Letter from the Special Issue Editors

The quality of data has always been important to businesses and intelligence, as policies and business decisions are often made based on analysis performed on data. This issue contains articles that explore different aspects of data quality that frequently arise in the context of multiple (Web) data sources providing overlapping, complementary, and sometimes contradictory information about the same concept or real world entity. It is therefore important to provide techniques for understanding the truthfulness and trustworthiness of data sources to the extent possible, and reconcile their differences in order to create a clean integrated view of the underlying data sources. Techniques for entity resolution, mapping, fusion, and data cleaning are important "ingredients" towards achieving such a clean unified view.

The first three articles in this issue describe different approaches for understanding and improving data quality in terms of truthfulness, trustworthiness, and respectively, eliminating (unnecessary) nulls. In the article *Truthfulness Analysis of Fact Statements Using the Web*, Li, Meng, and Yu describe an approach for determining the truthfulness of fact statements obtained from the Web. Whether a fact statement is considered true depends in part on the degree of truth of alternatives to the fact statement. In the next article *Corroborating Information from Web Sources*, Marian and Wu describe how one can exploit corroboration (i.e., similar answers provided by different data sources) to provide a measure of trust in the answers obtained from different sources. In the article *Eliminating NULLs with Subsumption and Complementation*, Bleiholder, Herschel, and Naumann describe how data quality can be improved by data fusion operators, subsumption and complementation, that minimizes the number of nulls in data. In addition, the authors described a set of rewrite rules based on these two operators and other database primitives that can be used to optimize the process of eliminating nulls.

The next two articles describe how constraints can be used to understand the quality of data and how different constraints that have been used for improving data quality can be captured uniformly in one language. In the article *Efficient and Effective Analysis of Data Quality using Pattern Tableaux*, Golab, Korn, and Srivastava describe the *Data Auditor* system, which can efficiently compute patterns of subsets of the underlying data that satisfy or violate a given constraint. The patterns derived are useful for understanding the semantics of the data. Fan and Geerts, in the article titled *A Uniform Dependency Language for Improving Data Quality*, describe how different dependency formalisms that have been used in the context of studying data quality can all be captured under their language of Quality Improving Dependencies (QID). They describe how a modification of the chase process can be used to enforce QIDs and thus, provide a uniform framework for reasoning and understanding data cleaning mechanisms behind different existing formalisms.

The rest of the articles in this issue describe generic frameworks towards data cleaning, entity resolution, and data integration in general. In the article *Towards a Domain Independent Platform for Data Cleaning*, Arasu *et al.* describe how various generic data cleaning primitives can be consolidated into a tool for facilitating the process of data cleaning. The primitives consist of various generic operators for determining textual similarity, clustering, and parsing. While these primitives are generic, the platform allows these operators to be customized to specific domains. Whang and Garcia-Molina described their efforts in a generic framework for entity resolution. They extended the core entity-resolution primitives for scalable iterative blocking, joint resolution, and incremental resolution in the article *Developments in Generic Entity Resolution*. They have also described how different existing quality measures could lead to different evaluations of the entity-resolution results and hence, proposed a single but configurable measure. Finally, the article *Extracting, Linking and Integrating Data from Public Sources: A Financial Case Study* showcases a real-world example of integrating various data sources from the financial domain. The integration process requires applications of techniques from mapping, entity resolution, and data fusion. As Burdick *et al.* described, the integrated result can be analyzed to provide valuable insights that is difficult to obtain otherwise. From this experience, the authors plan to develop and extend capabilities towards a generic, high-level infrastructure for large-scale data integration.

Xin Luna Dong and Wang-Chiew Tan
AT&T Labs-Research, IBM Research-Almaden and UC Santa Cruz