Editorial: Special Issue on Quality Aspects of Data Preparation

MARCO CONSOLE and MAURIZIO LENZERINI, University of Rome, Sapienza, Italy

This Special Issue of the Journal of Data and Information Quality (JDIQ) contains novel theoretical and methodological contributions as well as state-of-the-art reviews and research perspectives on quality aspects of data preparation. In this editorial, we summarize the scope of the issue and briefly describe its content.

Additional Key Words and Phrases: Data quality, data preparation, ontologies

ACM Reference format:
Marco Console and Maurizio Lenzerini. 2023. Editorial: Special Issue on Quality Aspects of Data Preparation. *ACM J. Data Inform. Quality* 15, 4, Article 40 (October 2023), 2 pages.
https://doi.org/10.1145/3626461

Data preparation is the process of gathering, transforming, and cleaning raw data prior to processing and analysis. It is an important first step in any data engineering and data science project, and it involves a diverse host of activities, both conceptual and practical. These activities include but are not limited to, understanding, collecting and reformatting data, aggregating, combining, and enriching information coming from autonomous data sources, and making modifications and corrections in order to meet quality standards in the target information system.

To its core, data preparation is closely related to data quality, as its aim is to ensure levels of quality of data that are acceptable for the task at hand. However, despite this close connection, a systematic attempt to relate data quality and data preparation has yet to appear in the literature.

This special issue collects recent theoretical and methodological contributions to the field of data quality that are connected to data preparation activities as well as state-of-the-art reviews and research perspectives. The aim of the issue is to pave the way for novel approaches to data preparation that are grounded in the data quality framework.

1 ARTICLES INCLUDED IN THIS SPECIAL ISSUE

This special issue contains four articles that provide a broad and diverse contribution to the field of data quality for data preparation.

The article **"Completing and Debugging Ontologies: State-of-the-Art and Challenges in Repairing Ontologies"** focuses on the problem of identifying and removing errors in conceptual specifications, an important component of many data preparation processes. Specifically, the article focuses on ontologies, i.e., conceptual specifications written in terms of logical axioms, and assumes a workflow divided into two steps: identifying errors and repairing errors. In this context, the article presents a thorough review of techniques for repairing ontologies that are based on abductive reasoning and discusses open problems and research perspectives in the field.

The article **"A Method to Screen, Assess, and Prepare Open Data for Use"** discusses data preparation methodologies that are specifically targeted to open data, i.e., data made available

1936-1955/2023/10-ART40
https://doi.org/10.1145/3626461

online by public or private organizations under open licenses. Open data sources form an important source of information in several data analysis tasks but, due to their nature, are prone to error and incompleteness. In this context, the article presents a thorough data preparation methodology tailored to open data used in enterprise settings. Moreover, the authors conceptualize the process of data preparation for open data in terms of a value-creating process.

The article **"Pipeline Design for Data Preparation for Social Media Analysis"** studies the problems connected to data preparation with a specific focus on social media data, one of the major sources of information in several engineering and analysis tasks. Extracting information from social media is often challenging due to its heterogeneous nature, often consisting of multimedia content, and its dependence on the context of creation and use. In this context, the article proposes a systematic approach to designing data preparation pipelines for social media data. These pipelines allow the use of metadata for performances and quality information and support the automatic generation of provenance metadata for their results. Finally, the article presents the use-case of a pipeline for image extraction in social media. The aim of this pipeline is to create a dataset to analyze behavioural aspects of the COVID-19 pandemic.

The article **"A data centric AI framework for automating exploratory data analysis and data quality tasks"** studies methodological aspects of data preparation workflows related to Artificial Intelligence systems, one of the most widespread application scenarios for data preparation techniques. Specifically, the article discusses open areas and pain points of these workflows and presents the results of a field study based on user interviews carried out in real-world corporate environments. Starting from the results of this study, the authors propose a framework and a set of algorithms to perform exploratory data analysis for Artificial Intelligence systems. Finally, the article presents concrete implementations of these algorithms and discusses the results of a thorough experimental evaluation of these implementations.