13th International Workshop on Quality in Databases: Preface

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

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

Sourav S Bhowmick
Nanyang Technological University
Singapore, Singapore
assourav@ntu.edu.sg

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 emerging challenges posed by AI-based systems. DQ is tackled from different perspectives by different research communities, including database, machine learning (ML), and information systems. We believe it is important to bring together these communities to foster a vital discussion about the future of DQ assessment and improvement.

Considering the large number of participants (>50) at QDB'23, QDB'24 aims to (1) continue to host the vital discussions about data quality, and (2) exchange best practices and novel methods for (semi-)automated (ML-based) data quality assessment and improvement in the context of AI-based systems. Since the workshop addresses the needs of the AI era, it is of interest to both industry and academia (exemplified by the data-centric AI trend) as reflected in the accepted papers and in the workshop format.

VLDB Workshop Reference Format:
Lisa Ehrlinger, Hazar Harmouch, and Sourav S Bhowmick. 13th International Workshop on Quality in Databases: Preface. VLDB 2024 Workshop: 13th International Workshop on Quality in Databases (QDB'24).

1 MOTIVATION AND SCOPE

The workshop aims to provide a platform for researchers with different kinds from background 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) held last year proved to be a noteworthy success, attracting over 50 participants (see Section 4). As QDB is specifically dedicated to the topic of data quality, it allows this very specific community to meet and exchange new ideas. The event brings together experienced and senior-level data quality researchers with junior researchers and PhD students. In the course of the 2024 edition of the workshop, we want to further foster the interaction between academia and industry with a dedicated real-world use cases session on data quality.

Proceedings of the VLDB Endowment. ISSN 2150-8097.

Topics of Interest

The topics include, but are not limited to:

- Data preprocessing
- Data profiling for data quality measurement
- Explainable data cleaning
- DQ requirements for generative AI systems
- DQ using generative AI
- Data quality assessment for AI-based systems
- DQ improvement / data cleaning for AI-based systems
- Benchmark data sets to evaluate DQ assurance methods
- Automation of DQ assessment and improvement methods
- Methods to scale data quality assessment and cleansing
- ML-powered methods for improving data quality
- Data quality in graph-structured or time-series data
- Metadata management to improve data quality
- Data quality in different data science domains
- Human-in-the-loop approaches for DQ
- Post-training quality / fact checking
- FAIRness in data quality

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

- Ihab Ilyas (Apple, University of Waterloo, USA)
- Felix Naumann (Hasso Plattner Institute, University of Potsdam, Germany)

Program Committee

- Ziawasch Abedjan (TU Berlin, Germany)
- Antoon Bronselaer (Ghent University, Belgium)
- Felix Biessmann (Einstein Center Digital Future, Germany)
- Ismael Caballero (University of Castilla La Mancha, Spain)
- Cinzia Capiello (Politecnico di Milano, Italy)
- Chang Ge (University of Minnesota, USA)
- Christine Legner (University of Lausanne, Switzerland)
- Sebastian Link (University of Auckland, New Zealand)
- Elizabeth Pierce (University of Arkansas at Little Rock, USA)
- Kai-Uwe Sattler (TU Ilmenau, Germany)
- Sebastian Schelter (University of Amsterdam, Netherlands)
- John Talburt (University of Arkansas at Little Rock, USA)
- Panos Vassiliadis (University of Ioannina, Greece)
- Wolfram Wöß (Johannes Kepler University Linz, Austria)

3 WORKSHOP FORMAT

The full-day workshop on August 26, 2024, will include the paper presentations, an invited keynote by Sebastian Schelter (TU Berlin), as well as an industry experience session with two invited speakers that will participate in a panel discussion about data quality in practice. The full program with updates and details is available on our website: https://hpi.de/naumann/s/qdb2024.html

Keynote Speaker

Sebastian Schelter (TU Berlin) will speak about "How Data Management Research Helps to Improve Real World ML Applications".

Accepted Papers

This year, we received in total 13 submissions (11 regular research papers and 2 demos) and received 3 additional submissions that where handed over from the DeCo (3rd International Workshop on Data Ecosystems) workshop. From these papers, we accepted five regular research papers, two demo submissions as well one DeCo workshop paper. In total, this resulted in an acceptance rate of 50%.

- Accelerating the Data Cleaning Systems Raha and Baran through Task and Data Parallelism (Fatemeh Ahmadi, Yusuf Mandirali, Ziawasch Abedjan)
- Towards Semi-Supervised Data Quality Detection In Graphs (Rubab Zahra Sarfraz)
- Valuation-based Data Acquisition for Machine Learning Fairness (Ekta and Romila Pradhan)
- AutoFAIR : Automatic Data FAIRification via Machine Reading (Tingyan Ma, Wei Liu, Bin Lu, Xiaoying Gan, Yunqiang Zhu, Luoyi Fu, Chenghu Zhou)
- Compute Engine Testing with Privacy-Compliant Production-Like Synthetic Data (Yu Liu, Jiangnan Cheng, Steve Chuck, Lyublena Antova, Yurgis Baykshtis, Matt David, Ge Gao, Mehrdad Honarkhah, Kuan-Sung Huang, Chen-Kuei Lee, Usman Muhammad, Shihao Peng, Andrii Rosa, Rebecca Schlussel, Michael Shang, Kelvin Silva, Brandon Vo, Zac Wen, Yihao Zhou)
- Process Model-based Access Control Policies for Cross-Organizational Data Sharing (Liam Tirpitz, Leon Gentges)
- Tracking Consistency over Data Streams with InkStream [Demo] (Samuele Langhi, Angela Bonifati, Riccardo Tommasini)
- A Data Generator to Explore the Interactions Between Concept Drifts and Anomalies [Demo] (Jongjun Park, Akanksha Nehete, Tammy Zeng, Fei Chiang)

Industry Experience Session

The industry session consists of two industry talks by Quanqing Xu (Oceanbase) and Divesh Srivastava (AT&T) and a panel discussion with Quanqing Xu, Divesh Srivastava, and Fatma Ozcan (Google) on DQ research in the intersection between academia and industry.

4 HISTORICAL INFORMATION ABOUT QDB

We are building on an established tradition of twelve 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 in Vancouver 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'24.

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.