Editorial for the Inaugural Issue of the *ACM Journal of Data and Information Quality* (*JDIQ*)

MOTIVATION FOR *JDIQ*

A growing component of organizational operations today involves the collection, storage, and dissemination of unprecedented vast volumes of data. However, this expansion comes not without growing pains. Organizations are often unable to translate this data into meaningful insights that can be used to improve business processes and change the way we work. The reasons for this difficulty can often be traced to issues of data and information quality, involving both problematic symptoms and their underlying causes. Previously collected data can turn out to be inconsistent, inaccurate, incomplete, or out-of-date. Organizations can have inappropriate or conflicting strategies across the "pockets" of an enterprise that interfere with the ability to get the right information to the right stakeholders in the right format at the right place and time. To make matters worse, the boundary of stakeholders is broadening and increasingly involves extended enterprises often reaching a global inter-enterprise scale. The time horizon for the use of information also becomes an open and moving target. In recent years, several terms have emerged to refer to these issues, such as Information Quality and Data Quality. We have chosen to name this journal Data and Information Quality to cover the full range of issues and will generally use these terms interchangeably.

Complicating matters is the fact that today's organizations need to do more with their data if they are to compete effectively. Data quality as measured by its fitness for use in a particular application is a major consideration and possibly a thorny issue when discussing issues such as data privacy and protection, data lineage and provenance, enterprise architecture, data mining, data cleaning, as well as data integration processes such as entity resolution and master data-management. Particularly in the area of data integration processes, organizations must grapple with how to deal with incomplete customer data, inaccurate or conflicting data, and fuzzy data as they strive to develop measures of confidence for the information produced in this environment.

Even more daunting is the reality that even if organizations get the creation and management of information right for current stakeholders, there is always the prospect of unexpected future stakeholders to consider. How does one ensure that over the long term information will remain accessible, trustworthy, and meaningful in the face of rapidly changing computing and storage technologies and corresponding demands for use? What types of models, methods,

and metadata will be needed to represent, preserve, and query data lineage and provenance, possibly for centuries to come?

Research on information quality that addresses these issues is not entirely new. Several disciplines such as statistics, library sciences, accounting, computer science, and management information systems have examined some of these issues. What is new and important, however, is that there is now a movement towards a unified body of knowledge that addresses information quality in its entirety rather than in a piecemeal fashion. This *ACM Journal of Data and Information Quality* (*JDIQ*) is inspired by and a tribute to this movement.

A NEW FORUM FOR LOOKING INWARD AND OUTWARD

A world of increasing technological capability is a world of increasing demands on data. Data is created, used, and disseminated in astoundingly complex ways and at dizzying rates, thereby increasing the need for the most useful, usable, and trustworthy information possible. We see examples of this need as data travels from the reaches of space back down to earth. It is hard to imagine answering questions such as "is there life on Mars?" without all the data collected on the surface of the Red Planet and sent back 35 million miles to anxiously waiting NASA scientists and a global audience. Indeed, it is this global audience that constitutes the backbone of the mounting demand for high-quality data and information. Here on earth, a doctor can monitor the heart of a rural farmer thousands of miles away partly because data can be transmitted effectively at such distances, but more importantly, because the doctor can trust the quality of the data enough to use it to diagnose and monitor the patient's heart.

In these and many other cases, we see a common trajectory for information quality research and practice. The spiraling effects of our collective quest to understand ourselves (e.g., the genome human project) and to improve the world (e.g., global information integration projects) demand ever-widening research into how to improve the quality of data and information. *JDIQ* seeks to respond to the need of various communities engaged in the dialog about and inquiry into data and information quality by providing a forum for innovative theories that can be tested in the pragmatic environments of organizations.

As virtual distance reduces physical distance and our global village flattens, our dependence on each other for data and information increases all the more. This dependency on shared information is arguably as great as our dependency on physical resources. Nearly every realm of human activity, from science and technology to news and business, generates both formal and informal information. In such environments, decisions are often based on data from numerous and sometimes unknown origins. Furthermore, this data is constantly transformed by a variety of processes for storage, maintenance, retrieval, and use. Despite the complexities involved, sharing data and information has become a default policy and practice in many organizations and leads to the often asked question: how good is the quality of our data and information?

Both the academic and practitioner communities are responding to these challenges with fresh ideas. It is our hope that *JDIQ* will provide a much

needed forum for such ideas and related research. Readers will find useful articles exploring innovative theories and pragmatic solutions in data and information quality, while authors will have the opportunity to sharpen their ideas and dig deeper into new studies, producing and sharing pioneering theories and solutions that can be further developed into a pragmatic agenda for today's and tomorrow's companies, public agencies, and users across our interconnected world.

As information is shared and distributed and as organizations use data in more versatile ways, problems and issues for study grow more complex, and solutions to these problems demand novel approaches. As the relatively new, multi-disciplinary field of data and information quality has advanced and become more demanding, it has become evident that the field needed a home. Institutionally backed forums give a powerful boost to developing fields, bringing research to the next level and providing a broad academic and professional community to exchange ideas and test hypotheses. It is our hope that *JDIQ* will be one such forum for the data and information quality field.

WELCOME TO AUTHORS AND COLLABORATORS

JDIQ welcomes innovative research questions with pragmatic implications in mind. We particularly encourage interdisciplinary approaches to conducting research that draws from a variety of theoretical backgrounds and practical evidences for insights. We welcome diverse research methods that span from quantitative to qualitative approaches used for both theory-building and theory-testing. Topics of interest range from technical prescriptions to organizational and normative explanations. In addition to research contents and results, articles should articulate a clear research question related to data and information quality, report the research method used, discuss the theoretical foundations, and provide implications for future research and practice.

Along with regular research articles, we also plan to provide selective, well thought-out industry business cases that document compelling experiences, successful or failed, in actual organizations. We welcome suggestions for ways in which the journal can improve as it seeks to provide a critical forum for advancing the field's research and practice. More information about *JDIQ* and instructions for authors can be found on the *JDIQ* Web site at http://jdiq.acm.org/.

JDIQ has also established relationships with several conferences and workshops that incorporate themes relevant to *JDIQ*. We are working with those program committees to identify high-quality papers that, when expanded to journal length, would be appropriate for fast-track processing by *JDIQ*. We encourage discussions with other program committees that are interested in pursuing this type of collaboration.

INTRODUCTION TO THE ARTICLES IN THE FIRST ISSUE

With this first issue, we are venturing into an exciting world of dialogue with our readers and authors on data and information quality. We appreciate your contributions, feedback, and cheers.

We invite our readers to delve into the four articles of this inaugural issue. Article one gives an overview of the current literature and recommends future directions. Article two asks how we can solve the problem of missing data using experimental research as a key method. Article three uses the action-research method to look at how data governance in industry is practiced. Article four asks how to devise a procedure for assessing the currency of data using a modified design-science method.

The first article, "Overview and Framework for Data and Information Quality Research" by Stuart Madnick, Yang Lee, Richard Wang, and Hongwei Zhu, offers an overview of the evolution and current landscape of data and information quality research. The authors introduce a framework and use it to characterize the research along two dimensions, topics and methods. Representative works are cited for purposes of illustrating the issues addressed and the methods used in earlier data and information quality research. The authors also identify and discuss challenges to be addressed in future research.

Researchers and prospective authors are encouraged to further their research based on the frameworks and examples introduced in this article. We particularly encourage researchers to apply multiple and diverse research methods and to ask new questions. Practitioners are encouraged to find a way to implement various theoretical findings provided in this article in their organizations. We would also like to hear from practitioners about new problems and issues they face, lessons learned, and ways to improve the theoretical findings.

The second article, "A Bayesian Approach for Estimating and Replacing Missing Categorical Data" by Xiaobai Li, proposes a new approach for estimating and replacing missing categorical data. Applying the Bayesian method, the posterior probabilities of a missing attribute value belonging to a certain category are estimated. The results of this experimental study demonstrate the effectiveness of the proposed approach.

Researchers are encouraged to conduct similar experiments in different settings to confirm the efficacy of the proposed approach or propose a different approach inspired by this example. Practitioners can find a useful application from the theoretical findings in treating and solving missing data for critical datasets in their organizations.

"One Size Does Not Fit All – A Contingency Approach to Data Governance" by Kristin Weber, Boris Otto, and Hubert Oesterle is the third article. It discusses data governance by applying IT governance and organizational theory. The article presents findings from a community-action research project involving six international companies in various industries. It outlines a data governance model that consists of three components, data quality roles, decision areas, and responsibilities. The authors also describe a data-governance contingency model and demonstrate the influence of performance strategy, diversification breadth, organization structure, competitive strategy, degree-of-process harmonization, degree-of-market regulation, and decision-making style on data governance.

We encourage researchers to further their research based on this work on data governance. We see that there are many avenues to untapped

organizational, psychological, socio-economical, and managerial issues in the information quality areas. Practitioners can implement some of the findings from this article and further articulate and operationalize the suggestions from this article.

The fourth article is "A Procedure to Develop Metrics for Currency and its Application in CRM" by Bernd Heinrich, Mathias Klier, and Marcus Kaiser. This article provides a procedure for developing metrics for quantifying the currency of data. The authors suggest a six-step procedure which involves probability-based metrics for quantifying data currency. The six steps include, (1) selection of an attribute, (2) identification of impact factors, (3) acquisition of data on the impact factors, (4) selection of data distribution and estimation of the distribution parameters, (5) definition of a metric for the attribute, and (6) application of the metric. The authors also describe the results from applying this metric in a major German mobile service provider.

Researchers are encouraged to identity various ways to assess the quality of data and to further identify costs and benefits of differing qualities of data. Practitioners can find a way to implement a part of the suggested metric and find a way to improve its implementation in their organizations and industries.

PREVIEW OF PAPERS IN THE SECOND ISSUE

We also invite our readers to the forthcoming second issue of *JDIQ*, a special issue on Data Quality in Database Systems, organized by the guest editors, Professors Louiqua Raschid and Felix Naumann. The four articles in this issue will be:

—"Mining in Large Noisy Domain" by Manoranjan Dash,
—"Optimal Stopping: A Record-Linkage Approach" by Vassilios Verykios,
—"Representing Data Quality in Sensor Data Streaming Environments" by Anja Klein, and
—"Incorporating Domain-Specific Information Quality Constraints into Database Queries" by Suzanne Embury.

ACKNOWLEDGMENTS

We also thank the journal's senior advisory board members chaired by Professor Richard Wang. They are Professors Robert Austin, Peter Chen, Benn Konsynski, Allen Lee, Jay Nunamaker, Tamer Özsu, Edgar Sibley, Richard Wang, Andrew Whinston, and Jennifer Widom. We particularly appreciate our senior advisors who provided valuable and timely advice on the research ethics, lessons learned from their experiences as editors-in-chief, and suggested lists of topics and researchers for special issues.

We acknowledge Raïssa Katz-Hass for her valuable advice on editing, Chris Yarawich for editing manuscripts, and Michael Cullinan for maintaining the Web site for the journal. Our special appreciation goes to the journal's current Information Officer and Administrator, Ms. Jiun Hsu, and our first Managing Editor and Information Officer, Professor Elisabeth Pierce, for their tireless and devoted work on many fronts, ranging from preparing numerous proposal drafts to dealing with tedious glitches for a smooth and successful launch of the journal. We cannot thank them enough for their weekend devotion and late-night conference calls.

TO OUR AUTHORS AND READERS

To prospective authors and readers, we want to thank you for keeping this emerging field alive and exciting by demanding and producing exceptional articles. We look forward to hearing from you.

<div align="right">

Stuart E. Madnick
Massachusetts Institute of Technology
smadnick@mit.edu
and
Yang W. Lee
Northeastern University
y.lee@neu.edu, ylee@mit.edu
Editors-in-Chief

</div>