Editorial

The term *provenance* refers, broadly, to information about the origin, context, derivation, lineage, ownership, or history of some artifact. The provenance of data is more specifically a form of structured metadata that records the activities involved in data production. Data provenance applies to a broad variety of data types, from database records, to scientific datasets, to business transaction logs, web pages, social media messages, and more. At the same time, different definitions and measures of quality apply to each of these data types in different domains. While the relationship between quality and quality assessment methods have been widely explored in this journal over the past few years, we felt that the relationship between provenance, the quality of data, and quality assessment methods deserved further study.

This special issue of *JDIQ* is an attempt to formally investigate the multiple facets of this relationship. The call for articles for this issue was an attempt to identify a set of key questions and broad areas around the concepts of provenance and data quality.

Quality through provenance: Is it possible to formalize the intuition that knowledge of the history of a dataset can be instrumental to assessing the quality of the underlying data? How general can such a formal model be? What would a formal definition of provenance-based information quality metrics look like?

Quality of provenance: Next, and complementary to the previous point, are there useful metrics and methods for describing and assessing the quality or information content of data provenance itself?

Quality of query results: Third, how can the theory of data provenance be used to better understand the results of a (relational) database query?

Quality improvement: Finally, can provenance be used to inform interventions to improve upon the quality of a dataset? Also, are there interesting application use cases associated with each of these questions?

In response to our call, we received nine submissions and selected four articles, following two rounds of highly qualified peer reviews. Three of the articles directly answer some of the previous questions. The first article presents a model for using provenance in data quality assessment, specifically of Linked Data. The second presents a framework for assessing the quality of provenance itself. The third includes a formal model for explaining missing answers from relational queries using data provenance. The fourth addresses a common infrastructure problem, namely, how to embed references to provenance within the data.

In the first article in this collection, **QUAL: A Provenance-Aware Quality Model**, Baillie, Edwards, and Pignotti address one of the core themes of this special issue, namely, how knowledge of the provenance of a dataset can help assess the quality of that dataset. Their contribution is a data quality assessment model, Qual-DM, and associated software framework, which relies on a variety of contextual metadata around the data—including provenance. Furthermore, the software generates the provenance of the quality assessment activity itself.

Qual-DM is aligned with and extends PROV, the W3C recommendation for a conceptual model of provenance. Qual-DM concepts are formally defined as classes in the

OWL2 Description Logic, and inference rules are introduced to describe how certain concepts can be derived from others. The Qual-O operational model, an extension of the W3C PROV-O ontology, realizes the formal model and provides a concrete foundation for the Qual quality assessment service. The service functionality is evaluated on datasets from two case studies, namely, for a public transport passenger information service and for an “invasive species monitoring” scenario. The authors also show that the functionality of the Qual model compares favorably with those of other quality assessment models from the literature and present performance data on the OWL reasoning tasks associated with the assessment.

In the next article, **Provenance Quality Assessment Methodology and Framework**, Cheah and Plale address a complementary problem. The authors note that assessing the quality of provenance itself should be a primary concern, if provenance is to be used reliably in data quality assessment. Following this observation, the authors focus on provenance traces that describe the execution of complex, long-running, and possibly distributed processes. They observe that collecting provenance traces for these processes is error-prone, due to the distributed and possibly heterogeneous nature of the underlying computing infrastructure. Errors can be introduced in the provenance trace that describes complex, long-running processes at various stages during its lifetime, namely, during capture, storage, query, and “stitching.” The latter is the process of collecting multiple provenance fragments from independent systems and joining them together. When the traces are collected through unreliable message passing after process execution, provenance traces may also be incomplete. Finally, independently collected trace fragments may be *temporally inconsistent*, that is, relative to the constraints defined in the PROV-CONSTRAINT specification.

With this motivation, the article describes a framework for the analysis of the quality of the provenance itself, along three of the typical “dimensions” cited in information quality management: correctness (the extent to which provenance is error free), completeness (the extent to which the trace completely describes, as a set level of abstraction, the underlying process execution), and relevance. The latter is more of a subjective dimension having to do with the usefulness of the provenance trace given a target analytical task.

The article employs different techniques, mostly heuristic in nature, to detect *likely anomalies* in a collection of provenance fragments, to detect temporal constraint violations, and to estimate the completeness of a provenance graph. A score model that combines these estimates is then proposed as an overall measure of quality for a complex provenance graph.

The framework is demonstrated on one synthetic dataset, previously developed by the authors, as well as two real-world provenance datasets, a data-ingesting workflow from NASA (AMSR-E), and a previously published dataset for data integration in the tourism domain.

Our third question was how to use provenance to improve the quality of query results. This challenge is tackled by the third article, **A Hybrid Approach to Answering Why-Not Questions on Relational Query Results**. In this work, Melanie Herschel tackles the problem of explaining missing answers in results of queries on relational databases. While traditional lineage techniques provide an explanation for data that is in the query output (e.g., a set of justifying source tuples), missing-answer explanations provide explanations for data that is not in the output. This information is useful to improve the quality of the source data, as it can drive the users’ attention on the part of the data where the errors are most likely to be.

The author reviews related techniques by dividing them into two categories, those that are based on highlighting missing data (i.e., source data that is needed to make the missing answer appear) and those that propose query changes to make the missing

answer appear in the output. Since both have limitations with regard to completeness or conciseness, the proposed hybrid approach subsumes previous techniques by allowing highlighting both missing data and relevant parts of queries. This intuition is captured in a framework and the existing techniques are modeled as instances of the framework. The formalism makes use of conditional tuples (c-tuples) to keep track of conditions relevant to the execution of a query on a tuple, as it has proven to be effective for this line of research. The main algorithm computes hybrid explanations in four steps: computing a generic witness, annotating the witness to identify parts of the query that may block tuples, computing derivations, and computing explanations. An experimental study shows that the proposed cost metric helps in identifying the explanations that are valuable for human interpretation, and that the algorithms have competitive runtime compared to the state-of-the-art systems.

Self-Identifying Data for Fair Use explores the association between data and its provenance. Chong, Skalka, and Vaughan start from the observation that research data repositories routinely state some kind of fair use policy on the datasets they host. However, they report on anecdotal evidence of routine violations of fair use policy, despite research scientists being generally well intentioned. The authors argue that ensuring a persistent coupling between datasets and their provenance information, during normal scientific usage of the data, may help alleviate the problem.

Their approach to establishing such a permanent and reliable association involves a new watermarking technique, whereby a short metadata payload is embedded within the dataset.

Watermarking has been widely used to facilitate copyright protection, typically on images, using steganography techniques: an invisible signature is embedded in the main image, in such a way that (1) it can be retrieved, provided one has access to details of the algorithm used for the embedding, and (2) it cannot otherwise be easily detected or erased.

It is well known that watermarking inevitably alters the data. In typical imaging applications, this is normally acceptable. With scientific data, however, any change to the data may render it unusable or invalid for experimental purposes. Thus, the main technical challenge faced in this research involves finding a watermarking scheme that does not interfere with normal scientific use. The main insight is to use the least significant bits of the numerical data in the file in order to minimize the effect of data corruption when the data is used in calculations. The approach is tested and validated on a real environmental research dataset.

The embedding of provenance metadata is only one of the many techniques made possible by this general approach. One limitation is the size of the metadata payload that can be embedded, however. Thus, an ideal application scenario for this technique involves storing the provenance trace itself in a separate repository and embedding a reference to the trace inside the data.

The systematic collection, analysis, and usage of data provenance are only now entering the mainstream, bringing both opportunities and new challenges. We highlight two such challenges. The first relates to an understanding of the cost and benefit framework around the collection of provenance data. The cost of collecting provenance data depends both on the characteristics of the data (its structure, size, heterogeneity) and on the data management environment that is responsible for its production and consumption (e.g., automated vs. manual data processing, structured queries vs. search interfaces, etc.). A comprehensive framework must be developed to investigate the tradeoffs between the cost of collecting and managing provenance and its benefits vis-à-vis quality assessment and improvement.

The second challenge revolves around novel applications of data provenance. While the primary motivation to obtain provenance data was to improve the quality of query

results, we are witnessing novel practices, such as providing clear and concise *explanations for problems*. An example is using provenance in the case where the result of a query includes values that appear to be outliers; the explanation may involve an unusually high price for specific items. The data quality tools that are available to detect errors in the data are not adequate to provide an explanation for these occurrences, but provenance-enabled systems today can compute their lineage across data sources. Can these systems be extended to provide more useful explanations of those errors? An example may be the ability to identify processes or employees that are more likely to produce erroneous data values. This kind of concise description would provide better insight to the users but cannot be easily determined without access to provenance data. The opportunities for exploiting data provenance in this context have not been fully explored.

We wish to thank Editor in Chief Louiqa Raschid and Senior Associate Editor Felix Naumann for their support and guidance throughout this process, and Ms. Janet Cavanagh, administrative assistant for the journal, for her help in coordinating this special issue.

Finally, we thank the following reviewers who donated their time and offered their expertise to make this issue possible. Their contribution is greatly appreciated.

Khalid Belhajjame, University Paris - Dauphine
 Domenico Beneventano, Università degli Studi di Modena e Reggio Emilia
 Laure Berti-Equille, Qatar Computing Research Institute
 James Cheney, University of Edinburgh
 Victor Cuevas, Universidad de Guadalajara, Mexico
 Tom De Nies, Ghent University
 Helena Galhardas, University of Lisbon
 Ashish Gehani, SRI International
 Boris Glavic, Illinois Institute of Technology
 Paul Groth, Elsevier
 Olaf Hartig, University of Waterloo
 Scott Jensen, San Jose State University
 Andrea Maurino, Università degli Studi di Milano-Bicocca
 Gianni Mecca, Università della Basilicata
 Renee Miller, University of Toronto
 Luc Moreau, University of Southampton
 Mourad Ouzzani, Qatar Computing Research Institute
 Jaehong Park, The University of Texas at San Antonio

Dr. Paolo Missier
 Newcastle University

Dr. Paolo Papotti
 Qatar Computing Research Institute

Guest Editors