ISSN 2281-5147

EVALUATION AND HARMONIZATION OF DATA QUALITY CRITERIA: INSIGHTS FROM EXPERT INTERVIEWS FOR LEGAL APPLICATION

Louis Sahi*

Abstract

This article presents a framework for assessing data quality, highlighting its growing importance in today's data-driven organisations. With the emergence of regulatory frameworks such as the GDPR and the EU's Open Data Regulation, the need for robust data quality standards has never been more important. The study begins by addressing the inconsistencies in current data quality criteria (DQCs) and proposes a unified list, derived from an extensive literature review. By aligning these standards with the broader context of data processing, including governance and lifecycle management, the research aims to create a coherent approach to data quality. Expert interviews were conducted with data management and legal professionals to validate the framework. This involvement not only consolidates the DQCs, but also ensures their compliance with EU regulations. The findings underline the need for collaborative data processing (CDP) in decentralised environments, such as the European Common Data Spaces, and highlight the importance of trust, legal compliance and reliability of shared data. Ultimately, this research contributes to bridging the gap between academic methodologies and practical industrial applications of data quality assessment, fostering a more secure and efficient data landscape.

*Louis Sahi, IRIT, Université de Toulouse, UT3, Toulouse, France, sahilouis@gmail.com
Louis graduated in Cybersecurity (Master) from Côte d'Ivoire. He has a bachelor's degree in Network and Computer Sciences and experience in the industry in Cote d'Ivoire and Morocco. This work is supported by the European Union's funded project Legality Attentive Data Scientists (LeADS) under Grant Agreement no. 956562.

ISSN 2281-5147

Table of Contents

ISSN 2281-5147

Keywords

Data Quality Criteria – Collaborative Data Processing – Trust– Reliability – Compliance

1. Introduction: Unlocking the Potential of Data, Ensuring Quality for Competitive Advantage

In today's economy, data is a valuable asset that is rapidly being integrated into business processes across all sectors. For data-driven organisations, data isn't just useful, it's essential for developing innovative strategies and products that ensure competitive advantage (Hupperz et al. 2021). By collecting and analysing data, these organisations are able to make more informed decisions and respond efficiently to market changes (Fabijan et al. 2017). However, effective use of data remains a challenge, often due to the quality of the data itself.

The Cost of Poor Data Quality

Despite significant investments in data infrastructure, many organisations face problems due to poor data quality. This is more than just a technical issue: data quality affects productivity, decision-making and even financial performance (Haug, Zachariassen, et van Liempd 2011). As Gartner's Ted Friedman highlighted in 2018, organisations that go digital risk a crisis of trust in data, which can reduce business value and harm financial outcomes (Moore 2018). Reliable, high-quality data is critical to maintaining trust in an organisation's information systems, particularly in light of regulatory requirements around data management.

ISSN 2281-5147

How to Define Data Quality?

The question of what makes data 'data quality' remains complex. To address this, we conducted a systematic review (Sahi et al. 2023) to explore academic perspectives on data quality criteria (DQC). Our findings show that there is no universal set of data quality standards, with different authors emphasizing different criteria and sometimes using inconsistent terms and definitions. From this literature review, we finally identified 30 essential DQCs and proposed a single definition for each, with the aim of standardising the assessment of data quality across systems. While this list is a valuable foundation, it needs to be validated in practice to bridge the gap between theory and practice. Furthermore, data quality standards need to be aligned with evolving regulations, such as the GDPR and the EU Open Data Regulation, highlighting the need for a compliance-focused approach to data quality.

Working With Data Management Experts

To ensure that our criteria are applicable in industry, we worked with data management experts, including legal experts, to review and assess the relevance of each DQC within the European regulatory framework. This collaboration provided a multi-faceted understanding of data quality and strengthened the practical relevance of our framework. Overall, this research highlights the importance of high quality data as the backbone of successful data-driven organisations. A systematic approach to data quality, informed by both academic research and practical insights, is essential for organisations seeking to remain competitive and compliant in an increasingly data-centric world.

Structure of the Article

The rest of this article is structured as follows. Section 2 outlines the motivation and contributions of the expert interviews. Section 3 describes interview preparation and participant selection. Section 4 presents the background and data management expertise of each interviewee. Section 5 summarises the feedback, definitions and

ISSN 2281-5147

relevance of the DQCs. Finally, section 6 provides a comprehensive analysis of the findings.

2. Motivation & Contribution: Ensuring Data Quality for Effective Data-Driven Decision Making and Compliance

2.1 Motivation: Contributing to Compliance in the European Data Space

With the European Union's initiative to establish *European Common Data Spaces*, ensuring the quality of shared data has become increasingly important. In a data space, data is managed at its source and shared only when needed, involving different stakeholders such as data providers, intermediaries and users. This collaborative model, known as collaborative data processing (CDP), spans the entire data lifecycle and emphasises community-driven interactions between users and systems (Gan et al. 2017).

However, CDP introduces specific challenges, such as:

1. **Compliance with Legal Requirements**: Data quality assurance strategies need to be aligned with regulatory frameworks, such as GDPR, to ensure secure and responsible data sharing.
2. **Trust in Decentralized Governance**: Building trust within a distributed system is essential for reliable data use and sharing.
3. **Reliability of Distributed Systems**: Ensuring that data remains accurate and consistent across a distributed infrastructure is critical.

2.2 The Challenge: Lack of Standardization in Data Quality Criteria

Many surveys and research studies have proposed various Data Quality Criteria (DQCs), each focusing on specific domains such as health information, information security and business performance. However, these studies vary in scope and context, resulting in different sets of DQCs - each with unique terminology, interpretations

ISSN 2281-5147

and criteria definitions. Our previous review highlighted a major issue: *there is no universal set of DQCs that can be applied across all domains*. In response, we conducted a systematic review (Sahi et al. 2023) to propose a comprehensive set of DQCs that can be applied across domains. However, bridging the gap between academic research and industry practice requires validation by data management experts from different sectors. In addition, collaborative data processing brings its own challenges. Ensuring data reliability in distributed systems, establishing trust within decentralised governance, and complying with legal requirements are essential for successful collaborative data processing (Jonathan et al. 2017).

2.3 Long term objective: Towards automated data quality assessment

A key goal of our work is to enable automated data quality assessment through standardised and unified DQCs. Automated assessment uses algorithms to evaluate data quality, eliminating the need for manual oversight. However, without consistent and universal DQCs, this vision remains out of reach. Previous research has classified DQCs based on data dimensions (Foote 2022), but few have explored the contextual aspects of data processing. Our framework aims to address this gap by including broader aspects such as data lifecycle, governance and regulatory requirements. This approach will provide a solid foundation for future automation efforts and improve data reliability, trustworthiness, and compliance across multiple domains.

2.4 Long Approach: Gathering Expert Insights to Refine DQCs

These challenges raise important questions about how DQCs can be formalised and enforced within a regulatory framework. While regulatory requirements vary by data type and use, no current framework analyses DQCs and data management regulations, such as GDPR or Open Data, together. This study fills this gap by developing a framework that aligns DQCs with regulatory requirements, with a focus on the evolving European data environment.

To address these challenges, we worked with data management experts with expertise in regulatory data governance. Through interviews, we gathered insights into relevant DQCs and refined our framework to align with key EU regulations. This study presents the findings of these regulatory data experts and identifies DQCs that incorporate critical regulatory requirements, ensuring that data quality in collaborative

ISSN 2281-5147

spaces meets both operational and regulatory standards. This research is a step towards a universal data quality framework that will serve as a foundation for future efforts to automate data quality assessment and compliance in Europe's collaborative data ecosystems.

3. Methodology: Consolidating Data Quality Insights from Legal Data Experts

To gather expert opinion on data quality criteria (DQCs), we conducted semi-structured interviews with data professionals specialising in data law and compliance. This process allowed us to consolidate their insights into a validated set of criteria. Here's how we approached this research:

1. **Formulating Relevant Questions**
2. **Selecting Qualified Data Professionals**
3. **Conducting the Interviews**
4. **Analyzing Results**

3.1 Step 1: Developing Key Questions

Our aim was to validate and refine a list of 30 DQCs, focusing on the relevance, thedefinition and the impact on trust, reliability and compliance of each criterion. To conduct effective interviews, we designed open-ended, semi-structured questions to encourage thoughtful, experience-based feedback. Key questions included:

- What aspects of data management have you explored?
- Have you encountered legal challenges in data management, like privacy or compliance issues?
- What kind of data do you work with, and how is it processed?
- What defines a "quality" dataset in your view?
- For each DQC, we asked:

ISSN 2281-5147

- How would you define this criterion?
- How should it be evaluated?
- Does it enhance trust, reliability, or legal compliance?

3.2 Step 2: Selecting Data Experts with Legal Backgrounds

The interviews were hosted by a partner organisation involved in the LeADS project. The Security and Technology Policy Director of the host organisation helped us to select participants based on their expertise and relevance to the study. Five professionals from different European organisations with expertise in data management, privacy and compliance were chosen:

- **Interviewee 1:** A Senior Legal Counsel focused on ensuring global data protection compliance within a major financial organization.
- **Interviewee 2:** A privacy expert in a semiconductor company, overseeing the processing of telemetry data to ensure regulatory adherence.
- **Interviewee 3:** Chief Officer of Privacy Protection in a consumer goods company, advising on GDPR compliance and privacy strategies.
- **Interviewee 4:** An academic specializing in data quality and governance, with a focus on regulatory challenges in AI and data integrity.

3.3 Step 3: Conducting the Interviews

The interviews were conducted via Webex between February 7 and 22, 2024. Each session, which lasted between 30 minutes and one hour, was recorded and supported by handwritten notes. Transcriptions were made to assist in the analysis, while maintaining strict confidentiality.

3.4 Step 4: Analysis and Results

The insights gathered from these professionals have been systematically analysed to identify key themes, validate the DQCs and align them with the practical compliance

ISSN 2281-5147

and governance requirements of the European regulatory landscape. This feedback is invaluable in refining a DQC framework that supports effective and compliant data management across multiple sectors. By consolidating expert feedback, this study advances the development of universally applicable data quality standards that prioritise trust, reliability and compliance.

4. Insights and Findings from Legal and Privacy Data Experts on Data Quality

4.1 Insights from Interviewee 1: Senior Legal Counsel in Financial Privacy

Our first expert, with a background in financial privacy, emphasised that high quality data is precisely tailored to its purpose and must be rigorously governed and protected. This respondent identified critical attributes such as granularity, relevance and security, as well as the importance of governing data through tagging and taxonomy for reliable use. Key DQCs identified include:

1. **Appropriate amount of data:** Data must contain the necessary attributes for accurate use, ensuring no excessive details that might clutter its intended purpose.
2. **Governance:** Effective governance is essential to ensure high-quality data, including protocols for maintaining integrity and usability.
3. **Understandability:** Data should be clearly labeled and tagged to facilitate accurate interpretation across contexts.
4. **Consistency:** Data should remain uniform across platforms to be correctly interpreted by all users.
5. **Currency:** The data is up-to-date.
6. **Timeliness:** Data must be available, accessible, and usable within required timeframes, ensuring timely actions and decisions.

ISSN 2281-5147

7. **Uniqueness:** Avoiding redundancy and duplication, ensuring each data element is unique and specific.
8. **Ease of manipulation:** High-quality data should be reusable and adaptable for various purposes without compromising its integrity.
9. **Free of error:** Data must be accurate and reliable, without distortions that might mislead its users.
10. **Integrity:** Data must be secured against unauthorized access, ensuring only designated individuals can access and modify it.
11. **Interpretability:** Data should be interpretable in ways that allow for multiple perspectives and uses.

These DQCs emphasize the importance of governance and clarity in data to support financial operations while safeguarding privacy and integrity.

4.2 Insights from Interviewee 2: Privacy and Legal Compliance Expert in Semiconductor Manufacturing

The second expert, from a semiconductor manufacturing background, highlighted the need for standardised data references across teams to avoid misunderstandings and compliance risks. This expert emphasised the value of metadata and advocated adaptable data frameworks that can accommodate evolving privacy regulations. Key DQCs from his perspective include:

1. **Understandability:** Inconsistent data referencing across teams can lead to misunderstandings and compliance risks. Metadata plays a key role in ensuring clarity and consistency, helping teams to use data correctly across different functions.
2. **Authorization:** Only authorized individuals should handle sensitive data, with adherence to legal bases and privacy regulations
3. **Objectivity:** Data should remain impartial and collected consistently to avoid bias, ensuring equal treatment across individuals and contexts.

ISSN 2281-5147

4. **Relevancy:** Data relevancy is crucial for organizations seeking to derive value from their data assets. By ensuring that data outputs align with business objectives, fit the context, and meet customer needs, organizations can enhance decision-making processes and drive successful outcomes (Micheli et al. 2020).
5. **Value-added:** Data should deliver tangible benefits; for example, data from current products should inform improvements in future designs, providing real value.
6. **Communication:** Data must be clear, timely, and accessible only to authorized users, preventing security and privacy breaches.

This feedback emphasizes data consistency, privacy, and relevance within a technologically complex environment where accuracy and compliance are paramount.

4.3 Insights from Interviewee 3: Chief Officer on Privacy Protection

Focusing on data protection in a company that produces a wide range of consumer goods, this expert highlighted the importance of balancing data utility with privacy obligations, particularly in the context of AI applications that require high accuracy. She highlighted some of the challenges posed by Europe's stringent data protection regulations, particularly in relation to the handling of sensitive data. Key DQCs identified by this expert include:

1. **Accuracy:** Data must be correct and reliable, free from errors that could mislead decision-making.
2. **Accessibility:** Data should be easy to locate and accessible only to authorized personnel, ensuring streamlined retrieval processes.
3. **Authorization:** Access rights should be flexible, allowing tailored levels of access based on roles within the organization.
4. **Relevancy:** Particularly in AI, it's crucial to distinguish between relevant data and "noise" to ensure models are built on accurate information.
5. **Objectivity:** Data collection and processing must be unbiased and neutral, avoiding discrimination based on personal characteristics.

ISSN 2281-5147

6. **Ease of manipulation:** The usability of data depends on its format and inherent restrictions; for example, personal data should not be reused for unrelated purposes without consent.
7. **Traceability:** Tracking changes to data is essential, requiring individual accountability and monitoring to ensure transparency and security.

These criteria reflect a strong commitment to both data utility and the safeguarding of personal privacy.

4.4 Insights from Interviewee 4: Researcher in Data Protection Principles

Our final expert, a privacy and data governance researcher, offered a philosophical perspective on the relationship between privacy and data quality. He stressed that privacy includes the right to be unobserved and stressed that individuals do not always want their data to be accurate. His proposed DQCs include:

1. **Safety:** Poor quality data processing of sensitive information can lead to significant risks. A thorough risk analysis is essential to mitigate potential harm, especially under frameworks like GDPR.
2. **Free of error:** This criterion should apply exclusively to factual data, emphasizing the need for objective validation methods.
3. **Reliability:** Ensuring accuracy in data processing and the procedures that support this reliability is critical.
4. **Accuracy:** Data must represent its meaning accurately without bias, maintaining the privacy of individuals while delivering reliable outputs for decision-making.
5. **Value added:** This principle balances legal compliance with the costs of data processing, helping data controllers determine when it is worth processing certain data.

Together, these insights illustrate the intricate balance between data quality, privacy rights, and the ethical considerations inherent in data governance.

ISSN 2281-5147

5. Summary of the Study

5.1 Overview of Analysis

This study explores the perspectives of legal data professionals on key DQCs in light of EU data management regulations. Through the insights gathered from four experienced professionals, we identified 28 key comments highlighting 20 critical DQCs. Notably, 95% of these criteria (19 in total) are consistent with our previously established list. Several DQCs, including **accuracy, authorisation, ease of manipulation, freedom from error, objectivity, relevance, understandability and value added**, were highlighted multiple times, reflecting their importance in the data management landscape..

The responses contributed to a comprehensive classification of DQCs across a generic data lifecycle, confirming the thoroughness of our DQC list and its relevance to current data management practices.

This research aims to enhance the way data controllers should prioritize data quality management and delineate responsibilities to ensure high-quality data outputs. Furthermore, it highlights the connection between data quality management and legal compliance, as a lack of awareness in this domain can lead to significant disadvantages for organizations. The insights provided by our interviewees emphasize the necessity of integrating key EU regulatory points into DQCs.

5.2 Classification of DQCs within a Generic Data Lifecycle

Our primary objective is to assess the trust, reliability and legal compliance of collaborative data processing by DQCs. This analysis places significant emphasis on the legal aspects of data handling in collaborative ecosystems. It facilitates the classification of DQCs within the data lifecycle, enables a clearer assessment of data quality within information systems, and supports the development of automated data quality assessment systems.

Figure 1: Classification of Data Quality Criteria in the data lifecycle

Discussions with the data experts revealed that their responsibilities and expertise in data management correspond to essential steps in the data lifecycle (Shah, Peristeras,

ISSN 2281-5147

et Magnisalis 2021), which include 1) **Collection, 2) Preparation, 3) Analysis, 4) Sharing, and 5) Reuse.** These steps form the backbone of effective data management practices. The insights shared by the experts allowed us to categorize and structure DQCs according to these lifecycle stages, as illustrated in Figure 1.

6. Conclusion

Data quality is a multi-faceted concept that encompasses regulatory compliance, confidence in governance and the reliability of data processing. Achieving these goals requires a thorough understanding of data quality criteria (DQCs). In this study, we have proposed a standardised and unified list of DQCs derived from a comprehensive literature review, aiming to bridge the gap between academic methodologies and practical industrial applications.

As the regulatory landscape for data management continues to evolve in Europe, there is a growing interest in privacy and data protection legislation. This dynamic environment highlights the need for a balanced approach that integrates data quality, system reliability and regulatory compliance. Our findings suggest that trust in data processing can only be established through this fairness.

This paper presents a consolidated framework for assessing data quality, based on insights from open and semi-directive interviews with European data experts from multinational companies. Through these discussions, we refined our initial list of 30 DQCs, identifying the most relevant criteria that align with key points of European data management regulations. We also mapped these DQCs to different stages of the data lifecycle, providing a roadmap for building a trusted, collaborative data processing ecosystem.

7. Future Direction

Looking ahead, our focus will shift to implementing a decentralised, blockchain-based solution for sharing DQCs throughout the data lifecycle. This innovative approach aims to increase the transparency and traceability of data processing activities, enabling all stakeholders to effectively assess data quality. By leveraging such

ISSN 2281-5147

technologies, we can support distributed systems governed by a collaborative governance ecosystem that fosters trust among all data stakeholders.

Through these efforts, we hope to advance the field of data quality management and contribute to a more secure and reliable data processing environment that meets both regulatory requirements and the expectations of all stakeholders involved.

8. Selected Readings

Fabijan, Aleksander, Pavel Dmitriev, Helena Holmström Olsson, et Jan Bosch. 2017. « The Evolution of Continuous Experimentation in Software Product Development: From Data to a Data-Driven Organization at Scale ». P. 770-80 in *2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE)*.

Foote, Keith D. 2022. « Data Quality Dimensions ». *DATAVERSITY*. Consulté 4 octobre 2023 (https://www.dataversity.net/data-quality-dimensions/).

Gan, Wensheng, Jerry Chun-Wei Lin, Han-Chieh Chao, et Justin Zhan. 2017. « Data Mining in Distributed Environment: A Survey ». *WIREs Data Mining and Knowledge Discovery* 7(6):e1216. doi: 10.1002/widm.1216.

Haug, Anders, Frederik Zachariassen, et Dennis van Liempd. 2011. « The Costs of Poor Data Quality ». *Journal of Industrial Engineering and Management (JIEM)* 4(2):168-93. doi: 10.3926/jiem.2011.v4n2.p168-193.

Hupperz, Marius, Inan Gür, Frederik Möller, et Boris Otto. 2021. *What is a Data-Driven Organization?*

Jonathan, Albert, Muhammed Uluyol, Abhishek Chandra, et Jon Weissman. 2017. « Ensuring reliability in geo-distributed edge cloud ». P. 127-32 in *2017 Resilience Week (RWS)*.

Kahn, Beverly K., Diane M. Strong, et Richard Y. Wang. 2002. « Information quality benchmarks: product and service performance ». *Communications of the ACM* 45(4):184-92. doi: 10.1145/505248.506007.

ISSN 2281-5147

Laranjeiro, Nuno, Seyma Nur Soydemir, et Jorge Bernardino. 2015. « A Survey on Data Quality: Classifying Poor Data ». P. 179-88 in *2015 IEEE 21st Pacific Rim International Symposium on Dependable Computing (PRDC)*.

Micheli, Marina, Marisa Ponti, Max Craglia, et Anna Berti Suman. 2020. « Emerging Models of Data Governance in the Age of Datafication ». *Big Data & Society* 7(2):2053951720948087. doi: 10.1177/2053951720948087.

Moore, Susan. 2018. « How To Create A Business Case For Data Quality Improvement ». *Gartner*. Consulté 28 juillet 2023 (https://www.gartner.com/smarterwithgartner/how-to-create-a-business-case-for-data-quality-improvement).

Sahi, Louis, Romain Laborde, Mohamed-Ali Kandi, Michelle Sibilla, Giorgia Macilotti, Benzekri Abdelmalek, et Afonso Ferreira. 2023. « Towards Reliable Collaborative Data Processing Ecosystems: Survey on Data Quality Criteria ». P. 2456-64 in. IEEE Computer Society.

Shah, Syed Iftikhar Hussain, Vassilios Peristeras, et Ioannis Magnisalis. 2021. « DaLiF: a data lifecycle framework for data-driven governments ». *Journal of Big Data* 8(1):89. doi: 10.1186/s40537-021-00481-3.

ISSN 2281-5147