

PREFACE

The *First International Workshop on Data Quality in Collaborative Information Systems* was held in conjunction with the 13th International Conference on Database Systems for Advanced Applications, 19-22nd March, 2008, in New Delhi, India. The present volume contains the texts for four accepted papers and two invited papers presented at the workshop, as well as one additional paper presented at the co-located workshop on analysis of high dimensional discrete data.

Poor data quality is known to compromise the credibility and efficiency of commercial as well as public endeavours. Several developments from industry as well as academia have contributed significantly towards addressing the problem. These typically include analysts and practitioners who have contributed to the design of strategies and methodologies for data governance; solution architects including software vendors who have contributed towards appropriate system architectures that promote data integration and; and data experts who have contributed to data quality problems such as duplicate detection, identification of outliers, consistency checking and many more through the use of computational techniques. The attainment of true data quality lies at the convergence of the three aspects, namely organizational, architectural and computational.

At the same time, importance of managing data quality has increased manifold in today's global information sharing environments, as the diversity of sources, formats and volume of data grows. In this workshop we target data quality in the light of collaborative information systems where data creation and ownership is increasingly difficult to establish. Collaborative settings are evident in enterprise systems, where partner/customer data may pollute enterprise data bases raising the need for data source attribution, as well as in scientific applications, where data lineage across long running collaborative scientific processes needs to be established. Collaborative settings thus warrant a pipeline of data quality methods and techniques that commence with (source) data assessment, data cleansing, methods for sustained quality, integration and linkage, and eventually ability for audit and attribution.

The workshop was intended to provide a forum to bring together diverse researchers and make a consolidated contribution to new and extended methods to address the challenges of data quality in collaborative settings. Topics covered by the workshop include at least the following: Data linkage and fusion, Entity resolution, duplicate detection, and consistency checking, Data profiling and preparation, Use of data mining for data quality assessment, Methods for data transformation, reconciliation, consolidation, Data lineage and provenance, Models, frameworks, methodologies and metrics for data quality, Application specific data quality, case studies, and experience reports.

We would like to extend our appreciation and gratitude to the program committee of the workshop, who provided valuable reviews of submitted papers, and ensured the quality of the accepted papers and success of the workshop.

Nick Koudas, University of Toronto Markus Helfert, Dublin City University Floris Geerts, University of Edinburgh Xuemin Lin, University of New South Wales Rosanne Price, The University of Melbourne Wasim Sadiq, SAP Research Graeme Shanks, University of Melbourne Diane Strong, Worcester Polytechnic Institute Kerry Taylor, CSIRO ICT Centre Harry Zhu, Old Dominion University

We would also like to thank the keynote speaker of the workshop, *Prof. Divesh Srivastava*, from AT&T Labs-Research for his inspiring talk on “The Bellman data quality browser”.

Last but not least, we would like to acknowledge the authors of the papers presented at the workshop, and thank them for their valuable contributions to the workshop as well as to the challenging and diverse area of data quality.

Shazia Sadiq, Xiaofang Zhou
Organizing Co-chairs

The University of Queensland.
April, 2008