Editorial

Special Issue on Human in the Loop Data Curation

Demartini, Gianluca; Sadiq, Shazia; Yang, Jie

DOI
10.1145/3650209

Publication date
2024

Document Version
Final published version

Published in
Journal of Data and Information Quality

Citation (APA)
Demartini, G., Sadiq, S., & Yang, J. (2024). Editorial: Special Issue on Human in the Loop Data Curation. *Journal of Data and Information Quality*, *16*(1), 1-2. Article 3. https://doi.org/10.1145/3650209

Editorial: Special Issue on Human in the Loop Data Curation

GIANLUCA DEMARTINI, The University of Queensland, St Lucia, Australia
SHAZIA SADIQ, The University of Queensland, St Lucia, Australia
JIE YANG, TU Delft, Delft, Netherlands

This Special Issue of the Journal of Data and Information Quality (JDIQ) contains novel theoretical and methodological contributions on data curation involving humans in the loop. In this editorial, we summarize the scope of the issue and briefly describe its content.

CCS Concepts: • **Information systems → Data cleaning;**

Additional Key Words and Phrases: Data quality, data curation, crowdsourcing, human in the loop

ACM Reference Format:
Gianluca Demartini, Shazia Sadiq, and Jie Yang. 2024. Editorial: Special Issue on Human in the Loop Data Curation. *ACM J. Data Inform. Quality* 16, 1, Article 3 (March 2024), 2 pages. https://doi.org/10.1145/3650209

Although data quality is a long-standing and enduring problem, data quality problems have recently received a resurgence of attention due to the fast proliferation of data analytics, machine learning, and decision-support applications built upon the wide-scale availability and accessibility of (big) data. Particularly, the success of machine learning heavily relies on the quality and quantity of training data. Data curation which may include ingestion, annotation, cleaning, integration, and so on, is a critical step to provide adequate assurances on the quality of analytics and machine learning results. Such data preparation activities are recognized as time and resource intensive for data scientists as data often comes with a number of challenges that need to be tackled before it can be used in practice. Data re-purposing and the resulting distance between design and use intentions of the data, is a fundamental issue behind many of these challenges. These challenges include a variety of data issues such as noise and outliers, incompleteness, representativeness or biases, heterogeneity of format or semantics, and the like. Mishandling these challenges can lead to negative and sometimes damaging effects, especially in critical domains like healthcare, transport, and finance.

We present *Information Resilience* [1] as a means through which data pipelines can be protected from failures and risks arising from such challenges across the various tasks where data is sourced, shared, transformed, analysed, and consumed. An observable distinct feature of Information Resilience in these contexts is the increasingly important role played by humans, being often the

This work is partially supported by the Australian Research Council (ARC) Training Centre for Information Resilience (Grant No. IC200100022).
Authors' addresses: G. Demartini and S. Sadiq, The University of Queensland, St Lucia, QLD, Australia, 4072; e-mails: demartini@acm.org, shazia@eecs.uq.edu.au; J. Yang, TU Delft, Delft, Netherlands, 2628 XE Van Mourik Broekmanweg 6; e-mail: J.Yang-3@tudelft.nl.

ACM 1936-1955/2024/03-ART3
https://doi.org/10.1145/3650209

source of data creation and active players in data curation. This special issue looks at the interdisciplinary overlap between manual, automated, and hybrid human-machine methods of data curation. The need for new research effort on involving humans in the loop of the data curation process is exacerbated by the importance of developing methods that can scale to large amounts of data while also maintaining a human touch. This means designing processes that can deliver a high level of transparency in the data curation process (e.g., explaining why certain values have been dropped), deal with ethical data challenges like the decision to use or discard certain attributes (e.g., applicants' gender) in decision making processes, and, overall, increase the quality and trust in the outcome.

1 ARTICLES INCLUDED IN THIS SPECIAL ISSUE

This special issue contains three articles that provide a broad and diverse contribution to the field of human in the loop data curation.

The article "Enhancing Human-in-the-Loop Ontology Curation Results through Task Design" focuses on improving the quality of ontology and on the curation process required to achieve this goal. Authors focus on the task of verification of ontology restrictions, which is addressed by means of human computation. The novel contributions of this work include task design guidelines for this sub-domain of data curation including preferred ways to present the ontology and to perform qualification tests with the crowd.

The article "Validating Synthetic Usage Data in Living Lab Environments" focuses on the use of log data from end users as data to perform system effectiveness evaluations. Authors show that click models and their reliability and robustness depend on the underlying data quality and quantity. Their results show that the more complex the model, the more log data is required.

The article "Cleenex: Support for User Involvement during an Iterative Data Cleaning Process" proposes a framework for iterative data cleaning where users are performing data cleaning tasks. Authors evaluate their proposed framework both by means of simulations as well as involving real users. Their experimental results show that the activity of the end users is significantly reduced thanks to the use of the proposed tools.

REFERENCE

[1] Shazia Sadiq, Amir Aryani, Gianluca Demartini, Wen Hua, Marta Indulska, Andrew Burton-Jones, Hassan Khosravi, Diana Benavides-Prado, Timos Sellis, Ida Someh, Rhema Vaithianathan, Sen Wang, and Xiaofang Zhou. 2022. Information resilience: The nexus of responsible and agile approaches to information use. *The VLDB Journal* (2022), 1–26.

Received 22 February 2024; accepted 27 February 2024