Systematic Mapping of Data Quality in the Public Sector: Dimensions, Metrics and Practices

Maria Inês Vale Silva*
Federal University of Ceará (UFC)
Ceará State Treasury Department
Fortaleza, CE, Brazil
ines.vale@sefaz.ce.gov.br

Ramon Nicolas Gomes Luna*
Federal University of Ceará (UFC)
Fortaleza, CE, Brazil
ramon.g.nicolas2@gmail.com

Amanda K. B. Cavalcante*
Federal University of Ceará (UFC)
Fortaleza, CE, Brazil
amanda.kevillyn@alu.ufc.br

Valéria Lelli
Federal University of Ceará (UFC)
Fortaleza, CE, Brazil
valerialelli@ufc.br

Rossana M. de Castro Andrade
Federal University of Ceará (UFC)
Fortaleza, CE, Brazil
rossana@ufc.br

José Maria Monteiro
Federal University of Ceará (UFC)
Fortaleza, CE, Brazil
monteiro@dc.ufc.br

Ismayle de Sousa Santos
State University of Ceará (UECE)
Fortaleza, CE, Brazil
ismayle.santos@uece.br

ABSTRACT

Data quality plays a critical role in ensuring the reliability and effectiveness of information systems in public sector organizations. However, fragmented data ecosystems, legacy systems, and a lack of standardized assessment practices challenge the implementation of consistent data quality strategies. This study presents a comprehensive systematic mapping of data quality dimensions, metrics, practices, and challenges across public institutions. A total of 53 peer-reviewed studies were analyzed, revealing the most frequently used data quality dimensions and associated metrics. The research highlights a wide spectrum of techniques, ranging from traditional validation rules to emerging applications of machine learning for anomaly detection and data imputation. In addition, it identifies key barriers — including limited automation, lack of governance, and scarce resources — and proposes actionable recommendations for public sector entities. The findings serve as a conceptual and practical foundation for improving data quality management in complex government environments, with the ultimate goal of supporting the implementation of data quality assessment strategies in a state-level public finance institution in Brazil, the Ceará State Treasury Department (Sefaz-CE).

KEYWORDS

Data Quality, Public Sector, Systematic Mapping, Government Data

1 Introduction

This paper addresses the urgent need for a structured understanding of data quality practices in the public sector, where data volume is growing rapidly with the digitalization of the services offered.

However, fragmentation and lack of standardization of systems, limited automation and indicator monitoring tools [1, 8, 25, 49, 56], the scarcity of consolidated data governance practices [26, 54], and the high costs associated with maintaining proprietary solutions [5] compromise the reliability of information provided by governments.

This research responds to this need by systematically mapping the literature on data quality in public organizations. The study identifies and categorizes the most commonly used dimensions, metrics, and methods, revealing how governments in different contexts conceptualize and operationalize data quality. The study explores traditional assessment techniques, such as metrics-based and rule-based validation, and recent advances, such as the use of Machine Learning (ML) for anomaly detection. It also reveals critical gaps and challenges, offering practical recommendations for advancing data quality initiatives in government contexts.

The systematic mapping was guided by research questions that address current approaches, relevant dimensions, assessment methods, and challenges related to data quality in the public sector. The review covered 53 studies from the IEEE, ACM, and Scopus databases, and reflects the growing interest in the topic.

The most frequently discussed data quality dimension was Completeness, cited in 40 studies, but other dimensions are also discussed. Overall, the results offer strategic insights for improving governance and data quality practices in public administration, pointing the way to developing more effective strategies that are sensitive to the context of public administration.

The paper is structured as follows: Section 2 presents the theoretical foundation. Section 3 outlines related work. Section 4 details the methodology used in conducting the systematic mapping. Section 5 presents and discusses the results. Section 6 discusses the threats to validity, and, finally, Section 7 presents the final remarks and directions for future research and practical implementation.

2 Background

Early conceptualizations of data quality, such as those proposed by Wang and Strong [48], emphasized the idea of "fitness for use," arguing that the value of data depends on its ability to support users specific tasks and decisions. Their framework introduced intrinsic, contextual, representational, and accessibility dimensions—an enduring conceptual base for subsequent research. Redman [35] expanded this view by relating data quality directly to organizational

*The authors contributed equally to this research

performance and institutional trust, framing quality as not only a technical property but also a managerial responsibility. These foundational works established that data quality is inherently multidimensional and context-dependent.

Building upon these perspectives, data quality standards and models have evolved to support both theoretical understanding and operational assessment. ISO/IEC 25012:2008 [18] defines fifteen quality characteristics, divided into inherent and system-dependent categories, while ISO/IEC 25024:2015 [19] complements it by providing measurable indicators and methods for evaluating structured data. Together, these standards formalize the link between conceptual attributes and quantifiable metrics, offering a reference framework for practitioners. Although they represent a conceptual baseline for most studies on data quality, in this work, they are used only as theoretical support to interpret and categorize the most frequently mapped metrics, rather than as a guiding framework for the research design. This decision reflects the diversity of approaches found in the literature, which often adapt quality dimensions pragmatically rather than adhering to a single normative model.

Beyond ISO standards, other frameworks contribute to the theoretical landscape. The DAMA-DMBOK [24] broadens the discussion by situating data quality within the broader discipline of data governance, advocating a lifecycle perspective and emphasizing stewardship roles, metadata management, and policy enforcement. Likewise, Madnick et al. [19] propose a unifying framework that connects data and information quality research to business processes, highlighting dependencies between technical standards and institutional capacity. However, these governance-oriented models are underrepresented in public sector applications, where fragmented systems and heterogeneous data sources challenge their full implementation.

Therefore, this study bridges the gap between conceptual models and practical challenges by mapping how data quality dimensions and metrics have been applied in the public sector. It aims to clarify which theoretical constructs have been effectively operationalized and where adaptation or simplification occurs in real-world contexts.

3 Related Work

Specifically, a subset of review and research articles was selected to inform the comparative analysis and provide a broader understanding of the existing research landscape.

The systematic review conducted by Zainuddin and Akhir [55] examined 37 studies on data quality in Open Government Data (OGD), focusing on key dimensions such as completeness, timeliness, and consistency. Their review highlights common challenges related to missing or incomplete values and discusses emerging solutions. While their work focuses on OGD, this review extends the analysis to public sector information systems, classifying metrics or indicators into categories standardized with ISO/IEC 25012:2008 [18].

The survey conducted by Zhang et al. [56] provides a systematic overview of quality assurance (QA) techniques for big data applications, addressing the challenges posed by the 4Vs (volume, velocity, variety, and veracity). Their review highlights six primary QA approaches, including testing, Model-Driven Architecture (MDA), monitoring, and fault tolerance, emphasizing the need for automation and scalability. While that survey focuses on the technical assurance of big data systems, this review extends the discussion by focusing on data quality practices and challenges within public sector organizations, combining both technical and governance-oriented perspectives.

The systematic review conducted by Serra et al. [40] explores how context influences Data Quality Management (DQM), examining how elements such as user requirements, metadata, and application domains shape data quality dimensions and metrics. Their review proposes a taxonomy that connects contextual factors to different stages of DQM, emphasizing the importance of context-aware approaches. In contrast, this study focuses on practical dimensions, metrics, and challenges of data quality in public sector organizations.

The study conducted by Ijab et al. [23] applies a Systematic Literature Review (SLR) to propose a Big Data quality framework tailored to Malaysia's Public Sector Open Data Initiative (MyPS-ODI). The framework consolidates multiple data quality dimensions, such as availability, usability, reliability, concordance, presentation, and correctness, and links them to open data principles to improve transparency and public services. While that work offers a valuable context-specific model for big data and open data, this review focuses on a broader range of public sector information systems, emphasizing both technical practices and organizational challenges of data quality management.

Priestley et al. [32] conducted a literature survey to evaluate how traditional data quality frameworks apply to modern ML pipelines. Their review emphasizes the need for a stage-specific approach to data quality, organized around four key dimensions—intrinsic, contextual, representational, and accessibility. While their survey focuses on the technical aspects of data quality within ML lifecycles, this study addresses broader challenges, particularly those related to data quality management in public sector environments, which are influenced by governance and regulatory factors.

In summary, the analyzed surveys and systematic reviews address data quality from various perspectives, such as open government data, big data assurance techniques, ML pipelines, context-aware approaches, and national open data initiatives. While those studies provide valuable contributions, they are often limited to specific domains or technical scopes. This mapping differentiates itself by offering a comprehensive mapping of data quality dimensions, metrics, practices, and challenges across public sector organizations, combining both technical and governance aspects. This broader approach enables actionable recommendations for improving data quality management in complex governmental ecosystems.

4 Methodology

This systematic mapping aims to identify the dimensions, metrics, and frameworks or methodologies employed to evaluate data quality in the public sector. The PICO approach [50] — which stands for Population, Intervention, Comparison and Outcome — was adopted to formulate the research questions and define the search strings. The review process was managed using the Parsifal[1] , a tool designed to support the planning and execution of systematic reviews.

[1]Parsifal: https://parsif.al/

The overall procedure is illustrated in Figure 1. In total, 660 papers were initially retrieved from the selected databases, and after the screening and eligibility assessment stages, 53 studies were included in the data extraction phase, as detailed in Section 5. To begin the analysis of the 53 selected articles, the steps illustrated in Figure 1 were followed. This collaborative and iterative process was supported by ongoing guidance from the professors, ensuring a balanced distribution of tasks and the reliability of the results.

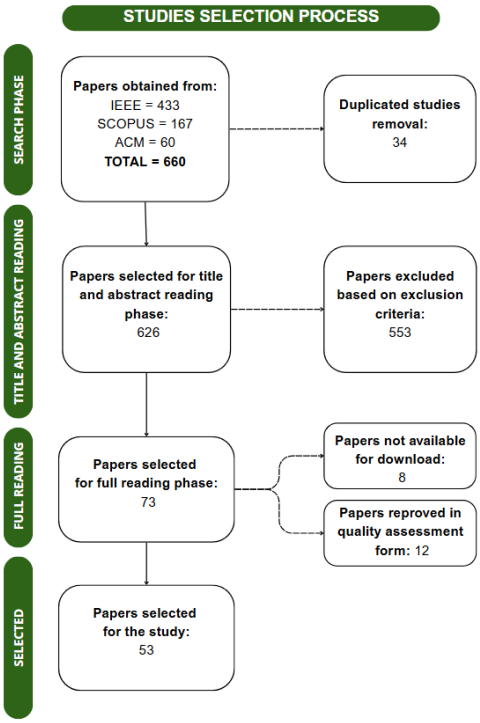

Figure 1: Studies selection process.

4.1 Research Questions

To guide the systematic mapping, a set of research questions (RQ) was defined to explore how data quality is ensured in public domains. These questions were formulated to identify the key dimensions and metrics, methodologies, and challenges associated with evaluating data quality in this context. The following 4 research questions were defined:

- **RQ1** - What are the existing approaches to data quality in public sector organizations?
- **RQ2** - What data quality dimensions are most relevant for a public agency?
- **RQ3** - What measures, metrics and indicators were used to assess and ensure data quality mainly in the public sector?
- **RQ4** - What are the main challenges and limitations faced by public sector organizations in implementing data quality practices?

4.2 Search Strategy

The literature review was conducted using three well-established and widely recognized scientific databases in the field of computing and information systems: the ACM Digital Library, Scopus, and IEEE Xplore Digital Library. These sources were selected because of their relevance, breadth, and academic rigor in publishing high-quality research related to data management and public sector technologies. There was no filter related to the time period.

The first three authors are undergraduate and graduate students at the Federal University of Ceará and are the main investigators of the study. They were involved in conducting the secondary study, under the ongoing guidance of four PhD professors, who are co-authors of the research. The researchers divided the reading of the 53 articles almost equally among themselves, applying the systematic mapping methodology and prioritizing the main points to be included in the article. To ensure rigor and methodological consistency, they held regular knowledge-sharing meetings to discuss the material read, align interpretations, and decide on the most appropriate way to present and discuss the results.

The student researchers also shared other tasks, such as refining the search strategy and applying inclusion/exclusion criteria, extracting and consolidating dimensions, metrics, and practices from the studies, and synthesizing the challenges, recommendations, and implications for the public sector.

To guide the construction of the search string (see Table 1), the PICO elements were defined as follows:

- **Population (P)**: Public agency and related synonyms;
- **Intervention (I)**: Data quality;
- **Outcome (O)**: Metric, Indicator, Measure, Process, Procedure, Framework, Method, and Methodology;
- **Comparison (C)**: Not applied in this study, as the objective was to identify and map the state of the art, rather than perform comparative analyses.

Table 1: Search string created for the study selection.

IEEE (Base String)	("data quality") AND ("public agency" OR "government" OR "national agencies" OR "national agency" OR "public administration" OR "public institution" OR "public sector" OR "state agency" OR "governmental body" OR "governmental bodies" OR "government department")
ACM	Title:("data quality") AND Fulltext:("public agency" OR "government" OR "national agencies" OR "national agency" OR "public administration" OR "public institution" OR "public sector" OR "state agency" OR "governmental body" OR "governmental bodies" OR "government department")
Scopus	(Title ("data quality") AND Abs (("data quality") AND ("government" OR "national agency" OR "national agencies" OR "public agency" OR "public administration" OR "public institution" OR "public sector" OR "state agency")) OR KEY (("data quality") AND ("government" OR "national agency" OR "national agencies" OR "public agency" OR "public administration" OR "public institution" OR "public sector" OR "state agency")))

During the pilot phase, the inclusion of Outcome terms in the search string proved too restrictive, significantly reducing recall and omitting relevant foundational studies. To enhance coverage, the

Outcome (O) component was therefore removed, resulting in a more general query centered on the intersection between Population (public sector) and Intervention (data quality).

The base search string was then adapted for each database (see Table 1). In Scopus and the ACM Digital Library, the term "data quality" was fixed in article titles to filter out irrelevant results, whereas in IEEE Xplore the original query was applied without modification. This strategy balanced precision and comprehensiveness across databases.

4.3 Eligibility Criteria

Inclusion and exclusion criteria were applied to ensure the selection of relevant articles. The inclusion criteria comprised studies that specifically addressed data quality in the public sector, including those that presented metrics, dimensions, challenges, solutions, or open questions related to data quality assessment and management in governmental or public administration contexts. The inclusion criteria are listed below:

(1) Context in the public sector;
(2) Coverage of data quality challenges, techniques, or frameworks;
(3) Discussion on heterogeneous databases or hybrid environment;
(4) Focus on measures and indicators;
(5) Relevance to data quality or data governance;
(6) Use of Artificial Intelligence (AI)/Machine Learning (ML) in measuring or controlling data quality.

The exclusion criteria were specified as follows:

(1) Does not present measures and indicators;
(2) Duplicate publication;
(3) Focus on non-related technical aspect;
(4) Irrelevant research focus;
(5) Not available in English or Portuguese;
(6) Not available in full text;
(7) Short paper (four pages or less).

4.4 Study Selection

The article selection was managed using Parsifal. In the first stage, titles and abstracts of all retrieved papers were screened and either advanced or excluded according to predefined criteria. The remaining studies then underwent full-text review and quality assessment, as detailed in the next subsection.

4.5 Quality Assessment

To ensure the rigor and consistency of the selected studies, a structured quality assessment form was developed and applied using the Parsifal tool. The form comprised the following questions:

(1) Does the paper explicitly discuss data quality as a central concept?
(2) Are the methods, techniques, and tools used in the study clearly described and appropriate for assessing or improving data quality?
(3) Does the study propose or analyze methods for measuring or controlling data quality using quantitative data quality indicators?
(4) Does the study use real-world data for validation?
(5) Does the paper use a clear and replicable methodology for data quality assessment?
(6) Does the study focus on the public sector (specially finances)?
(7) Does the article mention challenges, suggestions, limitations and gaps for future researches?

The form evaluated key aspects such as relevance, methodological soundness, and reliability, ensuring that only robust contributions were included in the review. Each question was aligned with the research objectives and offered three possible answers: "Yes" (2 points), "Partially" (1 point), and "No" (0 points). The total score for each study was obtained by summing the assigned values.

To advance to the next screening stage, a study had to achieve a minimum score of seven points, ensuring adherence to the established quality standards. Studies scoring below this threshold were excluded. After completing the selection and quality assessment process, 53 studies were retained for final analysis.

The sixth question specifically emphasizes financial contexts, as the ultimate goal of this mapping is to instantiate the identified measures for assessing data quality within Sefaz-CE, a public institution responsible for managing state-level revenues and expenditures in Brazil. Consequently, studies applied to real-world scenarios of tax collection or public spending—whether municipal or state-level—received higher scores.

This approach reflects the particular demands and constraints of public administration, especially in fiscal and financial domains, where data quality directly influences transparency, accountability, and governance.

4.6 Data Extraction

Finally, during the data extraction process, relevant information was systematically collected from the selected studies. A structured data extraction form, whose questions are listed below, was created to capture key details, including study objectives, methodologies, and main findings.

(1) What are the main objectives of the study related to data quality?
(2) Which data quality dimensions were considered?
(3) What methods, techniques, or tools were utilized to assess data quality?
(4) What specific metrics or indicators are used in the study to assess data quality?
(5) Are any new indicators developed in the study to assess data quality?
(6) Does the study provide recommendations for public practices or policies?
(7) What specific ML techniques are employed in the study to assess or improve data quality?
(8) Does the study identify specific gaps in the literature or current practices?

The type of all response fields is String. This step aimed to standardize the extracted data, facilitating comparison and synthesis across studies. The procedure was also conducted in Parsifal and supported by spreadsheets[2] to maintain control and organization throughout the process.

[2]Supplementary materials of the study are available at: https://doi.org/10.5281/zenodo.17334803

5 Results and Discussion

This section presents a synthesis of the answers to the four research questions that guided this systematic mapping of data quality practices in the public sector.

The temporal distribution of the 53 selected studies (Figure 2) confirms a growing interest in data quality within the public sector, with a notable concentration of publications from 2022 to 2024 (peaking at 6 articles in 2023). Furthermore, the geographical analysis (Figure 3) highlights the global, yet regionally concentrated, nature of the research. The majority of the studies originate from Europe (35.8%) and Asia (32.1%), a factor relevant to the external validity and transferability of the findings to contexts with distinct governance maturity.

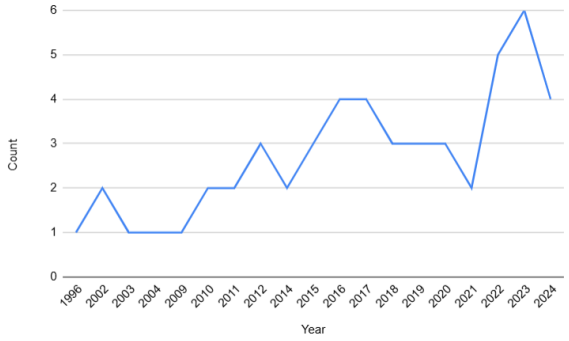

Figure 2: Articles per Year

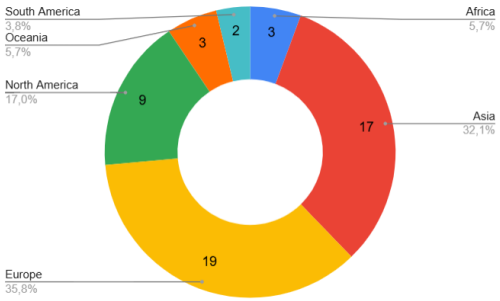

Figure 3: Articles per Continent

The results demonstrate a broad spectrum of approaches, metrics, and challenges, reflecting the complex and heterogeneous nature of public administration data environments. From foundational validation techniques to more advanced and participatory frameworks, the studies reviewed highlight how public organizations conceptualize, operationalize, and struggle with data quality.

Before analyzing the findings in extracting responses to the research questions, the main objectives of the studies related to data quality will be presented, grouped by thematic focus:

- **Assessment of Data Quality**: Data quality assessment methods include quantitative analyses [13, 52], validation procedures [21, 38], and case-based measurements [2, 4, 5, 7, 8, 28, 39, 54] to determine, primarily, the level of data reliability, completeness, timeliness and accuracy. Some articles propose models or systems that measure data quality automatically [10, 25, 47] or through manual inspection [10, 33, 49] of datasets.
- **Data Quality Improvement**: Some studies focus on methods for improving or cleaning data [10, 17, 32, 42], either by proposing correction mechanisms [5, 7, 15, 27] or improving data collection [7, 27] processes. Other articles present practical techniques for feedback loops [44], or proactive quality monitoring [9, 26] to ensure more consistent and usable datasets. These contributions often draw on lessons learned from field implementations or pilot studies, focusing on data quality as an evolving operational challenge.
- **Frameworks and Governance Models**: Many studies focus on developing conceptual frameworks or operational models to support data governance and quality strategies [1, 8, 12, 15, 29, 44, 45, 54, 58]. Among these, there are articles that describe structured approaches that align data quality practices with organizational objectives, particularly in government and public finance institutions. These models typically integrate multiple dimensions of quality and are designed to be adaptable to diverse institutional contexts.
- **Automation and Tooling**: Some studies explore the development of automated systems and tools to assess or improve data quality [1, 2, 6, 9, 13, 14, 21, 25–28]. These tools are especially relevant in complex data environments, where manual inspection is impractical and where automation can accelerate quality control in real-time applications.
- **Open Data and Transparency**: Many of the articles analyzed address Open Government Data (OGD) [1, 9, 13, 33, 38, 47, 55], but few explicitly mention objectives related to public transparency. These studies highlight the role of data quality in fostering trust and accountability in public data platforms, linking technical practices to governance and civic engagement objectives.

5.1 RQ1 - What Are the Existing Approaches to Data Quality in Public Sector Organizations?

This systematic mapping reveals that rule-based validation techniques and metric-based frameworks are the most prevalent approaches for assessing data quality in this context [14, 28, 38, 53]. Several studies adopt standardized models such as ISO/IEC 25024:2015 [19], which provides structured quality characteristics, or use specific quality dimensions like completeness, consistency, and accuracy to construct measurement frameworks [38].

SQL-based stored procedures [38], Open Data Indicators (ODI) scoring methodologies [13], and schema validation [47] are frequently applied to detect anomalies and validate structural integrity in datasets. Some articles integrate statistical techniques such as the Tukey test[3] [1] or employ custom scoring formulas and composite indices that aggregate various quality indicators into a single score [13, 20].

Some studies develop dashboards or quality monitoring systems tailored to public administration contexts [14, 26, 29]. For example,

[3]Tukey test is a statistical test used to compare multiple group means after an analysis of variance indicates a significant difference between them.

the QUALYST system [38] incorporates quality dimensions into institutional workflows, enabling recurring assessments. Others adopt process-centric approaches, incorporating quality evaluation into data governance routines or linking it to data lifecycle management policies [8, 29, 36, 54].

Although automation is not yet widespread, a number of articles suggest increasing reliance on software tools to operationalize data quality evaluations. Examples include the use of RapidMiner [1], DQ tools with visualization capabilities [25, 26, 49], and the integration of quality assessment with business intelligence platforms to monitor quality in real time [14, 21].

A significant portion of studies rely on expert knowledge or stakeholder input to define what constitutes quality data within the specific context of public administration. In these cases, data quality assessment often takes a more qualitative or participatory approach, incorporating surveys and interviews [1, 2, 9, 11, 32, 34, 43, 51, 54, 56, 57].

Despite the heterogeneity of tools and frameworks, common challenges emerge. These include data fragmentation across departments, lack of standardization, and the difficulty of maintaining quality over time. Several studies [8, 29, 31, 34, 54] emphasize the importance of institutional commitment and governance structures to sustain DQ initiatives.

5.1.1 *Machine Learning Techniques for Data Quality Enhancement*

Although traditional methods remain predominant in the assessment and improvement of data quality within the public sector datasets, recent studies have increasingly investigated the application of ML techniques as complementary or alternative approaches. Among the 53 studies included in this review, 15 (28.3%) explicitly reference the use, or the potential use, of ML techniques for tasks such as data quality assessment, anomaly detection, or automated data cleansing. This trend reflects a growing recognition of the capacity of ML models to address complex patterns of inconsistency and incompleteness that may not be easily captured through rule-based or manual methods.

Supervised learning algorithms have been widely adopted in this context. For example, decision trees are employed to identify and classify anomalous data patterns [3, 11, 37, 42, 56], while naïve Bayes classifiers [11, 56] and random forests [15, 20] are used to predict the likelihood of data quality issues. In some cases, ML is integrated into data cleaning pipelines to support tasks such as data augmentation and annotation during pre-processing stages [32].

In addition, unsupervised learning techniques, particularly clustering algorithms such as k-means, are applied to uncover latent patterns and detect outliers. These approaches allow the identification of anomalous records that do not conform to the structure of the main data clusters, thus facilitating the detection of potential quality issues without the need for labeled data [42].

An emerging trend is the use of deep learning models such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term memory (LSTM) for imputing missing or incomplete data [37], especially in contexts involving textual or unstructured data. In a Systematic Literature Review of Data Quality in Open Government Data [55] various uses of CNN, RNN and LSTM are also cited.

Elouataoui et al. [15] focus on anomaly detection using Isolation Forests or k-NN algorithms in metadata spaces. In contrast, Davidson et al. [11] explore ensemble techniques such as the Ensembles of Quality-Matrix Perturbed Data (EQPD), which enhance classifier robustness by perturbing the training data through quality matrices.

Berndt et al. [4] present a work that uses an extensive corpus extracted from an Electronic Health Record (EHR) system to investigate issues related to task complexity, sample size, and goal quality, using a large number of data mining models, including simple keyword or regular expression matching, Natural Language Processing (NLP), and ML algorithms.

Even when not directly employing ML, many articles signal its potential. For instance, Batini et al. [1] use the Tukey test in RapidMiner and suggests AI integration for future automation, while Croft et al. article [10] evaluate the impact of data quality on ML model performance using a transformer-based line-level vulnerability prediction model (LineVul) to show how poor data can degrade accuracy.

Overall, while ML is not yet mainstream in public sector DQ workflows, its growing application shows promise—especially for large-scale anomaly detection, data imputation, and intelligent record linking.

5.2 RQ2 - What data quality dimensions are most relevant for a public agency?

Some dimensions found in the research for this work are considered classic [18, 19] when consulted in consolidated data quality frameworks [18, 24, 48]. They are: Completeness, Consistency, Accuracy, Uniqueness, Timeliness and Validity. Other dimensions are recognized in specific contexts or as subdimensions, such as Interpretability [1, 8, 13], which is treated as part of Representational Data Quality [48] and is related to the semantic and syntactic clarity of data. Related to interpretability and accessibility is Usability, recognized in interface and analysis contexts. Linkage, recognized in Linked Data and open data [12], is not classic, but relevant in interoperability. In addition to these, Conformity is cited in [27], similar to validity, format compliance or consistency with standard patterns.

Other data quality dimensions can be considered emerging, specific, or still debated, such as Security, Performance, and Readability. These are often considered system or presentation qualities, not core data quality dimensions. Security is not traditionally a dimension of data quality, but rather of data governance. Performance is related to the performance of data systems, but it is not a dimension of data quality per se, but rather of systems. Readability occasionally appears as part of representational quality, more common in document and metadata evaluation.

Provenance is increasing in importance with open and scientific data. ISO 25012:2008 [18] considers it relevant to trustworthiness. Believability is cited by Wang & Strong [48] as a dimension of trust in data. Authenticity is associated with reliability, especially in legal or scientific contexts. It is considered by some as a subdimension of trustworthiness.

According to Table 2, the most frequently mentioned data quality dimensions in the analyzed literature is Completeness (cited

in 40 articles, with 6 metrics or indicators classified in this dimension), which evaluates whether all necessary information is present. Consistency, which refers to the absence of contradictions and the coherence of data in relation to a specific context, both within the same data set and between different relevant data sets, was mentioned in 37 articles. Consistency is critical to the reliability of information systems, as it ensures that integrity rules (such as functional dependencies or disjoint classes) are not violated. For example, metrics such as Member of Disjoint Classes [12] or Implication Complexity [17] are used to verify logical contradictions between data. Consistency is also present in ISO/IEC 25012:2008 [18] as both an inherent and system-dependent characteristic, reflecting its importance in various contexts.

Table 2: Most mentioned dimensions

DQ Dimension	Number of Metrics classified in the Dimension	Number of Articles mentioning
Completeness	6	40
Consistency	12	37
Accuracy	10	29
Uniqueness	3	19
Timeliness	8	17
Validity	3	6
Accessibility	5	5
Interpretability	4	5
Usability	2	2

Accuracy, cited in 29 studies, represents the degree to which data have attributes that correctly represent the true value of the intended attribute of a concept or event, in a specific context of use. It is one of the most intuitive dimensions, often measured by comparison with sources considered "ground truth" or by manual inspection of correct labels, as observed in classification tasks. Works such as that of Batini et al. [3] highlight accuracy as an essential dimension for analytical and decision support applications, where incorrect information can lead to misguided actions. Uniqueness, which relates to unique records appearing only once in a dataset, appears 19 times.

Timeliness, mentioned in 17 articles, refers to the timely availability of data for its intended use. Metrics such as delay calculations [13] and update frequency [3, 6, 10, 22, 23, 30, 31, 43] are used to assess this dimension, especially in monitoring contexts, such as environmental or fiscal data. The distribution of the frequency of these dimensions across the analyzed years is detailed in Figure 4, where the heat intensity indicates the volume of publications per year.

Other relevant dimensions include Accessibility (mentioned 5 times) appears frequently in studies addressing open or distributed data, such as Dereferenceability in Linked Data. Interpretability (5 mentions), which concerns the ease of understanding of data by humans or systems, often associated with the presence of clear metadata. Dimensions such as Validity and Usability are also mentioned, particularly in studies focused on the integration of heterogeneous databases and interoperability. Although characteristics such as security, performance, provenance, and authenticity appear less frequently, they play an- important role in sensitive domains, such as healthcare or public safety, and tend to gain relevance as data governance becomes more mature.

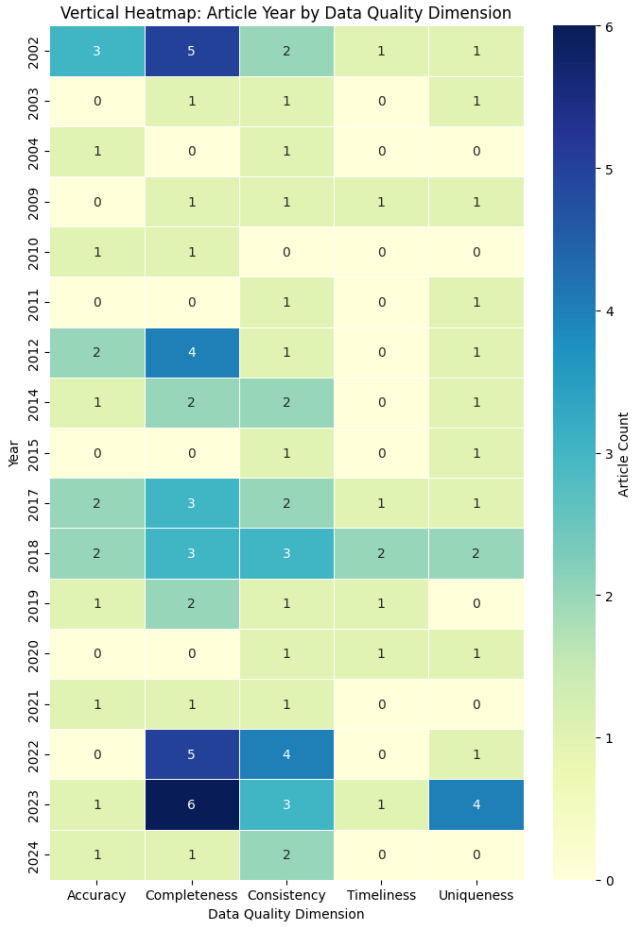

Figure 4: Heatmap of Article Year by Data Quality Dimension Frequency.

5.3 RQ3 - What measures, metrics and indicators were used to assess and ensure data quality mainly in the public sector?

The mapping revealed a wide variety of quantitative and qualitative metrics used to assess and monitor data quality in public sector datasets. Further elaborating on the quantitative instruments used to assess data quality in public sector datasets, approximately 70 distinct metrics were identified. These measures have been grouped into more common data quality metrics, while other emerging and new indicators will be briefly discussed below.

5.3.1 Most Common Metrics and Indicators Table 3 presents a consolidated set of data quality metrics identified across the reviewed studies. The metrics are organized according to their corresponding data quality dimensions, the references of the studies in which they were mentioned, their frequency of occurrence in the analyzed

literature, and the categorization of each dimension according to ISO 25012:2008 [18].

Some metrics are mentioned more than others because they are associated with recurring and universal problems in datasets, such as the presence of null fields or duplicates. The metric "% of null/missing values", for example, is one of the most frequently mentioned (18 times), as completeness is one of the most basic and easily measurable dimensions of data quality. Metrics such as "Update frequency", which indicate the frequency of data updates, while important, appear less frequently (8 times), possibly because they require metadata that is not always available in the analyzed datasets.

Whereas quality attributes are classified as either inherent or system dependent, the "Redundancy detection" metric, for example, falls under the "Uniqueness" dimension and is considered inherent because it assesses whether there are duplicate records within the data itself, regardless of context. Metrics such as "Update frequency" are "System-dependent", as they require knowledge of the data operational context, such as the expected frequency of updates. This distinction is essential for guiding data quality assessment and improvement strategies in organizational environments. It is interesting to note that some measures can be classified both from "Inherent" and "System dependent" point of view, depending on the business context considered.

Many studies adopt standard, widely accepted indicators, including:

- **Consistency and formatting**: syntactic rule violations, pattern frequency checks, standard format compliance and conflict label rate (in annotation tasks).
- **Accuracy and validity measures**: percentage of correct values, logical constraints (e.g., payment amount ≤ award value) and range checks and domain conformity.
- **Timeliness and currentness**: delay calculations, entry delay in days or weeks, update frequency and publication/content timeliness.
- **Completeness-related metrics**: percentage of complete cells, tuple and field-level completeness, Data Completeness Rate (DCR), presence of required elements (e.g., documentation, field population), missing value counts and null value ratios.
- **Uniqueness and duplication**: duplicate detection rates, unique key ratios, edit distance between names and addresses, and redundancy detection (duplicate rows/columns).
- **Rule-based quality indicators**: Conditional Functional Dependencies (CFDs) with support/confidence filtering, pass/fail rule execution metrics and violation frequency of constraints (e.g., 'graduated = yes' requires 'graduation year' not null).

5.3.2 Specialized and Contextual Metrics Some studies introduced or emphasized tailored quality metrics based on the characteristics of the dataset or the sector:

- **Streaming and time-series data quality**: time uniqueness, observation-level completeness, range-based accuracy and typicality (e.g., within 80%, 95%, and 99% confidence intervals) and moderation and conformity metrics specific to temporal patterns.
- **Metadata quality**: completeness and accuracy of metadata entries, standardization and multilingual inconsistencies in descriptors and data quality in monitoring and audit processes (reporting frequency and discrepancy analysis, concordance across datasets, verification factors and quality indexes with calculated weights).
- **User-perceived quality indicators**: survey-based metrics (1 to 5 scale) on dimensions such as clarity, usefulness, ease of use, and satisfaction.
- **Privacy-preserving and anomaly-based assessments**: mean squared error for anonymized datasets, certainty scores of anomaly class labels, mutual information between variables to detect structural anomalies.

5.3.3 Performance-Oriented Metrics When machine learning or rule-based inference is used to detect or improve data quality, several studies report performance metrics to evaluate these methods:

- **Classification and prediction accuracy**: Precision, Recall, F1-score, Accuracy, Confusion Matrix (True Positives - TP, False Positives - FP, True Negatives - TN, False Negatives - FN) and Area Under the Curve - AUC.
- **Efficiency and operational performance**: runtime performance, speed of anomaly correction, anomaly validation rate and behavior-based pass/fail metrics.

While many articles still refrain from explicitly quantifying data quality (often relying on qualitative assessments of errors or principles such as trust and interpretability), the trend across recent studies clearly points toward diversification of metrics to handle new data types (e.g., streaming, text, image, multilingual), greater use of composite indicators to support governance, decision-making, and public transparency, besides integration of ML evaluation metrics where automated quality control or repair is proposed.

Although still rare (15 % of studies), bespoke indicators allow organizations to internalize what "fitness for use" means inside specific public-sector processes, a trend expected to accelerate with the spread of data contracts.

The results reveal a maturing landscape in which traditional, interpretable metrics such as consistency, accuracy, and completeness remain dominant, while newer studies advance toward composite, context-aware indicators and ML-assisted approaches, reflecting the growing complexity of public sector data ecosystems.

5.4 RQ4 - What are the main challenges and limitations faced by public sector organizations in implementing data quality practices?

The implementation of robust data quality practices in public sector organizations is permeated by various challenges and limitations, which emerge from both technical and organizational and human issues. One of the main gaps identified in the literature and current practices is the lack of structured and efficient methods for continuous data quality monitoring, especially in contexts of limited resources, where exhaustive data review becomes unsustainable.

The limitations were systematically classified into nine distinct categories, allowing for a structured understanding of the recurring obstacles in this context. The nine categories, along with their respective challenges/limitations, are listed below:

Table 3: Most mentioned metrics and your dimensions

Metric or Indicator	Data Quality Dimension	Article References	Frequency of the Metric	ISO 25012:2008 Category
% of null/missing values	Completeness	[2, 3, 5, 6, 14, 20, 22, 25, 28, 30, 31, 37–39, 43, 44, 49, 55]	18	Inherent
% of cell completeness	Completeness	[1, 5, 23, 27, 31, 47, 55]	7	Inherent
% of variable/column completeness	Completeness	[1, 9, 16, 22, 27, 31, 42]	7	Inherent
% of tuple/row completeness	Completeness	[1, 9, 27, 47]	4	Inherent
Conformity Format	Consistency	[1–3, 5–7, 10, 11, 14, 17, 20, 25–28, 31, 36–38, 42–44, 47, 49, 55, 58]	25	Inherent
% of correct values	Accuracy	[5, 6, 11, 14, 15, 22, 23, 25, 31, 44, 47, 55]	12	Inherent
Ratios of accurate/non-duplicate entries	Accuracy	[20, 31, 38, 39]	4	Inherent
Outlier detection	Accuracy	[1, 6, 15]	3	Inherent
Range	Accuracy	[5, 25, 27]	3	Inherent
Redundancy detection	Uniqueness	[1–3, 5–7, 9, 10, 17, 22, 26, 28, 30, 38, 39, 42, 47]	17	Inherent
Time Uniqueness	Uniqueness	[27]	1	System-dependent
Update frequency	Timeliness	[3, 6, 10, 22, 23, 30, 31, 43]	8	System-dependent
Anomaly validation rates	Validity	[6, 7, 20, 25]	4	System-dependent

(1) **Models, Frameworks, and Standards:**
- Lack of holistic, lifecycle-based models [15, 29, 32, 44];
- Incompatibility of existing frameworks with domain-specific or national portals [1, 9, 38, 47, 52];
- Need for domain-independent and reusable systems [38];
- Fragmented practices and overuse of ad hoc metrics [1, 21, 28, 31, 38, 41];
- Weak alignment between QA frameworks and Big Data domain [15, 21, 23, 28, 56].

(2) **Metrics and Evaluation:**
- Absence of fine-grained, dataset-level quality assessments [47];
- Inconsistent indicator weighting [1, 52, 58];
- Overreliance on subjective or hardcoded metrics [6, 52];
- Lack of support for metric reuse and extensibility [12];
- Limited context-awareness in metrics [3, 8, 22, 25, 32, 40, 45].

(3) **Tools, Automation, and Monitoring:**
- Limited availability of automated tools [1, 8, 25, 49, 56];
- Lack of real-time monitoring [3, 22, 27, 32, 44, 49, 56] and visualization [14, 25, 29];
- Inadequate integration of DQ functions (e.g., profiling, deduplication, governance) [3, 16, 25, 26, 41];
- Lack of scalable [17, 22, 25] and user-friendly systems [6, 27].

(4) **Governance, Roles, and Processes:**
- Missing formal governance roles [26, 54];
- Lack of policies, coordination, and dedicated personnel [54];
- Underutilization of structured decision-making frameworks [52];
- Poor alignment between data quality and organizational Key Performance Indicators (KPI) [26].

(5) **Metadata and Semantic Structure:**
- Poor metadata management and classification [9, 15, 29, 47, 57];
- Insufficient semantic quality in implementation layers [29, 41];
- Lack of context models in DQ frameworks [40];
- Limited support for active metadata and unstructured data [15].

(6) **Sustainability and Resource Constraints:**
- High costs for maintaining proprietary approaches [5];
- Impracticality of exhaustive manual reviews in low-resource settings [22];
- Lack of scalable, long-term monitoring systems [25].

(7) **Privacy, Security, and Trust:**
- Lack of privacy-preserving solutions [20];
- Absence of trust metrics in sensitive domains [20, 41];
- No support for secure data sharing and evaluation [20].

(8) **Empirical Studies and Applied Validation:**
- Scarcity of longitudinal and real-world case studies [2, 5, 27];
- Need for applied research in public and low-resource sectors [28, 34, 37];
- Weak feedback and evaluation mechanisms [2, 28].

(9) **ML, NLP, and Data Science:**
- Poor alignment of DQ practices with ML workflows [32];

- Inadequate tools for evaluating complex/behavioral model failures [48];
- Limited guidance for cleaning and preparing noisy or inconsistent datasets [53].

Significant challenges persist in the evaluation of data quality within Open Government Data (OGD) initiatives. Studies highlight the lack of automated, granular assessments at the dataset level, limited adoption of quality standards, and insufficient interoperability due to proprietary formats and weak implementation of Linked Data principles.

Technological barriers also remain, such as low automation, absence of real-time monitoring tools, and limited adaptability of existing frameworks to big data and unstructured information. Overall, the literature reveals a pressing need for standardized, automated, and context-aware approaches to data quality management in the public sector.

5.5 Recommendations for public practices or policies

To improve the quality of public sector data, some suggestions are given: prioritize governance [1, 8, 15, 29, 34, 41, 42, 54, 56, 58], defining clear roles (e.g., Data Owner and Data Stewards), adopt policies, and align governance with the structure. Furthermore, implement standardization of metadata, identifiers, and publication formats, and achieve cross-domain alignment.

Embed continuous improvement and tools [1, 2, 6–8, 14, 16, 21, 25–28, 30, 39, 46, 56], building quality control into processes, pairing simple tools with domain expertise, and strengthening monitoring, and metadata systems (including user simulation tools). For big data, adopt enterprise-level Big Data Quality Assessment (BDQA) programs, certifications, and training for public data systems.

Invest in capacity building and human support [6, 8, 21, 30, 58], emphasizing training/supervision, improving application usability, integrating with national systems, secure government backing, and prioritizing adaptability and user-driven evolution of requirements.

Advance innovation and technology [5, 6, 25, 27, 39, 47], automating quality metrics for transparency/accountability; deploying digital tools (e.g., The Spatial Information Exchange—SIX[4]—and e-Planning) to boost efficiency and accuracy [46]; publishing in non-proprietary formats/RDF, link datasets, and exploring metadata-driven ML pipelines for large-scale governance [15].

Target domain gaps with regulatory reform, stronger validation, and infrastructure investment [39, 49]; reorganize local fiscal data processes to enable innovative services [5]; improve documentation, metadata transparency, and standard benchmarks for software-vulnerability data [10]; and apply Lot Quality Assurance Sampling (LQAS) to scale routine supervision and quality improvement in public health [22].

Technical constraints, opaque governance, and cultural factors within public institutions hinder effective data-quality implementation. Yet many studies also outline practical paths forward based on real-world experience and observed needs. Taken together, the challenges and proposed remedies provide a clearer view of the public sector's current data-quality landscape and actionable guidance for those seeking to advance it.

[4]SIX, developed by New South Wales Department of Lands, provided a shared framework to integrate foundation and business data on the web in a timely way.

6 Threats to Validity

To ensure transparency and credibility, threats to validity were examined in four dimensions: construct, internal, external and reliability. Construct validity concerns how faithfully concepts such as data quality, metrics, and dimensions were represented. Despite guidance from ISO/IEC 25012 and DAMA-DMBOK, the literature shows substantial variation in terminology and interpretation; ambiguous or domain-specific labels (e.g., "metadata richness," "information usability") required subjective judgment and may have introduced inconsistencies in grouping or naming indicators.

Internal validity addresses the accuracy and coherence of findings. Although generative AI tools (e.g., ChatGPT and Gemini) supported initial extraction and classification, all outputs were manually reviewed. Even so, bias may persist where metric definitions were unclear and inference guided classification; for example, treating "interpretability" as a subdimension of "representational quality" could have influenced frequency counts and synthesis.

External validity limits generalizability. Because the review targets the public sector in general, results may not reflect practices in specific domains (e.g., finance, health, education) or translate directly to the private sector. Moreover, the predominance of studies from developed countries constrains applicability to regions with lower digital infrastructure and governance maturity.

Reliability pertains to reproducibility: structured protocols (PICO) and the Parsifal platform supported standardization, but the lack of multiple independent reviewers at every stage and the reliance on expert judgment for metric classification—without external validation—leave room for interpretation bias.

7 Final Remarks

This study mapped the main dimensions, metrics, and practices related to data quality in the public sector. The results revealed a consistent focus on core dimensions — completeness, consistency, accuracy, uniqueness, and timeliness — which are essential for producing reliable and transparent, decision-ready data. Additional concerns such as provenance, interpretability, and linkability also emerged in open and linked data contexts.

The identified metrics range from basic checks for null values and metadata validation to more complex, context-dependent indicators of non-compliance. While traditional methods like SQL remain prevalent, there is a growing interest in applying artificial intelligence for anomaly detection and automated quality control.

The review also highlighted persistent challenges, including system fragmentation, lack of standardization, and limited automation. Addressing these issues requires stronger governance structures, capacity building, and the adoption of context-aware metrics. In general, this mapping synthesized best practices, trends, and gaps, offering actionable insights tailored to the realities of public administration. Based on these findings, public institutions can design a practical roadmap by selecting the most relevant dimensions and metrics, using the most cited methods as a foundation to identify quality issues and developing customized indicators for continuous improvement. As future work, Sefaz-CE will apply this roadmap to its data quality program and publish the results achieved.

ACKNOWLEDGMENTS

We thank Ceará Foundation for Scientific and Technological Development (FUNCAP) for their financial support. We also thank CNPq for the productivity grants to Prof. Rossana M. C. Andrade (306362/2021-0) and Prof. Ismayle S. Santos (312173/2025-3), as well as Sefaz-CE for encouraging the use of research to support problem-solving.

REFERENCES

[1] Nada Faisal Alogaiel and Omer Abdulaziz Alrwais. 2023. An Assessment of the Quality of Open Government Data in Saudi Arabia. *IEEE Access* (2023). doi:10.1109/ACCESS.2023.3285611

[2] Claudia Mihaela Balan. 2014. E-government quality of data. In *2014 First International Conference on eDemocracy & eGovernment (ICEDEG)*. IEEE, 91–95. doi:10.1109/ICEDEG.2014.6819958

[3] Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. 2009. Methodologies for data quality assessment and improvement. ACM Comput. Surv. doi:10.1145/1541880.1541883

[4] Donald J. Berndt, James A. McCart, Dezon K. Finch, and Stephen L. Luther. 2015. A Case Study of Data Quality in Text Mining Clinical Progress Notes. *ACM Transactions on Management Information System* (2015). doi:10.1145/2669368

[5] Mario Bochicchio and Antonella Longo. 2002. An Effective Approach to Reduce the ihAvalanche Effectln in the Management of Fiscal Data in Local Public Administration. *Software Maintenance, IEEE International Conference on* 0 (10 2002), 0560.

[6] Christian Bors, Theresia Gschwandtner, Simone Kriglstein, Silvia Miksch, and Margit Pohl. 2018. Visual Interactive Creation, Customization, and Analysis of Data Quality Metrics. *J. Data and Information Quality* (2018). doi:10.1145/3190578

[7] Isabelle Boydens. 2011. *Strategic Issues Relating to Data Quality for E-Government: Learning from an Approach Adopted in Belgium.* Springer New York, New York, NY, 113–130. doi:10.1007/978-1-4419-7533-1_7

[8] Sapa Chanyachatchawan, Krich Nasingkun, Patipat Tumsangthong, Porntiwa Chata, Marut Buranarach, and Monsak Socharoentum. 2023. Design and Implementation of a Data Governance Framework and Platform: A Case Study of a National Research Organization of Thailand. In *2023 20th International Joint Conference on Computer Science and Software Engineering*. doi:10.1109/JCSSE58229.2023.10201972

[9] Abiola Paterne Chokki, Antoine Clarinval, Anthony Simonofski, and Benoît Vanderose. 2023. Evaluating a Conversational Agent for Open Government Data Quality Assessment. In *29th Annual Americas Conference on Information Systems, AMCIS 2023*.

[10] Roland Croft, M Ali Babar, and M Mehdi Kholoosi. 2023. Data quality for software vulnerability datasets. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE, 121–133. doi:10.1109/ICSE48619.2023.00022

[11] Ian Davidson, Ashish Grover, Ashwin Satyanarayana, and Giri K Tayi. 2004. A general approach to incorporate data quality matrices into data mining algorithms. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 794–798. doi:10.1145/1014052.1016916

[12] Jeremy Debattista, Sören Auer, and Christoph Lange. 2016. Luzzu—A Methodology and Framework for Linked Data Quality Assessment. *ACM Journal of Data and Information Quality* (2016). doi:10.1145/2992786

[13] Ilie Cristian Dorobăț and Vlad Posea. 2021. Open Data Indicator: An Accumulative Methodology for Measuring the Quality of Open Government Data. In *2021 13th International Conference on Electronics, Computers and Artificial Intelligence*. doi:10.1109/ECAI52376.2021.9515147

[14] Mohammad Reza Effendy, Tien Fabrianti Kusumasari, and Muhammad Azani Hasibuan. 2019. Star Schema Implementation For Monitoring in Data Quality Management Tool (A Case Study at A Government Agency). In *Proceedings of 2019 4th International Conference on Informatics and Computing, ICIC 2019*. doi:10.1109/ICIC47613.2019.8985695

[15] Widad Elouataoui, Saida El Mendili, and Youssef Gahi. 2024. Active Metadata and Machine Learning based Framework for Enhancing Big Data Quality. *NISS 2024, April 18, 19, 2024, MEKNES, AA, Morocco* (2024). doi:10.1145/3659677.3659707

[16] Gregor Endler. 2012. Data quality and integration in collaborative environments. *SIGMOD/PODS'12 PhD Symposium* (2012). doi:10.1145/2213598.2213606

[17] Wenfei Fan. 2015. Data Quality: From Theory to Practice. *SIGMOD Rec.* (2015). doi:10.1145/2854006.2854008

[18] International Organization for Standardization. 2008. *ISO/IEC 25012:2008 — Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model.* Technical Report ISO/IEC 25012:2008. International Organization for Standardization.

[19] International Organization for Standardization. 2015. *ISO/IEC 25024:2015 — Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Measurement of data quality.* Technical Report ISO/IEC 25024:2015. International Organization for Standardization.

[20] Julien Freudiger, Shantanu Rane, Alejandro E Brito, and Ersin Uzun. 2014. Privacy preserving data quality assessment for high-fidelity data sharing. In *Proceedings of the 2014 ACM workshop on information sharing & collaborative security*. 21–29. doi:10.1145/2663876.2663885

[21] Jerry Gao, Chunli Xie, and Chuanqi Tao. 2016. Big Data Validation and Quality Assurance – Issuses, Challenges, and Needs. In *2016 IEEE Symposium on Service-Oriented System Engineering*. doi:10.1109/SOSE.2016.63

[22] Bethany L Hedt-Gauthier, Lyson Tenthani, Shira Mitchell, Frank M Chimbwandira, Simon Makombe, Zengani Chirwa, Erik J Schouten, Marcello Pagano, and Andreas Jahn. 2012. Improving data quality and supervision of antiretroviral therapy sites in Malawi: An application of Lot Quality Assurance Sampling. *BMC Health Services Research* (2012). doi:10.1186/1472-6963-12-196

[23] Mohamad Taha Ijab, Ely Salwana Mat Surin, and Norshita Mat Nayan. 2019. Conceptualizing big data quality framework from a systematic literature review perspective. *Malaysian Journal of Computer Science* (2019). doi:10.22452/mjcs.sp2019no1.2

[24] DAMA International. 2017. *The DAMA Guide to the Data Management Body of Knowledge (DAMA-DMBOK)* (2nd ed.). Technics Publications.

[25] Sean Kandel, Ravi Parikh, Andreas Paepcke, Joseph M Hellerstein, and Jeffrey Heer. 2012. Profiler: Integrated statistical analysis and visualization for data quality assessment. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*. 547–554. doi:10.1145/2254556.2254659

[26] Sesillia Fajar Kristyanti, Tien Fabrianti Kusumasari, and Ekky Novriza Alam. 2020. Operational Dashboard Development as A Data Quality Monitoring Tools Using Data Deduplication Profiling Result. In *Proceedings - 2020 6th International Conference on Science and Technology, ICST 2020*. doi:10.1109/ICST50505.2020.9732870

[27] Gómez-Omella Meritxell, Basilio Sierra, and Susana Ferreiro. 2022. On the Evaluation, Management and Improvement of Data Quality in Streaming Time Series. *IEEE Access* (2022). doi:10.1109/ACCESS.2022.3195338

[28] P. Missier, G. Lalk, V. Verykios, F. Grillo, T. Lorusso, and P. Angeletti. 2003. Improving data quality in practice: A case study in the italian public administration. *Distributed and Parallel Databases* (2003). doi:10.1023/A:1021548024224

[29] Per Myrseth, Jørgen Stang, and Vibeke Dalberg. 2011. A data quality framework applied to e-government metadata: A prerequsite to establish governance of interoperable e-services. In *2011 International Conference on E-Business and E-Government, ICEE2011 - Proceedings*. doi:10.1109/ICEBEG.2011.5881298

[30] Fred Nsubuga, Henry Luzze, Immaculate Ampeire, Simon Kasasa, Opar Bernard Toliva, and Alex Ario Riolexus. 2018. Factors that affect immunization data quality in Kabarole District, Uganda. *PLoS ONE* (2018). doi:10.1371/journal.pone.0203747

[31] Leo L. Pipino, Yang W. Lee, and Richard Y. Wang. 2002. Data quality assessment. *Commun. ACM* (2002). doi:10.1145/505248.506010

[32] Maria Priestley, Fionntán O'donnell, and Elena Simperl. 2023. A Survey of Data Quality Requirements That Matter in ML Development Pipelines. *J. Data and Information Quality* (2023). doi:10.1145/3592616

[33] Arie Purwanto, Anneke Zuiderwijk, and Marijn Janssen. 2020. Citizens' trust in open government data: a quantitative study about the effects of data quality, system quality and service quality. In *Proceedings of the 21st Annual International Conference on Digital Government Research*. 310–318. doi:10.1145/3396956.3396958

[34] Muhammad Badriansyah Putra, Fahmi Alaydrus, Ira Sulistyowati, Teguh Raharjo, and Riko Wijayanto. 2022. Issues and Challenges of the Data Analytics Development Project in The Center of Information System and Financial Technology. In *2022 1st International Conference on Information System & Information Technology*. doi:10.1109/ICISIT54091.2022.9872715

[35] Thomas C Redman. 1998. The impact of poor data quality on the typical enterprise. *Commun. ACM* 41, 2 (1998), 79–82.

[36] Fedri Ruluwedrata Rinawan, Afina Faza, Ari Indra Susanti, Wanda Gusdya Purnama, Noormarina Indraswari, Didah, Dani Ferdian, Siti Nur Fatimah, Ayi Purbasari, Arief Zulianto, Atriany Nilam Sari, Intan Nurma Yulita, Muhammad Fiqri Abdi Rabbi, and Riki Ridwana. 2022. Posyandu Application for Monitoring Children Under-Five: A 3-Year Data Quality Map in Indonesia. *ISPRS International Journal of Geo-Information* (2022). doi:10.3390/ijgi11070399

[37] Mujiono Sadikin, Purwanto S. Katidjan, Arif Rifai Dwiyanto, Nurfiyah, Ajif Yunizar Pratama Yusuf, and Adi Trisnojuwono. 2025. Improving the MSMEs data quality assurance comprehensive framework with deep learning technique. *Indonesian Journal of Electrical Engineering and Computer Science* (2025). doi:10.11591/ijeecs.v37.i1.pp613-626

[38] Tanapat Samakit, Chutiporn Anutariya, and Marut Buranarach. 2023. QUALYST: Data Quality Assessment System for Thailand Open Government Data. In *Proceedings of JCSSE 2023 - 20th International Joint Conference on Computer Science and Software Engineering*. doi:10.1109/JCSSE58229.2023.10202060

[39] Mariutsi Alexandra Osorio Sanabria, Ferney Orlando Amaya Fernández, and Mayda Patricia González Zabala. 2018. Colombian Case Study for the Analysis of Open Data Government: a Data Quality Approach. In *ICEGOV '18, April 4–6, 2018, Galway, Ireland*. Association for Computing Machinery. doi:10.1145/3209415.3209474

[40] Flavia Serra, Verónika Peralta, Adriana Marotta, and Patrick Marcel. 2024. Use of Context in Data Quality Management: A Systematic Literature Review. *J. Data and Information Quality* (2024). doi:10.1145/3672082

[41] Christian Sillaber, Clemens Sauerwein, Andrea Mussmann, and Ruth Breu. 2016. Data Quality Challenges and Future Research Directions in Threat Intelligence Sharing Practice. *WISCS'16, October 24 2016* (2016). doi:10.1145/2994539.2994546

[42] Ahmet Soylu, Óscar Corcho, Brian Elvesæter, Carlos Badenes-Olmedo, Francisco Yedro-Martínez, Matej Kovacic, Matej Posinkovic, Mitja Medvešček, Ian Makgill, Chris Taggart, Elena Simperl, Till C. Lech, and Dumitru Roman. 2022. Data Quality Barriers for Transparency in Public Procurement. *Information (Switzerland)* (2022). doi:10.3390/info13020099

[43] Justin St-Maurice and Catherine Burns. 2017. An Exploratory Case Study to Understand Primary Care Users and Their Data Quality Tradeoffs. In *J. Data and Information Quality*. Association for Computing Machinery. doi:10.1145/3058750

[44] Ikbal Taleb, Mohamed Adel Serhani, Chafik Bouhaddioui, and Rachida Dssouli. 2021. Big data quality framework: a holistic approach to continuous quality management. *Journal of Big Data* (2021). doi:10.1186/s40537-021-00468-0

[45] Jaak Tepandi, Mihkel Lauk, Janar Linros, Priit Raspel, Gunnar Piho, Ingrid Pappel, and Dirk Draheim. 2017. The data quality framework for the Estonian public sector and its evaluation: establishing a systematic process-oriented viewpoint on cross-organizational data quality. In *Transactions on Large-Scale Data-and Knowledge-Centered Systems XXXV*. Springer, 1–26. doi:10.1007/978-3-662-56121-8_1

[46] David Tien. 2010. Project management and data quality control. In *Proceedings - 2010 IEEE International Conference on Emergency Management and Management Sciences, ICEMMS 2010.* doi:10.1109/ICEMMS.2010.5563378

[47] Marco Torchiano, Antonio Vetrò, and Francesca Iuliano. 2017. Preserving the Benefits of Open Government Data by Measuring and Improving Their Quality: An Empirical Study. *2017 IEEE 41st Annual Computer Software and Applications Conference* (2017). doi:10.1109/COMPSAC.2017.192

[48] Strong Diane M Wang, Richard Y. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems* 12, 4 (1996), 5–33. doi:10.1080/07421222.1996.11518099

[49] Xinhao Wang, Lulin Xu, Qin Zhang, Da Zhang, and Xiliang Zhang. 2022. Evaluating the data quality of continuous emissions monitoring systems in China. *Journal of Environmental Management* (2022). doi:10.1016/j.jenvman.2022.115081

[50] Claes Wohlin, Per Runeson, Martin Host, Magnus C Ohlsson, Bjorn Regnell, and Anders Wesslen. 2012. *Experimentation in Software Engineering* (2012 ed.). Springer, Berlin, Germany.

[51] Hongjiang Xu. 2015. What Are the Most Important Factors for Accounting Information Quality and Their Impact on AIS Data Quality Outcomes? *ACM Journal of Data and Information Quality* (2015). doi:10.1145/2700833

[52] Li Ya, Song Heliang, and Xu Yingcheng. 2020. Method for Calculating the Weights of Internet + Government Service Data Quality Assessment Indexes Based on Analytic Hierarchy Process. In *Journal of Physics: Conference Series*. doi:10.1088/1742-6596/1584/1/012043

[53] Peter Z. Yeh and Colin A. Puri. 2010. An Efficient and Robust Approach for Discovering Data Quality Rules. In *Proceedings of the 2010 22nd IEEE International Conference on Tools with Artificial Intelligence - Volume 01 (ICTAI '10)*. IEEE Computer Society, USA, 248–255. doi:10.1109/ICTAI.2010.43

[54] Alivia Yulfitri. 2016. Modeling operational model of data governance in government: Case study: Government agency X in Jakarta. In *2016 International Conference on Information Technology Systems and Innovation*. doi:10.1109/ICITSI.2016.7858207

[55] Zahirah Zainuddin and Emelia Akashah P. Akhir. 2024. Systematic Literature Review of Data Quality in Open Government Data: Trend, Methods, and Applications. *IEEE Access* (2024). doi:10.1109/ACCESS.2024.3475577

[56] Pengcheng Zhang, Xuewu Zhou, Wenrui Li, and Jerry Gao. 2017. A Survey on Quality Assurance Techniques for Big Data Applications. In *2017 IEEE Third International Conference on Big Data Computing Service and Applications*. doi:10.1109/BigDataService.2017.42

[57] Ruojing Zhang, Marta Indulska, and Shazia Sadiq. 2019. Discovering Data Quality Problems: The Case of Repurposed Data. *Business and Information Systems Engineering* (2019). doi:10.1007/s12599-019-00608-0

[58] Zheng Zhu, Yingjie Tian, and Hongshan Yang. 2024. Research on Power Data Quality Analysis Method Based on Verification Rules in Big Data Environment. *CIBDA 2024 - China* (2024). doi:10.1145/3671151.3671359