14th International Workshop on Quality in Databases: Preface

Lisa Ehrlinger
Hasso Plattner Institute
Potsdam, Germany
lisa.ehrlinger@hpi.de

Lorena Etcheverry
Universidad de la República
Uruguay
lorenae@fing.edu.uy

Hazar Harmouch
University of Amsterdam
Amsterdam, Netherlands
h.harmouch@uva.nl

ABSTRACT

Data quality has been a major concern of organizations for decades, leading to the introduction of standards and quality frameworks. Recent advances in artificial intelligence (AI), e.g., generative AI, have brought data quality (DQ) back into the spotlight. In enterprises, it is particularly important to build data ecosystems that can cope with the emerging challenges posed by AI-based systems. DQ has been tackled from different perspectives: the database community has made significant advances in data profiling and data cleaning and still focuses on DQ issues like duplicate detection or missing data handling; the information systems community provides solutions for addressing DQ at an organizational level; the machine learning (ML) community focuses mainly on the development of robust models that can deal with issues in the data.

We build upon the success of QDB'24[1] and QDB'23[2] and continue offering an open format for joint discussions between different communities on the future of DQ assessment and improvement. QDB'25 focuses on novel and practical data quality assessment and improvement methods in the era of Large Language Models (LLMs).

VLDB Workshop Reference Format:
Lisa Ehrlinger, Lorena Etcheverry, and Hazar Harmouch. 14th International Workshop on Quality in Databases: Preface. VLDB 2025 Workshop: 14th International Workshop on Quality in Databases (QDB'25).

1 MOTIVATION AND SCOPE

The workshop aims to provide a platform for researchers with different backgrounds to exchange their challenges, ideas, and solutions. The high interest and recent articles in the ACM Journal on Data and Information Quality (JDIQ) demonstrate the importance and timeliness of data quality research, but offer no room for discussion. The International Workshop on Quality in Databases (QDB) proved in the last two runs (2023 & 2024) to be a noteworthy success, attracting over 50 participants (see Section 4). Through QDB, we provide the possibility for the specific community to meet and exchange new ideas, especially PhD students and junior researchers who are unsure to which venue to go for meeting "their"community.

This workshop aims to exchange novel ideas and best practices about data quality assessment and improvement in the era of AI. The event should unite experienced and senior-level data quality researchers with junior researchers and PhD students. We specifically expect junior researchers to benefit, since they get to meet the community and continue high-quality research on data quality.

Topics of Interest

The topics include, but are not limited to:

Foundational DQ methods and assessment

- Data profiling for data quality measurement
- Statistical methods to detect erroneous data
- Data lineage and provenance tracking
- Industry-specific data quality standards and compliance
- Data versioning and quality control
- Benchmark data sets to evaluate DQ assurance methods

AI/ML-specific data quality

- Data preprocessing
- Data quality for foundation models
- Data quality using generative AI
- Bias detection and mitigation in training data
- Data quality for few-shot and zero-shot learning
- Post-training quality/fact checking
- FAIRness[3] in data quality
- Explainable data cleaning
- ML-powered methods for improving data quality

Implementation and process optimization

- Automation of DQ assessment and improvement methods
- Real-time data quality monitoring
- Cost-benefit analysis of data quality improvements
- Integration of data quality tools in MLOps pipelines

We appreciate submissions on all these topics for different domains (e.g., healthcare, mobility, production) and for various types of data (e.g., graphs, time series).

2 COMMITTEE

We are grateful for the support from the steering committee and the thorough work done by the program committee in assessing the quality of the submissions.

Steering Committee

- Sourav S Bhowmick (Nanyang Technological University, Singapore)
- Felix Naumann (Hasso Plattner Institute, University of Potsdam, Germany)
- Ihab Ilyas (University of Waterloo, Canada)

[1] https://hpi.de/naumann/s/qdb2024.html
[2] https://hpi.de/naumann/s/qdb2023.html

Proceedings of the VLDB Endowment. ISSN 2150-8097.

[3] According to the FAIR principles, see https://www.go-fair.org/fair-principles

Program Committee

- Cinzia Capiello (Politecnico di Milano, Italy)
- Reynold C. K. Cheng (University of Hong Kong, Hong Kong)
- Chang Ge (University of Minnesota, USA)
- Thomas Hütter (University of Salzburg, Austria)
- Christine Legner (University of Lausanne, Switzerland)
- Sebastian Link (University of Auckland, New Zealand)
- Adriana Marotta (Universidad de la República, Uruguay)
- Paolo Papotti (Eurecom, France)
- Eduardo Pena (UTFPR , Brazil)
- Maria Angela Pellegrino (University of Salerno, Italy)
- Elizabeth Pierce (University of Arkansas at Little Rock, USA)
- Veronika Peralta (University of Tours, France)
- Sebastian Schelter (TU Berlin, Germany)
- Giovanni Simonini (University of Modena and Reggio Emilia, Italy)
- Wolfram Wöß (Johannes Kepler University Linz, Austria)

3 WORKSHOP FORMAT

The full-day workshop on September 1st, 2025, will include the paper presentations, two invited keynotes, as well as an interactive poster session. The full program with updates and details is available on our website: https://qdb-workshop.github.io/.

Keynote Speakers

Renee Miller and Paolo Missier will share with us their experience in topics including data-centric AI, provenance and table search in semantic data lakes.

Accepted Papers

This year, we received a total of 15 submissions, all of which were regular research papers. One paper was desk rejected and seven of the remaining 14 reviewed papers were accepted, resulting in an acceptance rate of 50%.

- Out in the Wild: Investigating the Impact of Imperfect Data on a Tabular Foundation Model (Vasileios Papastergios and Anastasios Gounaris)
- Exploring Privacy-Preserving Record Linkage: A Holistic Framework for Dataset Generation and Detailed Result Analysis (Florens Rohde, Victor Christen, and Erhard Rahm)
- Dynamic Knowledge Graph-based Measurement of Data Quality (Johannes Schrott, Rainer Meindl, Christian Lettner, Stefan Hammer, and Magdalena Leitner)
- Evolving Gracefully: Building Robust and Self-Adaptive Data Cleaning Pipelines for Schema Evolution and Uncertainty (Kevin Kramer, Valerie Restat, and Uta Störl)
- Label Flipping For Group Fairness (Shashank Thandri and Romila Pradhan)
- PBE Meets LLM: When Few Examples Aren't Few-Shot Enough (Shuning Zhang and Yongjoo Park)
- Towards an SLM-based Auditing of Relational Schemas and Data Quality for Practical Data Governance (Antony de Medeiros)

4 HISTORICAL INFORMATION ABOUT QDB

We are building on an established tradition of thirteen previous international workshops concerning data and information quality. This section provides an overview of the previous workshops with respect to year, venue, affiliated event, chairs, and submissions. Considering the recent advances in AI-based systems, QDB'23 and QDB'24 revealed the interest of many researchers from different communities to exchange their challenges, use cases, and ideas on data quality. We therefore believe that the momentum of discussing DQ in the context of AI is high and requires a venue such as QDB'25.

In 2024, the 13th edition of the workshop was held in Guangzhou, China, co-located with VLDB 2024 and attracted around 45 participants. The workshop featured a keynote by Sebastian Schelter (TU Berlin), 8 paper presentations (out of 16 submissions; which doubled since 2023), and an industry session with two talks by Quanqing Xu (Oceanbase) and Divesh Srivastava (AT&T) as well as a panel discussion between the two speakers and Fatma Ozcan (Google) on DQ research at the intersection of academia and industry.
Chairs: Sourav S Bhowmick, Lisa Ehrlinger, Hazar Harmouch
Website: https://hpi.de/naumann/s/qdb2024.html

In 2023, the 12th edition of the workshop was held in Vancouver, Canada, co-located with VLDB 2023 and attracted over 50 participants. The focus was on data quality and data cleaning in the context of AI-based systems. The workshop featured two keynotes by Renée J. Miller and Theodoros Rekatsinas, five research paper presentations (out of 8 submissions), and a very engaging breakout session with moderators to discuss the topics of (1) interdependency between DQ and ML, (2) DQ benchmark datasets, (3) ontologies and standards for DQ, and (4) explainable DQ and data cleaning.
Chairs: Lisa Ehrlinger, Hazar Harmouch, Ihab Ilyas, Felix Naumann
Website: https://hpi.de/naumann/s/qdb2023.html

In 2016, the 11th edition of the workshop took place for the last time in Delhi, India, co-located with VLDB 2016. It had a special focus on problems related to Big Data Integration and Big Data Quality. The workshop accepted four our of 10 research papers.
Chairs: Christoph Quix, Rihan Hai, Hongzhi Wang, Verikat N. Gudivada, Laure Berti
Report: https://publications.rwth-aachen.de/record/680764

In 2012, the 10th edition of the QDB workshop took place in Istanbul, Turkey, co-located with VLDB 2012. The focus was on data quality and data cleaning in the area of rule mining and data linking using Wikipedia. QDB'12 attracted 39 participants and matched the high quality and good submission level of its predecessors.
Chairs: Xin Luna Dong, Eduard Constantin Dragut
Report: https://sigmodrecord.org/publications/sigmodRecord/1212/pdfs/11.report.dong.pdf

In 2011, the 9th edition of the QDB workshop took place in Seattle, US, co-located with VDLB 2011. The focus was on problems of assessing, monitoring, improving, and maintaining DQ.
Chairs: Mourad Ouzzani, Paolo Papotti, Erhard Rahm
Website: http://qdb2011.dia.uniroma3.it/index.html

In 2010, the 8th edition of the QDB workshop (QDB10) took place on September 13, 2010, in Singapore, co-located with VLDB 2010.

The focus was on data quality assessment, entity matching, and information overloading. The workshop accepted 9 out of 12 papers.
Chairs: Andrea Maurino, Cinzia Cappiello, Panos Vassiliadis, Kai-Uwe Sattler
Report: http://sigmod.org/publications/sigmodRecord/1112/pdfs/09.report.maurino.pdf

Earlier version of the QDB workshop were co-located with VLDB from 2007-2009.

CleanDB. The first International VLDB workshop on Clean Databases (CleanDB) was held in Seoul, Korea on September 11, 2006, co-located with VLDB 2006.
Website: https://pike.psu.edu/cleandb06/

From 2004-2006, a related workshop was termed IQIS and co-located with SIGMOD.