Guest Editors’ Introduction

Data Quality in the Internet Era

In the Internet era, information is accessible to and published by everyone in a free and uncontrolled way. New technologies such as mashups and service-oriented computing let users search, query, and employ such data to provide new services. New data types, such as geographical and multimedia video, are becoming common in the day-to-day user experience. This incredible amount of available data introduces new challenging data quality problems, such as accessibility and usability. Low information quality is common in governmental, commercial, and industrial Web applications (including Web 2.0 tools). Consequently, information quality on the Web is one of the most crucial requirements for an effective use of Web data and the pervasive deployment of Web-based applications.

Emerging Challenges

The marriage of semantic technologies and the Internet represented a revolution in the way we think about Web applications and generated new challenges for data quality researchers. For example, linked data – that is, the use of RDF technologies to expose, share, and connect pieces of data — lets us easily create complex knowledge networks. From a data quality viewpoint, linked data represents a new and interesting challenge due to the amount of data that can be linked, the network’s dynamicity, and the possibility of using external knowledge thanks to semantic technologies.

Ontology instance matching is an example of a challenge from the Semantic Web. In fact, until now, one of the most important issues data quality researchers addressed was ontology alignment – that is, determining correspondences between concepts. A new research problem in this area is instance matching, or evaluating the degree of similarity among different descriptions of real objects across heterogeneous data sources. In the data quality field, this problem is known by different names, such as record matching, record linkage, the merge/purge problem, or entity resolution. A need exists for solutions combining semantic technologies and data quality ones.

Another significant challenge is...
evaluating trust in Web data. Indeed, the Web lets us access a huge amount of data, but it’s becoming more and more important to discriminate among them. It’s not only necessary to have information about a data source’s reputation but also that we model and assess trust in the specific data provided.

A very important issue when dealing with Web data is enforcing privacy constraints when they’re specified. In such cases, published data are often subject to data transformation that alter their quality. Trade-offs between the quality of Web-available data and data privacy must still be investigated and constitute a current research topic. The Web is not only a source of new challenges but represents a new powerful tool for solving well-known problems. For example, typical data standardization or data cleaning activities can be improved by using the information derived from the Web. As another example, geographical information about affiliation data are available on the Web and can be easily used thanks to mashup or RESTful applications.

In this Issue
This special issue received more than 20 submissions, but three articles in particular best represented the myriad data quality issues in existence today.

In “Information Quality in Mashups,” Cinzia Cappiello, Florian Daniel, Maristella Matera, and Cesare Pautasso focus on assessing the quality of mashup applications, which are gaining popularity even with users who have few programming skills. Mashups support users not only in creating content and annotations but also in “composing” applications starting from third-party content and functions. The article aims to assess the quality of the information a mashup provides, which requires understanding how the mashup has been developed, what its components look like, and how quality propagates from basic components to the final mashup application.

The article “Learning-Based Approaches for Matching Web Data Entities,” from Hanna Köpcke, Andreas Thor, and Erhard Rahm, addresses the record-linkage issue. Record linkage is the process of determining whether two different records represent the same real-world object. This problem is much harder in the context of Web data. Effective record linkage typically requires combining several match techniques and finding suitable configuration parameters, such as similarity thresholds. This article adopts a machine learning technique to semiautomatically determine suitable match strategies with a limited amount of manual effort for training.

Finally, “Toward Uncertain Business Intelligence: The Case of Key Indicators,” by Carlos Rodríguez, Florian Daniel, Fabio Casati, and Cinzia Cappiello, provides a different viewpoint with respect to data quality. It moves the attention from operational data to data warehouses. In the context of Web-enabled intercompany cooperation and IT outsourcing, the adoption of service-oriented architectures and the use of external Web services might hinder a comprehensive view over distributed business processes and raise doubts about computed outputs’ reliability. The article provides new definitions about uncertain events and uncertain key indicators, a model to express and store uncertainty, and a tool to compute and visualize uncertainty.

These three articles represent only some of the current research on data quality. Other directions not discussed here but worthy of further investigation include quality of user-generated content, data quality in Web services, trust, methodologies for Web data, and detection of source dependency. Data quality isn’t a young field, but the unprecedented availability of massive amounts of data from the Web offers new challenges in interdisciplinary research – computer science, statistics, psychology, and mathematics, to name a few. Moreover, a look at fields different from computer science also suggests the need for economic and social models that let us evaluate the impact poor-quality Web data can have on specific phenomena. For instance, from a social perspective, it’s interesting to study the usage of Web-available health data. Indeed, people are starting to have a high degree of confidence in such data, making the Web an increasingly important source for health information. This special issue will hopefully contribute to the emerging field’s advancement, and we look forward to the innovative ideas to come.

Acknowledgments
We thank the authors, whose submissions made this special issue possible and showed how active and exciting
this emerging field is, the reviewers who made the effort to read and offer constructive comments, enabling authors to see their work from a different perspective, the IEEE Internet Computing editorial staff members who helped us manage the submissions, and the magazine’s editorial board for its guidance in the process.

References

Elisa Bertino is a professor of computer science at Purdue University. She’s also the research director of the Center for Education and Research in Information Assurance and Security (CERIAS). Her research interests include data security and trustworthiness, digital identity management systems, and database systems. Bertino has a doctoral degree in computer science from the University of Pisa. She’s a fellow of IEEE and the ACM. Contact her via http://homes.cerias.purdue.edu/~bertino/.

Andrea Maurino is an assistant professor of computer science at the University of Milano Bicocca, Department of Informatics, Systems, and Communication. His research interests include methodology tools and techniques for data quality, nonfunctional properties in service-oriented architecture, and e-government applications. Maurino has a PhD from Politecnico di Milano. He’s a founder of the university spin-off NextTLab. Contact him at maurino@disco.unimib.it.

Monica Scannapieco is a researcher at the Italian National Institute of Statistics (Istat) and a lecturer at SAPIENZA — Università di Roma. Her research interests include data quality, privacy preservation, and data integration architectures. Scannapieco has a PhD in computer science and engineering from SAPIENZA — Università di Roma. Contact her at scannapi@istat.it.

Selected CS articles and columns are also available for free at http://ComputingNow.computer.org.

LISTEN TO GRADY BOOCH
“On Architecture”
podcast available at http://computingnow.computer.org