Protocol: A Rapid Systematic Review of Data Quality Frameworks and Evaluation

Jack D'Arcy, Emma Howard

Administrative Information

Title: A Rapid Systematic Review of Data Quality Frameworks and Evaluation

Registration:

Authors: Jack D'Arcy (JD), Emma Howard (EH)

Contact: School of Computer Science and Statistics, Trinity College Dublin.
jadarcy@tcd.ie, emhoward@tcd.ie

Contributions:

Introduction

Rationale: Data quality has become increasingly important to businesses, yet it remains a complex and multifaceted field, which can be seen in the varying definitions within academic literature. As information systems have evolved from monolithic to network-based frameworks, data characteristics such as volume, sources, and generation methods have become progressively more complex [1, 2]. Despite a lack of standardisation in defining or assessing data quality, its importance is widely recognised; poor data can impair decision-making and operations, while high-quality data is essential for organisational success [3, 4].

In general, data quality is evaluated through frameworks or methodologies that assess processes from input to evaluation and improvement [5, 3]. Usually these frameworks address key dimensions such as completeness, timeliness, accuracy, consistency, and accessibility. However, despite the considerable research, and various frameworks proposed for data quality assessment, there is limited critical evaluation of them, particularly under real-world environments. This gap is particularly evident in complex environments like the Internet of Things (IoT), where traditional frameworks may not fully address continuously evolving challenges.

The rapid expansion of IoT has introduced additional complexity to data quality management. Initially described as a network of internet-connected objects, primarily sensors [3], the IoT has grown into a global network of interconnected devices that generate large volumes of data across diverse industries [2]. This increase in data volume further highlights the importance of data quality, as errors and inconsistencies in IoT data can have broader impacts. However, research in this field often has two significant limitations. Firstly, the increase of the heterogeneity of data sources has led to more semi-structured and unstructured data in IoT, despite research being more focused on structured data [3]. Additionally, existing research on data quality frameworks for IoT often remains very general resulting in less applicable outputs for specific sectors.

Given the breadth of research on this topic, a rapid systematic review is particularly suited to refining data quality frameworks, evaluation methods, and error-identification techniques for specific sectors [6]. This rapid systematic review aims to build upon two prior systematic reviews [7, 8], which examined data quality frameworks. Both studies give detailed overviews of data quality frameworks but their focus remained broad, primarily addressing the IoT industry as a whole and, in some cases, specific sectors like manufacturing and healthcare. That said, the findings by Goknil et al. [8] are very complementary to this review due to its recent publication and extensive scope, which includes evaluations of error-identification techniques and data quality frameworks. This review aims to provide a more tailored analysis, focusing specifically on the telecommunications (telecom) and software as a

service (SaaS) sectors, as well as expanding the focus for evaluation methodologies to include operational metrics in addition to the theoretical evaluation methods explored in [8].
Ultimately, this review seeks to identify implementable frameworks for assessing data quality, outline evaluation methodologies of these frameworks, and describe the relevant error detection techniques in these frameworks.

Objectives: The aim of this rapid systematic review of data quality frameworks and evaluation is to explore various data quality frameworks and how they are implemented, evaluated and identify errors. The
objectives are:

- To identify the data quality frameworks used in the IoT industry, especially relevant sectors such as the automotive, agricultural and transportation industry
- To evaluate how these frameworks are applied and assessed, encompassing both theoretical and operational metrics
- To summarise the methods used within the selected frameworks to identify errors

Methods

Eligibility criteria: Studies will be included (excluded) in the scoping review based on the following criteria:

- Studies must be published in English
- Studies published before 2015 will be excluded
- Studies must include a discussion of data quality frameworks implemented in IoT-related industries
- Studies must propose internal data quality frameworks that assess the quality of collected data
- Studies must focus on industries such as automotive, telecommunications, transportation or very closely related domains that offer broader relevance to these industries.
- Studies must evaluate the frameworks or, at a minimum, explain how these frameworks are applied
- The study must be a primary research study in its final publication stage

Information Sources: As the literature on data quality, especially in the IoT sector, is continuously evolving and the varied application of data quality frameworks across industries, the information sources for this review will be compiled from databases (listed below). Additionally, reference lists from foundational reviews in data quality and IoT, such as Goknil et al. (2023), will be reviewed to refine the search strategy.

The following databases will be searched:

- ACM Digital Library
- SCOPUS
- Web of Science
- IEEE Xplore

Search Strategy: The literature search will be limited to publications in English and will include a focus on quantitative studies. The search will examine article titles, abstracts, keywords, and publication titles, adapting to the search capabilities of each database. The general draft search string used as a starting point was:

("IoT" OR "Internet of Things" OR "teleco*" OR "software as a service" OR "SaaS" OR “telecommunication*”)
AND
("data quality" OR "data integrity" OR "data completeness" OR "DQ")
AND
("framework*" OR "evaluation") OR "error detection" OR "error identification")

Given database-specific search field limitations and query conventions (e.g., variations in field tags such as Title, Abstract, and different usage of Boolean operators), this general search string was adapted individually to each database. Full details of these database-specific search queries, including syntax and initial search results counts, are provided in Appendix A for transparency and replicability.

Study Records

Data Management: Literature search results will be downloaded into the Rayyan software tool. Firstly, all duplicates will be removed. Rayyan was chosen as it is a collaborative software tool, and allows authors to work on the search results list simultaneously.

Selection Process: Once duplicates have been removed, an initial sample of 50 abstracts will be screened independently by both authors using predefined inclusion criteria. Any discrepancies between the authors will be discussed until a unanimous consensus is reached. Then, the remaining abstracts will be screened individually by the first author. In the second round of screening, full-text publications will be reviewed. This round will begin with both authors independently screening the same five studies to ensure consistency in applying the inclusion criteria. Any discrepancies will be dealt with similarly to the first round. After this, the remaining publications will be screened individually by the primary reviewer. The PRISMA 2020 flow diagram will summarise the selection process, providing an explanation for any exclusions made during the second round [9].

Data Items: The following items will be collected from each study:

- Reference Details: Title, authors, year of publication, data source.
- Institution Details: Name of institution(s) (if given), type of institution, country of institution(s), relevant industry focus (e.g., automotive, agriculture, or telecommunications if specified).
- Framework Characteristics: Description of data quality framework(s) evaluated, including key dimensions assessed and type of framework (standardised or tailored), and if tailored, the specific industry or context.
- Data Quality Assessment Metrics: Specific metrics or criteria used for data quality assessment.
- Evaluation Method: Summary of how the specified framework was evaluated and if this method is theoretical, practical or both
- Error Identification Methods: A list of the various approaches to identify errors within the data streams
- Results: Key findings regarding framework effectiveness, applicability to IoT or telecom environments, and any statistical significance of results. Notable challenges or limitations identified in the framework’s application (e.g., issues with unstructured data, heterogeneity of sources).

Outcomes and Prioritisation: The primary outcome will be a comprehensive summary of data quality frameworks implemented in IoT industries, along with a quality evaluation of these frameworks. The review will also focus on the methods used within these frameworks to identify and address errors in data streams. Secondary outcomes will include an analysis aimed at identifying aspects of standardised frameworks that can promote uniform decision-making across IoT applications, as well as highlighting specific circumstances where certain frameworks may be less applicable or require adaptation.

Data Synthesis: Descriptive data from each study will be presented in a table format (see Table 1 - 3 below) and a narrative synthesis of the studies will be presented.

Table 1: Summary of Data Quality Frameworks and Characteristics

Source	Framework Type	Purpose	Key Components	Characteristics

Table 2: Data Quality Assessment Metrics and Evaluation Methods

Study	Evaluation Methods	Key Metrics to Assess the Framework	Effectiveness of Framework

Table 3: Error Identification Techniques

Study	Error Identification Techniques	Types of Errors Identified	Limitations

Appendix A

WEB OF SCIENCE: 576

(TI=("IoT" OR "Internet of Things" OR teleco* OR "software as a service" OR SaaS OR telecommunication*)
OR
AB=("IoT" OR "Internet of Things" OR teleco* OR "software as a service" OR SaaS OR telecommunication*))
AND
(TI=("data quality" OR "data integrity" OR "data completeness" OR DQ)
OR
AB=("data quality" OR "data integrity" OR "data completeness" OR DQ))
AND
(TI=("framework*" OR "evaluation" OR "error detection" OR "error identification")
OR
AB=("framework*" OR "evaluation" OR "error detection" OR "error identification"))

ACM DIGITAL LIBRARY: 19

(Abstract:(("IoT" OR "Internet of Things" OR "teleco*" OR "software as a service" OR "SaaS")) OR ContentGroupTitle:(("IoT" OR "Internet of Things" OR "teleco*" OR "software as a service" OR "SaaS")))
AND
(Abstract:(("data quality" OR "data integrity" OR "data completeness" OR "DQ")) OR ContentGroupTitle:(("data quality" OR "data integrity" OR "data completeness" OR "DQ")))
AND
(Abstract:(("framework*" OR "evaluation" OR "error detection" OR "error identification")) OR ContentGroupTitle:(("framework*" OR "evaluation" OR "error detection" OR "error identification")))

IEEE EXPLORE: 268

(("Abstract":"IoT" OR "Abstract":"Internet of Things" OR "Abstract":"teleco*" OR "Abstract":"software as a service" OR "Abstract":"SaaS" OR "Abstract":"telecommunication*")
OR
("Document Title":"IoT" OR "Document Title":"Internet of Things" OR "Document Title":"teleco*" OR "Document Title":"software as a service" OR "Document Title":"SaaS"))
AND
(("Abstract":"data quality" OR "Abstract":"data integrity" OR "Abstract":"data completeness" OR "Abstract":"DQ")
OR
("Document Title":"data quality" OR "Document Title":"data integrity" OR "Document Title":"data completeness" OR "Document Title":"DQ"))
AND
(("Abstract":"framework*" OR "Abstract":"evaluation" OR "Abstract":"error detection" OR "Abstract":"error identification")
OR
("Document Title":"framework*" OR "Document Title":"evaluation" OR "Document Title":"error detection" OR "Document Title":"error identification"))

SCOPUS: 758

(TITLE("IoT" OR "Internet of Things" OR teleco* OR "software as a service" OR SaaS OR telecommunication*)
OR ABS("IoT" OR "Internet of Things" OR teleco* OR "software as a service" OR SaaS OR telecommunication*))
AND

```
(TITLE("data quality" OR "data integrity" OR "data completeness" OR DQ)
OR ABS("data quality" OR "data integrity" OR "data completeness" OR DQ))
AND
(TITLE("framework*" OR "evaluation" OR "error detection" OR "error identification")
OR ABS("framework*" OR "evaluation" OR "error detection" OR "error identification"))
```

References

1. Taleb I, Serhani MA, Bouhaddioui C. Big data quality framework: a holistic approach to continuous quality management. Journal of Big Data. 2021;8(1):76. doi:10.1186/s40537-021-00468-0

2. Karkouch A, Mousannif H, Al Moatassime H, Noel T. Data quality in internet of things: a state-of-the-art survey. Journal of Network and Computer Applications. 2016;73:57–81. doi:10.1016/j.jnca.2016.08.002

3. Zhang L, Jeong D, Lee S. Data quality management in the internet of things. Sensors. 2021;21(17):5834.

4. Vetrò A, Canova L, Torchiano M, Minotas CO, Iemma R, Morando F. Open data quality measurement framework: definition and application to open government data. Government Information Quarterly. 2016;33(2):325–337. doi:10.1016/j.giq.2016.02.001

5. Micic N, Neagu D, Campean F, Zadeh EH. Towards a data quality framework for heterogeneous data. In: 2017 IEEE International Conference on Internet of Things (iThings), IEEE Green Computing and Communications (GreenCom), IEEE Cyber, Physical and Social Computing (CPSCom), and IEEE Smart Data (SmartData). IEEE; 2017. p.155–162.

6. Mulrow CD. Systematic reviews: rationale for systematic reviews. BMJ. 1994;309(6954):597–599.

7. Liu C, Nitschke P, Williams SP, Zowghi D. Data quality and the Internet of Things. Computing. 2019;102(2):573–599. doi:10.1007/s00607-019-00746-z

8. Goknil A, Nguyen P, Sen S, Politaki D, Niavis H, Pedersen KJ, et al. A systematic review of data quality in CPS and IoT for Industry 4.0. ACM Computing Surveys. 2023;55(14s):1–38.

9. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Systematic Reviews. 2021;10(1). doi:10.1186/s13643-021-01626-4