play a significant role in socializing data and this article provides an accessible exposition to the attribution feature of Google Fusion Tables, as well as the authors' experience with the system.

It is important for scientists to be able to pinpoint and cite data easily for proper scholarship. In the article *A Rule-Based Citation System for Structured and Evolving Datasets*, Buneman and Silvello discuss the requirements that must be met by a citation system in order to guarantee consistency and integrity of citations. The issues that arise in two data sets found in the fields of Digital Libraries and Curated Databases, as well as a rule-base system for automatically generating proper citations that addresses the issues are fully described in this article.

According to the December 2007 issue of IEEE Data Engineering Bulletin, connecting workflow and data provenance is said to be "the most interesting research challenge in the general field of provenance models". In *Panda: A System for Provenance and Data*, Ikeda and Widom report on their work-in-progress Panda project, whose goal is to develop a system that would seamlessly manage and query both data-based and process-based provenance. The authors have provided a simple and yet elaborate example for explaining various aspects of their vision, as well as the challenges associated with building such a system.

As scientists tend to follow the "extract, transform, and analyze" workflow model to process scientific data, many workflow systems aim to provide scientists with a platform for easily authoring workflows. The Trident scientific workflow system is described in the article *Provenance for Scientific Workflows Towards Reproducible Research.* Barga *et al.* give a good overview of various provenance-related components of Trident. An integral part of Trident is the automatic capturing of provenance in both control-flow and data-flow aspects of workflows, as well as in workflow evolution.

Finally, the article *Causality in Databases* takes a fresh perspective at the problem of explaining outputs of database transformations. Meliou *et al.* give an accessible description of how the notions of causality and responsibility, which are well-studied notions in philosophy, can be applied to databases to provide explanations of answers to database transformations, and point out several promising research directions.

Wang-Chiew Tan
IBM Almaden Research Center and UC Santa Cruz