Published: 11 December 2025
Accepted: 24 November 2025
Revised: 17 November 2025
Received: 17 November 2025

INTRODUCTION

Editorial: Special Issue on Data quality dimensions in Data FAIRification design and processes

ANNA BERNASCONI, Politecnico di Milano, Milan, MI, Italy

Anna Bernasconi is a postdoctoral researcher with the Dipartimento di Elettronica, Informazione e Bioingegneria at the Politecnico di Milano (Italy). She received her Ph.D. in Information Technology from Politecnico di Milano in February 2021. Her main research interest are Bioinformatics, Databases, and Search Systems dedicated to biological and genomic data. More specifically, during her Ph.D. research, she has focused on the modeling and integration of data and metadata of tertiary analysis human genomic datasets. The proposed methodologies are particularly useful to support complex biological query answering and semantic search functionalities. Her results have also been successfully applied to the domain of viral sequences, hence demonstrating the flexibility of the approach and notably contributing to the data-driven fight against the COVID-19 pandemic.

STEFANO CIRILLO, University of Salerno, Salerno, SA, Italy

ALBERTO GA S, Polytechnic University of Valencia, Valencia, Valencia, Spain

HAZAR HARMOUCH, University of Amsterdam, Amsterdam, Noord-Holland, Netherlands

Open Access Support provided by:

University of Salerno

University of Amsterdam

Politecnico di Milano

Polytechnic University of Valencia

Editorial: Special Issue on Data Quality Dimensions in Data FAIRification Design and Processes

ANNA BERNASCONI, Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milano, Italy
STEFANO CIRILLO, Department of Computer Science, University of Salerno, Fisciano, Italy
ALBERTO GARCÍA S., Valencian Research Institute for Artificial Intelligence, Universitat Politecnica de Valencia, Valencia, Spain
HAZAR HARMOUCH, Informatics Institute, University of Amsterdam, Amsterdam, Netherlands

This editorial summarizes the content of the Special Issue on Data quality dimensions in Data FAIRification design and processes of the Journal of Data and Information Quality (JDIQ).

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

Additional Key Words and Phrases: FAIR principles, findability, accessibility, interoperability, reusability, data quality, data curation, data repositories, fair data-systems, fair tools, fair assessment, fairification processes

ACM Reference Format:
Anna Bernasconi, Stefano Cirillo, Alberto García S., and Hazar Harmouch. 2025. Editorial: Special Issue on Data Quality Dimensions in Data FAIRification Design and Processes. *ACM J. Data Inform. Quality* 17, 4, Article 20 (December 2025), 5 pages. https://doi.org/10.1145/3778866

Introduction

The FAIR Data Principles, introduced in 2016 in the flagship paper by Wilkinson et al. [16], provide guidelines for metadata, data, and infrastructures to ensure that these are Findable, Accessible, Interoperable, and Reusable. The process of data FAIRification encompasses a series of activities, such as data modeling, cleaning, profiling, integration, preparation, and engineering, aimed at transforming unFAIR data into FAIR-compliant data. Depending on the structure and quality of data sources, FAIRification may involve a wide range of tasks, such as the development of semantic models, the design of metadata schemas, data linking, data analysis, and implementation/deployment of resources for data access.

The data property of being "FAIR" can be assessed at multiple levels of data granularity, spanning repositories, data sources, data lakes, datasets, data sheets, data models, and even individual data items. Moreover, the principles apply across numerous scientific domains, including the natural sciences (e.g., biomedicine, chemistry, astronomy, agriculture, earth and life sciences),

Authors' Contact Information: Anna Bernasconi (corresponding author), Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milano, Italy; e-mail: anna.bernasconi@polimi.it; Stefano Cirillo, Department of Computer Science, University of Salerno, Fisciano, Campania, Italy; e-mail: scirillo@unisa.it; Alberto García S., Valencian Research Institute for Artificial Intelligence, Universitat Politecnica de Valencia, Valencia, Valencian Community, Spain; e-mail: algarsi3@pros.upv.es; Hazar Harmouch, Informatics Institute, University of Amsterdam, Amsterdam, Netherlands; e-mail: h.harmouch@uva.nl.

ACM 1936-1955/2025/12-ART20
https://doi.org/10.1145/3778866

engineering, as well as the humanities and social sciences. In many research contexts, adherence to FAIR principles has become not only a hallmark of high-quality data but also a prerequisite for publication or funding. Here, we shall note that, in the information sciences community, there exists another concept of "fairness" [4], which is more concerned with Machine Learning and should not be confused with the property of FAIR data.

The close relationship between *data FAIRification* and *data quality* provides a rich foundation for further research. It inspires the development of new methods, tools, and best practices to effectively merge FAIR principles with advanced data quality frameworks, thereby strengthening the reliability, credibility, and usability of data assets in data ecosystems. This special issue brings together recent progress in both the theory and practice of data quality within the broader context of data FAIRification (the process through which data is made FAIR) and FAIRness (the measure of how much data is FAIR). While several aspects regarding FAIR have been analyzed in recent works, including their compliance in datasets [12], computational workflows [17], ecosystems [8], a focus on data quality has not been proposed yet. This special issue explores this new space and opens up to a community that is less used to talking about FAIR (typically employed in the life sciences [14, 15]).

Articles Included in the Special Issue

This special issue collected recent advances in exploring FAIR dimensions with a data quality perspective. A total of 10 submissions (8 technical papers, 1 survey, and 1 resource paper) were accepted from 21 submitted articles. We present the articles divided into three main groups: the first comprises works that focus on a specific principle (Interoperability or Reusability); the second includes works that describe architectures for managing FAIR data; the third contains works on the assessment of (non)FAIR-compliant data.

FAIR principles. The article "A principled approach to validation of persistent identifiers", by Bronselaer [2], proposes a formal framework to validate persistent identifiers in FAIR data pipelines, addressing Interoperability issues caused by incorrect identifier usage (with a strong link to Findability and Accessibility as well). The approach introduces *group expressions* (i.e., formulas using capture groups from regular expressions) to enable robust validation. The author develops a REST-based validation server implementing this framework and evaluates it on three real-world datasets. Results demonstrate that the method scales efficiently to millions of instances, offering a practical and reliable solution for ensuring identifier consistency in FAIR data management.

The article "Ontology-Based Schema-Level Data Quality: The Case of Consistency", by Cima et al. [5], presents a novel data quality framework grounded in Ontology-Based Data Management (OBDM), specifically designed for metadata analysis. Focusing on the Data Quality "Consistency" dimension, the authors propose techniques to assess the alignment between database integrity constraints and ontological (i.e., Interoperability-related) knowledge. The framework enables the evaluation of whether such constraints are both sufficiently protective and appropriately faithful to the encoded domain knowledge. Furthermore, the paper investigates the related computational problems, providing a detailed complexity analysis and proving decidability for widely used classes of OBDM specifications and constraints.

The article "Sustainable quality in data preparation", by Pernici et al. [6], explores data preparation sustainability through the lens of the circular economy, emphasizing practices that reduce waste, extend data life cycles, and promote reuse in line with the Reusability FAIR principles. It highlights the challenges of maintaining high data and metadata quality amid scalability and resource constraints. The authors propose an evaluation model that combines data quality metrics with sustainability parameters for both human and computational processes. Applied in a comparative study of data preparation methods, the model shows its value in balancing quality and sustainability within modern data pipelines.

FAIR architectures. The article "The BigFAIR Architecture: Enabling Big Data Analytics in FAIR-compliant Repositories", by Castro et al. [3], introduces BigFAIR, a novel architecture designed to enable large-scale, FAIR-compliant management of scientific data and metadata. Addressing the challenges of volume, variety, and velocity in research data, BigFAIR separates infrastructures for data and metadata to enhance flexibility, maintain data ownership, and reduce development efforts. The authors define pipelines and guidelines for their implementation, alongside a generic metadata warehouse model supporting analytical query processing. A case study on two real-world datasets demonstrates BigFAIR's scalability and effectiveness in supporting big data analytics within FAIR repositories.

The article "xFAIR: A Multi-Layer Approach to Data FAIRness Assessment and Data FAIRification", by Longo et al. [10], presents xFAIR, a multi-layer platform architecture designed to assess and enhance data FAIRness across multiple domains. Addressing the gap between FAIR principles and their practical implementation, xFAIR integrates modules for data acquisition, FAIRness evaluation, and ontology support. Its applicability is demonstrated through three real-world use cases: public administration, European data portals, and news media, each emphasizing domain-specific quality improvements. The results highlight the value of combining automated metadata validation with community-driven feedback to strengthen data quality and sustain FAIRification efforts.

The article "A GenAI System for Improved FAIR Independent Biological Database Integration", by Sakib et al. [13], presents FAIRBridge, an AI-driven system that helps life science researchers query and integrate data from diverse biological databases, including those not fully FAIR-compliant. It uses natural language processing to interpret user intent, identify relevant sources, and generate executable queries automatically. FAIRBridge improves data accessibility, accuracy, and responsiveness through intelligent resource mapping and quality control. The system also supports crowd curation and exploration of alternative data sources. Overall, FAIRBridge streamlines scientific data discovery and hypothesis testing through an intuitive, automated interface.

The resource paper "An Extract, Transform, Load foundation for Biobank Data Interoperability", by Cruoglio et al. [7], introduces a framework for converting biobank datasets into HL7-FHIR transaction bundles, enabling seamless connection to the BBMRI-ERIC Federated Search platform. The approach simplifies data harmonization and FAIR compliance by automating the extraction, transformation, and loading of data from local systems. Implemented through a Python-based toolkit, it generates FHIR-ready JSON files for integration into federated networks. Successfully adopted by three BBMRI.it biobanks, the framework enhances interoperability and supports standardized, privacy-preserving data sharing across research infrastructures.

FAIR assessment. The article "Towards Realistic Error Models for Tabular Data", by Jung et al. [9], tackles the challenge of understanding and modeling data errors in modern data pipelines, particularly their impact on downstream tasks such as machine learning. Addressing the lack of FAIR-compliant practices and the oversimplification of existing error models, the authors propose a latent factor model for tabular data errors that captures complex dependencies among error types and distributions. The model separates error sampling from error types, allowing flexible extensions and realistic simulations. Extensive experiments on data cleaning and ML benchmarks show that this approach effectively models real-world error behavior and enhances the reproducibility and robustness of data quality research.

The article "Quality Dimensions and Evaluation Framework for Machine-Actionable DMPs", by Arnhold et al. [1], presents a conceptual framework for the automated evaluation of machine-actionable Data Management Plans (maDMPs), emphasizing FAIR assessment and funder compliance. Building on community input and prior work, the authors define a taxonomy of evaluation goals, dimensions, and metrics, and introduce the DMP Quality Vocabulary (DMPQV) to standardize

quality measurements. A prototype implementation, validated through a case study with the Science Europe Practical Guide, extends the RDA maDMP standard to support funder-independent metrics such as completeness, feasibility, and compliance. The results show that automated evaluations can align with manual assessments and promote consistent, transparent DMP quality reporting.

The article "FAIRness of the Linguistic Linked Open Data Cloud: an Empirical Investigation", by Pellegrino et al. [11], assesses the compliance of Linguistic Linked Open Data (LLOD) datasets with the FAIR principles to evaluate their overall quality. Through a systematic literature review of 69 datasets published between 2014 and 2024, the authors employ the KGHeartBeat framework to analyze FAIR alignment across key data quality dimensions. Results show that LLOD datasets remain only partially findable and accessible, with limited interlinking and insufficient open licensing hindering reuse. The paper introduces a novel mapping between FAIR principles and data quality dimensions, offering a structured basis for improving LLOD accessibility and interoperability.

Acknowledgments

We are deeply grateful to all the authors for their outstanding contributions and to all the reviewers for their thorough work. We also thank Andrea Marrella (JDIQ Information Director) and Felix Naumann (JDIQ Editor-in-Chief) for their invaluable assistance.

References

[1] Lukas Arnhold, Tomasz Miksa, and Moritz Staudinger. 2025. Quality dimensions and evaluation framework for machine-actionable DMPs. *ACM Journal of Data and Information Quality* (Nov. 2025). DOI: 10.1145/3776555. Just Accepted.

[2] Antoon Bronselaer. 2025. A principled approach to validation of persistent identifiers. *ACM Journal of Data and Information Quality* (June 2025). DOI: 10.1145/3743145. Just Accepted.

[3] João Pedro de Carvalho Castro, Lucas Medeiros França Romero, Anderson Chaves Carniel, and Cristina Dutra Aguiar. 2025. The BigFAIR architecture: Enabling big data analytics in fair-compliant repositories. *ACM Journal of Data and Information Quality* (Nov. 2025). DOI: 10.1145/3774755. Just Accepted.

[4] Simon Caton and Haas Christian. 2024. Fairness in machine learning: A survey. *ACM Computing Surveys (CSUR)* 56, 7 (2024), 166. DOI: 10.1145/3616865

[5] Gianluca Cima, Marco Console, and Maurizio Lenzerini. 2025. Ontology-based schema-level data quality: The case of consistency. *ACM Journal of Data and Information Quality* (Oct. 2025). DOI: 10.1145/3770750. Just Accepted.

[6] Barbara Pernici, Cinzia Cappiello, Carlo Alberto Bono, Camilla Sancricca, Tiziana Catarci, Marco Angelini, Matteo Filosa, Matteo Palmonari, Flavio De Paoli, Sonia Bergamaschi, Giovanni Simonini, Angelo Mozzillo, and Luca Zecchini. 2025. Sustainable quality in data preparation. *J. Data and Information Quality* (October 2025). https://doi.org/10.1145/3769120

[7] Antonella Cruoglio, Federica Rossi, Davide Fragnito, Ramona Palombo, Alice Massacci, Martina Betti, Mattia D'Antonio, Massimiliano Borsani, Claudia Miele, Gennaro Ciliberto, et al. 2025. An extract, transform, load foundation for biobank data interoperability. *ACM Journal of Data and Information Quality* (Oct. 2025). DOI: 10.1145/3769117. Just Accepted.

[8] Daniel Jacob, François Ehrenmann, Romain David, Joseph Tran, Cathleen Mirande-Ney, and Philippe Chaumeil. 2025. An ecosystem for producing and sharing metadata within the web of FAIR data. *GigaScience* 14 (2025), giae111. DOI: 10.1093/gigascience/giae111

[9] Philipp Jung, Sebastian Jäger, Nicholas Chandler, and Felix Biessmann. 2025. Towards realistic error models for tabular data. *ACM Journal of Data and Information Quality* (Nov. 2025). DOI: 10.1145/3774914. Just Accepted.

[10] Antonella Longo, Marco Zappatore, Davide Colella, Francesca Zampino, and Antonella Calò. 2025. xFAIR: A multi-layer approach to data fairness assessment and data fairification. *ACM Journal of Data and Information Quality* (Oct. 2025). DOI: 10.1145/3769113. Just Accepted.

[11] Maria Angela Pellegrino, Pasquale Esposito, and Gabriele Tuozzo. 2025. FAIRness of the linguistic linked open data cloud: An empirical investigation. *ACM Journal of Data and Information Quality* (Oct. 2025). DOI: 10.1145/3769116. Just Accepted.

[12] Núria Queralt-Rosinach, Rajaram Kaliyaperumal, César H. Bernabé, Qinqin Long, Simone A. Joosten, Henk Jan van der Wijk, Erik L. A. Flikkenschild, Kees Burger, Annika Jacobsen, Barend Mons, et al. 2022. Applying the fair principles to data in a hospital: Challenges and opportunities in a pandemic. *Journal of Biomedical Semantics* 13, 12 (2022). DOI: 10.1186/s13326-022-00263-7

[13] Syed N. Sakib, Kallol Naha, Sajratul Y. Rubaiat, and Hasan M. Jamil. 2025. A GenAI system for improved fair independent biological database integration. *ACM Journal of Data and Information Quality* (Oct. 2025). DOI: 10.1145/3770753. Just Accepted.

[14] Zbyslaw Sondka, Nidhi Bindal Dhir, Denise Carvalho-Silva, Steven Jupe, null Madhumita, Karen McLaren, Mike Starkey, Sari Ward, Jennifer Wilding, Madiha Ahmed, et al. 2024. COSMIC: A curated database of somatic variants and clinical data for cancer. *Nucleic Acids Research* 52, D1 (2024), D1210–D1217. DOI: 10.1093/nar/gkad986

[15] Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, Jantina De Vries, Yukinori Okada, Alicia R. Martin, Hilary C. Martin, Tuuli Lappalainen, and Danielle Posthuma. 2021. Genome-wide association studies. *Nature Reviews Methods Primers* 1, 59 (2021). DOI: 10.1038/s43586-021-00056-9

[16] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, et al. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific Data* 3, 160018 (2016). DOI: 10.1038/sdata.2016.18

[17] Sean R. Wilkinson, Meznah Aloqalaa, Khalid Belhajjame, Michael R. Crusoe, Bruno de Paula Kinoshita, Luiz Gadelha, Daniel Garijo, Ove Johan Ragnar Gustafsson, Nick Juty, Sehrish Kanwal, et al. 2025. Applying the fair principles to computational workflows. *Scientific Data* 12, 328 (2025). DOI: 10.1038/s41597-025-04451-9

Received 17 November 2025; revised 17 November 2025; accepted 24 November 2025