Bulletin of Electrical Engineering and Informatics
Vol. 12, No. 6, December 2023, pp. 3736∼3757
ISSN: 2302-9285, DOI: 10.11591/eei.v12i6.5667

A systematic literature review on data quality assessment

Oumaima Reda, Naoual Chaouni Benabdellah, Ahmed Zellou
Software Project Management (SPM) Team, National School of Computer Science and System Analysis (ENSIAS),
Mohammed V University in Rabat, Rabat, Morroco

Article Info

Article history:

Received Jan 1, 2023
Revised Mar 24, 2023
Accepted Apr 17, 2023

Keywords:

Data quality
Quality assessment
Quality models
Systematic literature review

ABSTRACT

Defining and evaluating data quality can be a complex task as it varies depending on the specific purpose for which the data is intended. To effectively assess data quality, it is essential to take into account the intended use of the data and the specific requirements of the data users. It is important to recognize that a standardized approach to data quality assessment (DQA) may not be suitable in all cases, as different uses of data may have distinct quality criteria and considerations. In order to advance research in the field of data quality, it is useful to determine the current state of the art by identifying, evaluating, and analyzing relevant research conducted in recent years. In light of this objective, the study proposes a systematic literature review (SLR) as a suitable approach to examine the landscape of data quality and investigate available research specifically pertaining to DQA. The findings of our SLR clearly reveal and demonstrate the criticality of data quality and point to new directions for future study and have consequences for researchers and practitioners interested in defining and assessing data quality.

Corresponding Author:

Oumaima Reda
Software Project Management Team, National School of Computer Science and System Analysis (ENSIAS)
Mohammed V University in Rabat
Rabat, Morocco
Email: oumaima_reda@um5.ac.ma

1. INTRODUCTION

Reliable research demands data of known quality, however measuring the quality of data is a critical step in any data analysis task. Data quality has been a hot topic for many years and continues to receive a great deal of attention [1]-[3], as the value of data is highly dependent on its quality and ease of use. This means that the correct use of high-quality data can help to better plan, analyze, and decide. However, poor-quality data can be inappropriate for the intended purpose and the consequences can be serious. Studies on data quality have been carried out in various fields, ranging from data-driven domains such as big data and statistics to domain-driven domains such as healthcare, finance, and information systems.

To enhance research in the field of data quality, it is important to have a comprehensive understanding of the current state of knowledge. This involves identifying, evaluating, and analyzing relevant research conducted in recent years. Hence, this research paper is intended to provide an in-depth review and investigation on data quality in order to answer the question of how data quality can effectively be assessed. Through the systematic literature review (SLR), a rigorous review and analysis of empirical studies published between 2016 and 2021 were conducted. This process enabled the identification of existing research domains associated with data quality. Furthermore, the study compiled various data quality models proposed by other researchers, providing

a comprehensive overview of the different conceptual frameworks in use. Additionally, the research delved into proposed methodologies, metrics for measuring data quality, shedding light on the tools, and approaches available for data quality assessment (DQA).

The findings obtained through the SLR reveal the critical nature of data quality. They highlight the significance of ensuring high-quality data and emphasize the need for ongoing research in this area. The identified gaps and emerging trends in the field of data quality suggest new directions for future studies. These findings carry implications for both researchers and practitioners, offering valuable insights for defining, and assessing data quality effectively. Overall, the SLR contributes to advancing knowledge in the field of data quality by providing a comprehensive synthesis of existing research, identifying key areas for further exploration, and offering guidance to researchers and practitioners interested in enhancing their understanding and management of data quality.

The remainder of this research study is structured as follows: in section 2, we present the background and preliminaries, including an overview of data quality, data quality dimensions, and DQA. We discuss the research method and define the following comprehensive phases of the systematic review in section 3. Section 4 provides and analyzes our findings, while section 5 describes potential research areas and highlights the review's limitations. Lastly, in section 6 we give our findings and future work.

2. BACKGROUND

Data quality and DQA form the central focus of this systematic review, this section aims to provide an introductory understanding of these key concepts. By examining various dimensions and factors that contribute to data quality, this systematic review aims to shed light on effective approaches for assessing and improving data quality.

2.1. Data quality and data quality dimensions

The ISO 9000 standard defined the term "quality" as the extent to which the consumer needs are met, by considering all characteristics required by the customer for the product or service [4]. Generally, data quality concept refers to the adequacy of data to achieve the intended purpose [5]. In other words, it is a requirement that the user anticipates executing or a data value that the user anticipates obtaining. It is known as a multi-dimensional construct, since that several elements must be taken into account. These elements are described by data quality dimensions, which are evaluated using predetermined metrics [6].

Basically, it can be seen that the same data have various uses, this can lead some users to judge the quality of data to be high, while others judge it to be low. Thus, this quality of data may have a dual characteristic and be subjective, as a result, it may meet expectations and specifications. In other words, data quality is context dependent [7]. In order to make a decision, relevant information is sought from the data after it has been analyzed and processed in order to indicate the level of quality, by performing evaluative measures using a specific DQA model [8].

According to Wang and Strong [1], data quality dimensions are a collection of data quality attributes that each reflect a distinct feature or construct of data quality, i.e. each dimension concerns a specific aspect. Therefore, measuring data quality requires conducting DQAs to help determine how well this data adequately meet user needs [9]. The literature on data quality encompasses a range of dimensions, with accuracy, completeness, consistency, and timeliness emerging as the most frequently mentioned ones. These dimensions capture crucial aspects of data quality and serve as foundational elements for assessing data reliability and fitness for purpose. However, it is worth noting that researchers have proposed various categorizations and definitions for data quality dimensions, resulting in multiple categories being identified in the literature [10]-[13]. For instance, Weikum [14] develops a visionary classification of data quality criteria, distinguishing between system, process and data-focused criteria as shown in Table 1.

Table 1. Classification criteria of [14]

System-centric notions	Process-centric notions	Data-centric notions
Reliability, availability, integrity, security, performance and verifiability	Safety properties and liveness properties	Accuracy, comprehensiveness, timeliness, credibility, cost-effectivity and latency

2.2. Data quality assessment

DQA is a crucial component of every data analysis operation. It is the process of determining whether the data meet a user's information requirements in a particular use case [7]. It involves measuring the quality dimensions or criteria that are relevant and comparing the findings to the user's quality needs [15]. One of the challenges in DQA arises from the contextual nature of data quality. It is essential to consider the specific context in which data is used, as the evaluation of data quality is highly dependent on the intended purpose and the particular scenario. The same quality dimension may hold different relevance and require distinct evaluation methods in different contexts [7]. This contextual variability adds complexity to the assessment process and necessitates a nuanced understanding of the specific circumstances surrounding data usage.

Quality models play a significant role in DQA by providing detailed specifications of quality measures. These models offer guidance on the selection and application of appropriate measures for assessing data quality. They outline the definitions, scales, and formulas associated with each measure, enabling researchers and practitioners to determine the most suitable approach for measuring specific aspects of data quality. By indicating which measures are relevant and how they should be measured, quality models contribute to a standardized and systematic evaluation of data quality [16]. This ensures consistency and facilitates meaningful comparisons across different datasets and contexts. The utilization of quality models enhances the effectiveness and accuracy of DQA processes. Indeed, judging the quality of data frequently necessitates the computation of a large number of quality measures rather than a single measure [8].

3. RESEARCH METHOD

The goal of this research is to perform a comprehensive literature review on DQA using the original SLR guidelines given in [17]. According to these guidelines, our SLR consists of three essential activities: planning, conducting, and reporting the review [17]. Each activity is associated with several steps. Figure 1 depicts a summary of the study's implementation.

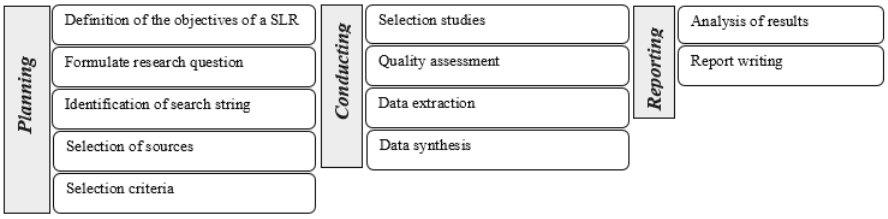


Figure 1. SLR process

3.1. Main objective

The main objective of this SLR is to provide a comprehensive overview of recent research conducted in the field of data quality. The focus is on understanding and evaluating the existing approaches, methodologies, and techniques used for DQA. By analyzing and assessing these approaches, the aim is to enhance the overall understanding of data quality and provide insights that can improve the accuracy and reliability of research outcomes. By achieving this objective, the SLR aims to provide valuable insights and recommendations for researchers, practitioners, and decision-makers in their pursuit of defining, evaluating, and improving data quality. Ultimately, the goal is to contribute to the overall improvement of DQA practices, enhance the credibility, and impact of research outcomes.

3.2. Research questions

As research questions are used to determine and guide the review process, refining them is the most challenging task of any systematic review. The following are our research questions:

- RQ1: what are the existing research domains related to data quality?
- RQ2: what data quality models are crucial for assessing data quality?
- RQ3: which methodologies were utilized to assess data quality?
- RQ4: what are the quality metrics used in DQA?

3.3. Search process

Besides the research questions that guide the SLR, some added questions need to be considered in order to begin and build a good research strategy [17]. First of all, it is necessary to specify which time period should be taken into account, what search strings to be searched, in which searching database or sources, and what are the criteria to be used for selecting studies. We answer respectively these questions in the next sections.

3.3.1. Timing

In our study, we focused on recent works by setting a specific timeframe for our search process. We limited our search to articles published between 2016 and 2021 to ensure that we captured the most up-to-date research in the field of data quality. This timeframe allows us to gain insights into current trends, advancements, and emerging areas of research within the specified time period.

3.3.2. Search string

The search string should be structured to reflect the words frequently used in the titles of relevant articles found by the reference search results. Keywords for searching articles were created using numerous criteria, including the identification of keywords based on research questions, use of synonyms, acronyms, and alternative spellings, and use of the Boolean 'AND' and 'OR' operators to connect keywords. This resulted in the following final search string: ("data quality" AND ("quality assessment" OR "quality evaluation")).

3.3.3. Sources and selection criteria

Manually searching for all relevant articles in conferences and journals is extremely time consuming, an online search was performed in three high quality digital libraries (Scopus, ACM, and DBLP) as they include all relevant publications and conferences proceedings. After doing preliminary research on the three digital libraries, the resulting collection of articles must be sorted based on titles, years of publication, keywords, and abstracts. We then used the following criteria shown in Table 2 to determine whether or not to choose preliminary research for further processing.

Table 2. Inclusion and exclusion criteria used

Inclusion criteria	Exclusion criteria
Studies published between 2016 and 2021	Duplicate studies
Studies published in English	Non-peer-reviewed studies
Full studies focusing on data quality area	Magazines, tutorials, and editorials
Studies pertaining to our research questions (RQ1-RQ4)	Research that do not consider data quality

3.4. Quality assessment

After the full-text reading of the articles, the quality of each publication was assessed to prove and strengthen the efficiency of our study. So that, we attempt to provide a quality assessment model that can be used to determine the relevance of each paper and whether it contains the requested information, based on its score. Thus, our model is provided as a formula that uses seven assessment criteria (AC) whose values are either 1 or 0 (i.e. yes or no) to compute and evaluate the final score of each paper by giving a coefficient θ to each AC based on its level of relevance. Our quality evaluation model is represented by:

$$S(m) = \sum_{i=1}^{7} \theta_i AC_i$$

High coefficient for papers proposing a novel quality assessment model (AC2) and suggesting evaluation metrics (AC3), medium coefficient for papers presenting a framework or a quality evaluation methodology (AC4) and include simulations and prototypes (AC5). We establish the lowest coefficient for articles that specify data quality (AC1), outline a state of the art (AC6), and perform polling (AC7). Table 3 summarizes our quality assessment model.

3.5. Data extraction and synthesis

Data extraction is conducted for each of the 100 publications chosen and the results are displayed in an Excel file. Titles, authors, year of publication, publication type, and the domain of application were the columns included in the collected data for all the sources in this study. Following the extraction step, the retrieved data is analyzed to answer the aforementioned research questions. During data synthesis phase, we collected, processed, and summarized the results of relevant studies in order to answer our research questions.

Data was presented in tables, graphs, plots, and charts. The results of this study are then shown as maps to help better understand the data quality landscape.

Table 3. Quality AC

AC	Description	Level	θ
AC1	Data quality definition	Low	1
AC2	Quality evaluation model	High	2
AC3	Assessment metrics	High	2
AC4	Methodology or framework	Medium	1.5
AC5	Simulation or prototype	Medium	1.5
AC6	State of the art	Low	1
AC7	Conducting a polling	Low	1

4. RESULTS

This section presents the findings of SLR, which was conducted using the research method described in section 3. We discuss the overview of the selected publications and provide an analysis of the results.

4.1. Overview of selected studies

A set of 987 publications that were found through automated searches of electronic data sources were utilized during the search procedure. The titles were then checked for duplicate research and 794 candidates were ready to move on to the next round of processing using the previously described inclusion and exclusion criteria. After the exclusion criteria were applied, the number of papers that needed to be read for the systematic review dropped from 794 to 340, only 138 of those 340 publications were deemed to be "relevant". As a consequence, the seven quality evaluation criteria listed in Table 3 were used to ensure that the incorporated findings will contribute significantly to SLR. As a result, 52 papers with a score of less than or equal to 2 were eliminated. Eventually, 100 papers were chosen to respond four research questions. Figure 2 gives a comprehensive overview of selection process.

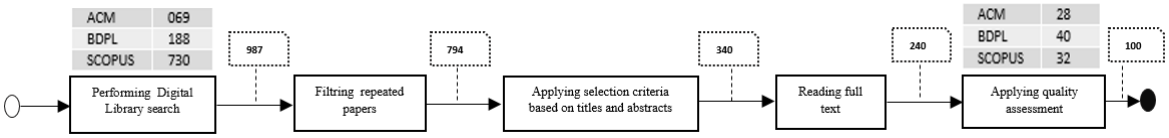


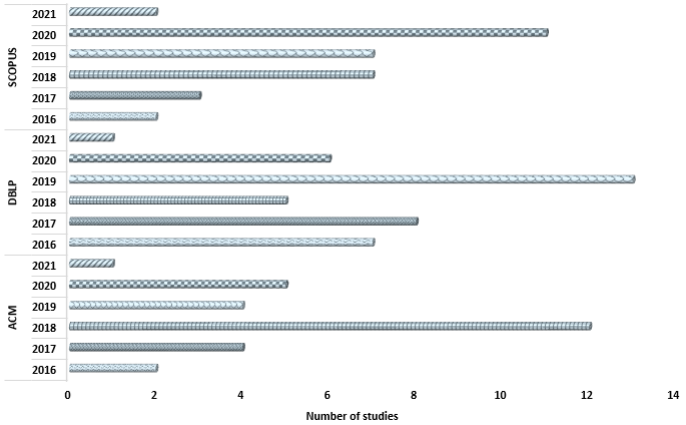
Figure 2. A summary of selecting process

4.2. Classification of selected research

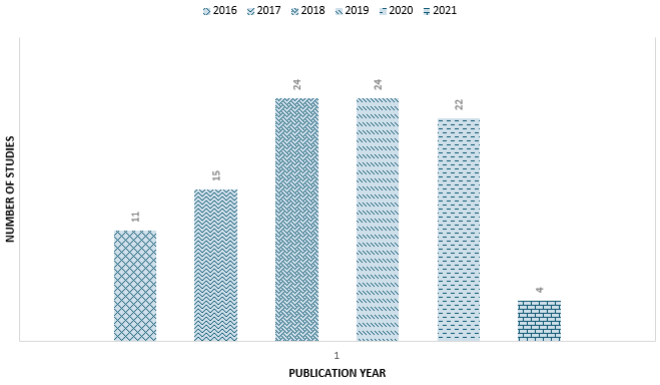
We further assessed the outcomes in order to provide a summary of the publishing trends in area of data quality. The results of search process are displayed in Table 4, along with the number of studies that were chosen based on the data sources and the years of publication. Figures 3(a) and (b) shows the distribution of 100 relevant publications by year (2016-2021). Our results show that the number of articles on data quality starts to increase from 2018 with 74 articles published during the last 4 years, although the results for 2021 are not conclusive because the research was done at the start of 2021. Nevertheless, in 2016 and 2017, 11 and 15 papers are published, which proves that data quality research is taking a great prominence and the number of research articles published is increasing every year. Next, as can be seen from Figure 3(c), which displays the distribution of papers by data sources, the majority of the articles that were chosen—40 publications, or 40%—were published in DBLP.

Table 4. Search process result

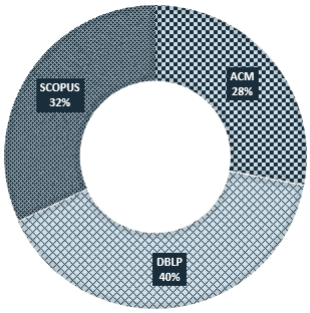
	2016	2017	2018	2019	2020	2021
ACM	2	4	12	4	5	1
DBLP	7	8	5	13	6	1
Scopus	2	3	7	7	11	2
Total	11	15	24	24	22	4



(a)



(b)



(c)

Figure 3. Reviewed paper statistics: (a) number of publications per source, (b) studies per years, and (c) source distribution of selected papers

4.3. Results of data extraction

Throughout the systematic review, we addressed four research questions to guide our investigation into the field of DQA. These questions were designed to explore the existing literature, analyze the methodologies employed, and identify key findings and trends. The findings we report in this section are directly related to these research questions and providing valuable insights into the current state of knowledge in the field.

4.3.1. RQ1: what are the existing research domains related to data quality?

Data quality field has captured the attention of researchers since many years. According to Bertossi and Rizzolo [18], the assessment of data quality is context-dependent, this means that data quality must be considered closely related to the intended use of the data and can only be assessed in that context. So that, the topics of discussion are very broad. Classifying research domains in the field of data quality is crucial for directing researchers' attention towards the least explored topics and identifying areas that require further investigation. Through our systematic review, we have observed that numerous studies have focused on the same topic but have taken different approaches and perspectives. This highlights the breadth and diversity of data quality research across various domains. Some of the domains that have been extensively studied include:

- Big data: with the exponential growth of data, research in data quality within the context of big data has become increasingly important. This domain explores the challenges and techniques for ensuring data quality in large-scale datasets [19].
- Artificial intelligence (AI): with the growing role of AI in data analysis, ensuring data quality becomes crucial for accurate decision-making. Research in this area focuses on the impact of data quality on AI performance and explores strategies to maintain high-quality data for AI applications [20].
- Healthcare: DQA in healthcare settings has gained significant attention due to the critical role of accurate and reliable data in patient care, clinical decision-making, and healthcare outcomes [21].
- Internet of things (IoTs) is a study area that provides limitless prospects and is vital to researchers worldwide. It is a technological revolution that will alter the way people work, think, live, and it is strongly reliant on the accuracy of data generated by IoT devices [22].
- Web: the vast amount of data that is now available on the web has increased user acceptance of web technologies in recent years, it become the main source of information. Nevertheless, even with a large amount of data, its quality remains doubtful [23]. As a result, investigations on data quality have been carried out to evaluate and improve the quality level of provided data.
- Industry: in the industry, many cases exist where data quality can degrade throughout the life of the production system. Beyond aging sensors, it is the entire data collection infrastructure that can introduce disruptions. These data collection infrastructures can perform poorly, specifically in terms of network parameters such as losses, delays, or traffic load resulting in a decrease in data quality [20].
- Information systems: assessing data quality within a complex information system can be challenging due to the fact that multiple data sources, both hardware and software are involved. These challenges include analyzing and processing a large collection of data in order to ensure data quality [24].
- Social media: the social media has emerged as a new source of helpful information, as data from social media is considered interesting. This is because if processed correctly, it can help to gain insight in business decision making, so that assessing the quality of social data is a context-dependent task [25].
- General context: besides these 8 research domains, there are also many other studies that focus on the issue of data quality in a general context without necessarily being domain specific. These studies are further detailed in the research questions.

Table 5 provides a comprehensive overview of the classification of selected studies based on their respective research domains. This classification enables a better understanding of the distribution and focus of research efforts in different domains related to data quality. The classification in this table serves as a valuable resource for researchers seeking to gain insights into the current landscape of data quality research across various domains.

Table 5. Data quality research domains

DQ domains	Studies	Number
Big data	[26]-[40]	15
IoT	[22], [41]-[53]	14
Information systems	[24], [54]-[66]	14
Web	[67]-[77]	11
Social media	[25], [78]-[81]	5
Healthcare	[21], [82]-[91]	11
Industry	[92]-[96]	5
AI	[97]-[101]	5
General context	[102]-[122]	21

4.3.2. RQ2: what data quality models are crucial for assessing data quality?

After selecting the research domains related to the quality of data, we now focus on the review of proposed data quality models to collect dimensions used by researchers and classify each model used with its specific domain.

- Big data: in an attempt to better understand the process of the national standard reference data (SRD) program, which enables the use of beneficial metrological concepts and methodologies to create trustworthy data and transform such data into national standards, Lee [30] presented the notion of data traceability with three dimensions known as the DQA matrix, which is based on the elements of a data production system and related evaluation criteria. According to Zhang *et al.* [32], recommendation and prediction systems should be used as examples for analyzing the quality of big data and identifying its problems at each level of the processing. Then, once these issues have been analyzed, the corresponding solution for each problem is given from the perspective of data quality. Taleb *et al.* [27] created a big data quality profiling model (BDQPM) with four dimensions: accuracy, completeness, consistency, and timeliness. They emphasized the ideas of data profiling, data quality profiling, and the model contains multiple modules to inspect the quality of data by offering a set of actions to be implemented in the pre-processing phase in particular that is related to the evaluation of data quality. Similarly, Cappiello *et al.* [26] in their data quality service (DQS) module used a model of seven dimensions; accuracy, completeness, consistency, distinctness, precision, timeliness, and volume. Table 6 presents all the quality dimensions used by authors.
- Healthcare: about medical data, a standard model for assessing data quality in primary health care electronic medical records (EMRs) proposed by Terry *et al.* [83], which contains three process of conceptualizing, developing, and testing. 11 metrics for assessing the quality of EMR data used in primary healthcare have been established and tested across three EMR datasets in the domains of comparability, completeness, accuracy, and currency. Likewise, Zan and Zhang [88] provided a data quality evaluation model with five dimensions; accuracy, consistency, integrity, timeliness, and normative to identify data affecting the credibility. A specific process is made to analyze the trust of data source and eliminate those that are untrustworthy. Finally, valid data are evaluated by calculating all dimensions of data quality. Similarly, quality model is for assessing the electronic health record (EHR) data quality in medical informatics in research and care in university medicine (MIRACUM) using standard EHR data quality metrics including plausibility, completeness, and conformance [89]. MIRACUM is a partnership of ten German university hospitals and business partners that aims to address the issues associated with digitization and future medical research. It is a member of the German medical informatics initiative (MII) [123]. Otherwise, Lee *et al.* [85] have implemented an existing DQA framework, using definitions of several data quality dimensions to propose the harmonized framework that focuses on the categories of conformance, completeness, and plausibility. Then, using the DQA framework shown in Table 6, they developed an inventory of common phenotypic data elements (CPDEs) obtained from the study datasets and analyzed it. Stoldt and Weber [91] proposed a quality model that aim assessing the quality of medical data and supporting the clinical decision making. The authors extend the fast healthcare interoperability resources (FHIR) model to enable data provenance annotations to be stored in EHRs. They used the fuzzy logic to determine the level of reliability of the data produced taking into account the level of trust of these data.
- IoT: in the context of the IoT, Zubair *et al.* [42] provided a survey on data quality in IoT when they identified the characteristics of IoT data inherent and specific to the domains. In addition, their classification of IoT data quality are grouped into seven dimensions namely: inaccuracy, completeness, inconsistency, ambiguity, uncertainty, timeliness, and credibility. Furthermore, besides those proposed in the literature related to IoT, other dimensions could be introduced to assess IoT DQ, such as accessibility, access security, and interpretability with two dimensions domains-specific e-health and smart grids such as duplicates and availability [22]. However, Korachi and Bounabat [53] used the DQSC-maturity model for providing the ability to define the maturity level of a smart city based on the quality of the data generated and consumed by the city, as well as for defining relevant recommendations and solutions required to reach the aimed level. Regarding remote sensing, acquisition, and decision-making processes, in [46] many data quality dimensions that are directly relevant to sensor data are described, then they proposed an approach of data quality evaluation for sensor data that supports domain knowledge aggregation and dissemination. However, goal quality model (GQM) is a goal-oriented method of defining software mea-

sures with five dimensions; accuracy, correctness, integrity, unique, and validation. For assessing data quality, this model begins by determining the quality objective of the data instances and then follows the objective until it reaches the problem. Finally, these problems must be able to define the target [97]. In order to describe the case of quality assessment of geographic information systems, Puentes *et al.* [45] suggested a quality model which use remote sensing products. Otherwise, concerning dynamic data quality, Labouseur and Matheus [52] explained the principle of the concept "one size does not fit all" and its relationship with dynamic data, especially dynamic data quality by providing some data quality dimensions to construct their model namely accessibility, ease of manipulation, and representation.

Table 6. Classification of DQ dimensions used

Studies	Categories	Dimensions
[52]		Accessibility, ease of manipulation, and representation
[68], [111]		Completeness, consistency, and accuracy
[79]		Validity, understandability, reference, uncertainty, and balanced/unbiased
[22]		Access security, accessibility, and interpretability
[102]		Completeness, redundancy, accuracy, and data type
[65]		Individual trustworthiness and global conclusiveness
[46]		Believability, accuracy, precision, free-of-error, completeness, timeliness, and consistency
[69]		Data completeness and data format
[58]		Accessibility, accuracy, completeness, consistency, timeliness, and relevance
[66]		Accuracy, completeness, consistency, conformity, integrity, and validity
[26]		Accuracy, completeness, consistency, distinctness, precision, timeliness, and volume
[31]		Accuracy, completeness, consistency, timeliness, validity, and uniqueness
[27]		Accuracy, completeness, consistency, and timeliness
[118]		Completeness, validity, accuracy, and currency
[120]		Uniqueness, validity, accuracy, completeness, and timeliness
[116]		Usefulness, complexity, and compliance
[107]		Accuracy, completeness, consistency, and timeliness
[83]		Comparability, completeness, correctness, and currency
[88]		Accuracy, consistency, integrity, timeliness, and normative
[89]		Plausibility, completeness, and conformance
[42]		Inaccuracy, completeness, inconsistency, ambiguity, uncertainty, timeliness, and credibility
[48]		Accuracy, correctness, integrity, unique, and validation
[91]		Reliability and trust
[79]		Validity, reference, understandability, balanced, and uncertainty
[32]	Data collection	Availability and relevance
	Data preprocessing	Usability and reliability
	Data storage	Usability and availability
	Data analysis	Reliability and usability
[85]	Conformance	Value conformance, relational conformance, and computational conformance
	Completeness	
	Plausibility	Uniqueness plausibility, atemporal plausibility, and temporal plausibility
[30]	Data properties	Completeness
	Method and procedure	Accuracy
	Data value and information	Consistency
[75]	Intrinsic	Syntactic validity, semantic accuracy, consistency, conciseness and completeness
	Accessibility	Availability, licensing, interlinking, security, and performance
	Representational	Representational conciseness, interoperability, interpretability, and versatility
	Contextual	Relevance, trustworthiness, understandability, and timeliness
[76]	Contextual	Relevancy
	Intrinsic	Consistency and accuracy
[47]	Intrinsic	Accuracy and believability
	Accessibility	Accessibility
	Representational	Consistency and interpretability
	Contextual	Timeliness and completeness
[109]	Contextual	Correctness and precision
	Intrinsic	Relevancy, content diversity completeness, and timeliness
[45]	Intrinsic	Source precision, readability, accuracy, resolution, objectivity, integrity, reputation, consistency, obsolescence, uniqueness, freshness, and acquisition cost
	Contextual	Real precision, timeliness, clarity, completeness, trust, concision, value added, volume, and believability
	Extrinsic	Accessibility, manipulation, security, interpretability, ease of use, compatibility, format, understandability, redundancy, and coherence
[78]	Availability	Accessibility, timeliness, and authorization
	Usability	Credibility, definition, metadata
	Reliability	Accuracy, integrity, consistency, auditability, and completeness
	Relevance	Fitness
	Presentation	Readability, structure
[105]	Syntactic level	Completeness, integrity, consistency, validity, maintainability, and timeliness
	Semantic level	Accuracy, coverage
	Pragmatic level	Relevance, usability, risks, currency and decay

– Web: Yi [69] addressed the topic of open data quality by concentrating on data completeness and data format in order to uncover concerns linked to open data. The author compared open data formats used by the governments of three countries: Korea, United Kingdom, and United States. After that, he offered instances of incomplete data across the three nations to demonstrate the presence of the data quality issue, as well as advice for acceptable data formats and data completeness to enhance open data quality. On the other hand, Zaveri and Rula [75] presented a quality rating survey for connected data. Their quality

dimension categorization is divided into four categories: intrinsic, accessible, contextual, and representational. Intrinsic dimensions define data quality from the standpoint of a data provider and independent of external variables. The amount to which data is available and retrievable is defined by the accessibility dimensions. Contextual dimensions are those that are heavily reliant on the task situation. Representation dimensions collect information on the data's design. From the same classification Luzzu's model [74] is constructed, accessibility, intrinsic, contextual, and representational categories with different dimensions for each category. Nevertheless, regarding the issues of Arabic DBpedia, [76] focused only on two categories (intrinsic and contextual) and creating a comparison of current quality assessment tools that demonstrates different attributes/functionalities of the available tools in order to offer a new linked DQA service that assists developers using Arabic language to swiftly rectify problems in DBpedia before utilizing it in a linked data application.

- Information systems: in data warehouse systems, Singh and Kawaljeet [66] presented a quality model of six dimensions, as described in Table 6, associated with four data quality problems, then classified a list of the causes of data quality issues at various stages of data warehousing to help users take care of these issues. However, Oliveira *et al.* [58] introduces an expanded version of the mapping between data mining challenges and data quality dimensions, with three procedures aiming to improve data quality and detect data anomalies using in their model six dimensions; completeness, accuracy, accessibility, consistency, timeliness, and relevance.
- Social media: recently, Salvatore *et al.* [78] addressed the issue of social media data quality, specifically using Twitter as a reference platform. They described new quality factors as reliability, usability, availability, relevance, and presentation quality. Likewise, Zengin and Onder [79] proposed a method for evaluating the quality of Youtube data concerning a specific issue which is the side effect of biologic therapy. The model used to asses the reliability and the quality of videos is presented in Table 6.
- General context: based on the idea of automating the verification of data quality, Schelter *et al.* [113] presented a declarative application programming interface (API) that enables users to recognize constraints on their datasets focusing on completeness, consistency, and accuracy. They explained how these constraints translate into computations of metrics on the data to effectively evaluate the constraints. Similarly, Jungbluth *et al.* [120] has chosen to use in their quality model five dimensions; uniqueness, validity, accuracy, completeness, and timeliness. Research by Berghe and Gaeveren [116] about the issue of data quality regarding the migration of data from an old to a new system, they ranked cleaning tasks according to several criteria, such as usefulness and complexity, while taking compliance into consideration. Furthermore, Ceravolo and Bellini [105] described a general methodology for DQA structured around the notion of matching, which aims at providing a configurable model supporting task composition. Their classification of three levels are composed as illustrated in Table 6. To give an illustration, three case studies are performed, the first one by [118], which performed a case study to analyze the quality of higher education data in one of Indonesia's institutions, according to the pangkalan data pendidikan tinggi (PDDikti is under the jurisdiction of the Ministry of Research, Technology, and Higher Education), such as records of personal data of students, teachers, achievements, and outcomes of professors' assessments. According to the results of this study, the dimensions of quality of higher education data specified by the ministry of research are completeness, validity, accuracy, and currency. The second one is by [107], they identified and examined the quality of data generated by a security incident response team in an organization. Then, following the collect of data, both the analysis of these data and the conclusions of the interviews conducted with the organization's security incident response team are presented according to the accuracy, consistency, completeness, and the timeliness of the data made available to the team. Finally, case report's elaborated at Manitoba Centre for Health Policy (MCHP) describing five key dimensions of data quality framework (DQF) which include accuracy, internal validity, external validity, timeliness, and interpretability [119]. This case report is intended to guide and provide best practices resource for other research institutes dealing with administrative data and trying to enhance their data quality evaluation process.

The literature on data quality outlines thorough definitions of the dimensions of data quality, however there is no agreement on which dimensions characterize data quality or what exactly each dimension means because it is contextual and depends on the perspective of each author [124]. Table 6 highlights the classifications of DQ dimensions employed in different models in the studies described.

4.3.3. RQ3: which methodologies were utilized to assess data quality?

The literature includes a broad variety of methods for assessing and improving data quality. Because of the diversity and complexity of these methodologies, recent research has focused on developing methodologies that aid in the selection and application of DQA and improvement techniques.

- Big data: large volumes of data are generated daily from heterogeneous sources, hence its quality is a major concern. The conflict between these large volume of data and the level of uncertainty in the quality of this data is always current. In this respect, Baldassarre *et al.* [35] presented a methodology called data quality to smart data (DQ2SD) that consists of four phases: data quality planning, analytics rules definition, data quality control, and data quality enhancement, which aims at extracting all of the value within the data, so that ensuring that collected data is both correct and appropriate. In other words, rather than getting a huge amount of meaningless and untruthful data, we will get valuable data. Besides, Klas *et al.* [34] introduced a new scalable quality assessment approach for big data (SQA4BD) with the goal of decoupling data quality analysis from the final individual quality evaluation, which is based on the particular quality demands of the future data consumer. As a result, starting with a basic model for data quality, then analyzing the data, and ultimately assessing quality in an inter-organizational situation is recommended. In such context, three major roles have been defined; data provider who is able to share specific data based on conditions to be negotiated, data consumer who can utilize specific data provided by the supplier based on whether these data contribute to satisfying his or her information and quality needs, and an authority determining the relevant aspects of quality and how they are calculated for various types of data. Furthermore, addressing the issue of big data related to data quality and data diagnosticity, Ghasemaghaei and Calic [37] employed [1]'s DQF to identify several data quality categories used in their study and they used organizational learning theory to characterize the influence of big data utilization on data quality categories as well as decision quality. Likewise, Cappiello *et al.* [26] have created a DQS module, seeking to evaluate the quality of a data source using a set of dimensions including the 4 Vs features (variety, volume, velocity, and veracity) of big data to get many insights about the quality of the examined big data sources.
- Healthcare: from a medical standpoint, more and more medical institutions are emphasizing the need of having an automated framework to maintain the quality of their data effectively. As a result, Pezoulas *et al.* [82] presented a web-based framework for medical DQA that focuses on improving clinical data completeness, relevance, and accuracy by providing a set of quantitative features for metadata extraction, data quality control, and data normalization. Similarly, for EHR data, the framework the care pathway-data quality framework (CP-DQF) suggested by [84], allow systematic management of data quality to assist more trustworthy process mining efforts in EHR research. Likewise, Kapsner *et al.* [89] presented a framework that supports a standardized, unified, and harmonized assessment of EHR data quality in MIRACUM utilizing common EHR data quality dimensions. In addition, this framework helps to systematically identify data quality issues by individual hospitals and then initiate reporting loops to improve its quality. Otherwise, Zaccaria *et al.* [86] proposed a methodology of five steps that consists of improving data quality of a large dataset extracted from a multicenter clinical trial, based on data preprocessing. Each step aims to solve one of the problems observed during the first step and after each step the improvement of the quality of data is evaluated. Within the same background, a scalable framework is recommended by [87] for organizing data quality rules, sharing them, and ensuring their reuse across health care facilities as rule templates, so that errors and discrepancies in data can be identified. Besides, Sun *et al.* [21] offered a provenance-based method for assessing data quality. A qualitative analysis is used to examine the current state of health data. Next, a model of instantiation provenance is included for health data analysis. They provided a framework for analyzing health data based on this model and used a prototype to verify the efficiency of the proposed method.
- IoT: regarding IoT field, there are a lot of methodologies and frameworks aiming at both assessing and improving the quality of data. For this reason, Aquino *et al.* [43] proposed hygieia framework for measuring data quality in IoT context for constrained smart sensor networks (SSNs) devices and helping at imposing low memory overhead and communication overhead in SSN devices. Generally, it is used to analyze the quality of IoT data in order to deliver information to IoT applications. On the other hand, there is valid. IoT, a framework whose architecture is partitioned into three packages, each one has a role to determines the quality of data sources and generate the quality of information vectors to mark-up the sensor information based on quality of data metrics [41]. Likewise, Luo *et al.* [44] proposed a cross

validation strategy to solve various data quality issues and give a fresh viewpoint on improving IoT data quality by fully using the power of crowds. Otherwise, Ge *et al.* [47] offered a framework for data quality management that was supposed to be particular to the area of smart grids. The authors created a list of data quality issues and attempted to utilize it to choose the many associated data quality dimensions in order to determine which ones are critical in the smart grid data quality improvement. Next they identify seven essential data quality characteristics, each of which is linked to specific smart grid data quality issues.

- Web: in order to measure the quality of linked data, Mihindukulasooriya *et al.* [77] proposed a solution based on Loupe API, a RESTful service configurable for profiling linked data by specifying user requirements and other configuration details. Profiling findings may be used to evaluate and validate the quality of a dataset in order to contribute to the quality improvement process by cleaning and fixing deficiencies in the data. Additionally, Ahmed [71] were interested in the issue of data integration in the context of linked open data (LOD), so they presented the problem of DQA during the integration process, then they presented a methodology for evaluating the quality of data in LOD sources during the integration process's phases. Recently, To *et al.* [73] introduced SydNet, a new framework for assessing the quality of linked data. This framework provides an approach which can help network analysts define the dimensions and metrics of quality that are necessary to provide an accurate consideration of the quality of data sources network and help also to merge data from various network data sources. Looking to increase the quality of data, a framework was published by [74] to enhance the quality of linked data. Luzzu with an extensive library of applied quality measures, as well as including a declarative language to create further domain-specific quality measures, as well as a full collection of ontologies for gathering and distributing data quality information. Based on a real case study of Colombian open data government, Sanabria *et al.* [68] proposed a methodological process that help to assess the quality of Sabaneta City's open government data (OGD). This process consists of three stages, beginning with the selection of the data set to be assessed and data profiling. The findings and analyses of the data quality evaluation are defined based on this data profiling, and the tree dimensions of correctness, consistency, and completeness were reviewed, and faults were detected.
- Information systems: Azeroual *et al.* [54] presented the different measures and techniques of data cleaning used both to enhance and increase data quality in research information systems (RIS). For that, knowing the reasons of poor data quality throughout data collection, data transmission, and data integration is critical in order to analyze and then remedy by data transformation and data cleaning. Focusing on the conceptual framework of [65], it is considered as an analytic tool that aims at helping users to understand and distinguish different concepts such as how quality issues could be presented and how potential data quality issues could be classified. This study take in consideration two dimensions of quality, global conclusiveness (GC) and individual trustworthiness (IT). Besides, Timmerman and Bronselaer [55] which suggested a rule-based framework, designed to identify and then address any issues with data quality brought on by improper data collecting and validation methods as well as poor execution.
- Social media: Berlanga *et al.* [25] proposed a methodology with the goal of identifying a reliable metric to judge and assess the general quality of a group of posts and user profiles from odd perspectives, as well as to include the metrics obtained from various quality criteria used to filter the relevance of posts.
- AI: Arbesser *et al.* [98] described and discussed Visplause, a system that aims at inspecting data quality problems in numerous time series by using time series meta-information and plausibility checks to flexibly structure and resume the results of data quality checks. He *et al.* [97] aimed to investigate the link between data quality and model quality, describing four elements of data quality that may be encountered in the field of deep learning. The results then reveal that all four criteria of data quality have a considerable influence on the quality of deep neural network (DNN) models.
- General context: in addition to the above-mentioned methodologies for measuring data quality, there are other interesting ones that are not specifically focused on one area. For instance, the data quality validation methodology (DQVM) has as its objective the assessment of the effects of bad data quality as well as the analysis of associated data quality actions on the results of processes, particularly scoring processes that produce as their output an evaluation that is a ranking or rating for an object. This methodology suggests a series of stages to examine the consequences of faults, injecting faults methodically throughout the process to identify various abnormal circumstances [104]. Furthermore, the total meteorological

data quality (TMDQ) framework, based on the total quality management (TQM) approach developed by [115], aims to offer observers with diverse meteorological data qualities from numerous perspectives according to four quality dimensions, accuracy, consistency, completeness, and timeliness. At the basis of this framework, a validation system is developed to assist meteorological observers in more efficiently improving and maintaining the quality of meteorological data. Similarly, Li *et al.* [109] developed a task-oriented data quality assessment (TODQA) framework which evaluates data quality from two aspects, intrinsic, and contextual quality. It defines, quantify, and fuses assessment metrics to rank candidate data sets based on their quality for specific tasks. While Simard *et al.* [114] have proposed a broad framework that focuses on the measurement and categorization of data and tries to offer a measure of the amount of uncertainty based on four major characteristics of accuracy, completeness, consistency, and timeliness. On the other hand, Kara *et al.* [108] presented a new approach which propose a global data quality model in order to obtain a general method of manipulating and evaluating different factors of data quality. This model combines many data quality models that are linked by five types of equivalence relations to indicate the relationship between two criteria of different models, this relations may be D-Similar, S-Similar, D-Same, S-Same, or S-different. Each one of this models has a tree structure and comprises of factors, criteria, and sub-criteria. Jungbluth *et al.* [120] suggested a quality data extraction methodology to assist users in the process of extracting data in order to increase the quality of data acquired throughout the process. This approach is divided into four primary stages, each with its own set of actions to be followed as a guide or reference. Finally, the methodology DMN4DQ used in [121] tries to simplify the description of data quality, which is separated into two levels: measurement and assessment. The final evaluation is then calculated by adding the scores from each dimension.

4.3.4. RQ4: what are the quality metrics used in DQA?

Having discussed the different methodologies and frameworks proposed for assessing data quality and gathered data quality models suggested by researchers, we now proceed to present the different assessment metrics used to measure these dimensions.

– Big data: as the goal of [28] is to evaluate data value in terms of data quality, they employed three data quality dimensions: accuracy, completeness, and redundancy as shown in Table 7, next based on these dimensions a linear model is established to calculate the quality scores. In addition, Liu *et al.* [36] proposed an approximate quality assessment model based on data set sampling to evaluate the quality of big data. Utilizing various sample sizes and sampling techniques, the authors chose three dimensions; completeness, accuracy, and timeliness to evaluate each sample. In Table 8, the three metrics provided, where S is a collection of data units, S_{acc} is the subset of accurate data units in S, S_{cp} is the subset of complete data units in S, N is the cardinality of S, and S_{acc} and S_{cp}'s combined cardinality is M.

Table 7. Quality metrics used in [28]

Accuracy	Completeness	Redundancy
$1 - \frac{\text{Nb of data with errors}}{\text{Nb total of data}}$	$1 - \frac{\text{Nb of incomplete data}}{\text{Nb total of data}}$	$1 - \frac{\text{Nb of duplicate data}}{\text{Nb total of data}}$

Table 8. Quality metrics used in [36]

Accuracy	Completeness	Timeliness
$Deg_{acc} = \sum_{i=1}^{M} \frac{1}{N} = \frac{M}{N}$	$Deg_{cp} = \sum_{i=1}^{M} \frac{1}{N} = \frac{M}{N}$	$Deg_{tim} = \frac{1}{N} \sum_{i=1}^{N} Deg_{tim}(a_i)$

Taleb *et al.* [33] suggested a big data quality evaluation system in their study that attempts to improve data quality by estimating and assessing data before beginning analysis. As a result, they employed a model containing the most frequently utilized quality dimensions in the context of big data; accuracy, completeness, and consistency with the metrics shown Table 9. Likewise, Mylavarapu *et al.* [29] developed a data accuracy assessment model without needing domain knowledge to evaluate the accuracy of both intrinsic and contextual data. Based on machine learning techniques, they selected the best data from a collection of data sets and considered it as the correct one. The formula used in this study is as follow, where ACC_{IA} represents the new dataset's intrinsic data accuracy (N), M, and K indicate the number of records and variables in the dataset, respectively, l_{ij} and d_{ij} denote the data item of the i^{th} record and the j^{th} variable of the new and correct datasets.

$$ACC_{IA} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{K}(1-\frac{|l_{ij}-d_{ij}|}{max(l_{ij},d_{ij})})}{M*K}$$

Table 9. Quality metrics used in [33]

Accuracy	Completeness	Consistency
$\frac{\text{Nb of correct values}}{\text{Total Nb of values of sample data}}$	$\frac{\text{Nb of missing values}}{\text{Total Nb of values of sample data}}$	$\frac{\text{Nb of values that respects the constraints}}{\text{Total Nb of values of sample data}}$

– IoT: the metrics utilized by [41] in Valid.IoT model are formalized in Table 10.

$$Artificiality = \begin{cases} 1, & \text{If the sensor data is derived from a single IoT hardware sensor and is not aggregated or interpolated} \\ 0, & \text{An unnamed data source that aggregates data using unidentified } algorithms \end{cases} \quad (1)$$

With $C_o(x_0, x_i)$ the individual concordance and a distance function based on infrastructure and propagation between sensor locations a and b for sensors I and j is called $d(a_i, b_j)$.

Table 10. Quality metrics used in [41]

Completeness	Timeliness	Concordance
$\frac{1-\text{Nb of missing data}}{\text{Nb of expected data}}$	Current-timestamp − Message-timestamp	$\frac{\sum_{i=1}^{n}\frac{C_o(x_0,x_i)}{d(x_0,x_i)}}{\sum_{i=1}^{n}\frac{1}{d(x_0,x_i)}}$

– Web: according to the study conducted on three Chinese local government datasets in Beijing, Guangzhou, and Harbin, there are sixteen types of quality problems (Pi, 1¡i¡16) that could affect data availability such as; misalignment, data are too coarse and too granular, text is jumbled, missing values, and date formats vary. For this reason, Li *et al.* [67] classified seven quality characteristics and metrics at various levels and then associate each dimension with the appropriate quality problems to score these three datasets. The results of this evaluation show that the total score for completeness, correctness, and consistency is poor, which will lead consumers to make the incorrect conclusion. In what follow, Table 11 presents the classification of quality evaluation metrics.

Table 11. Quality metrics used in [67]

Dimensions	Formulas
Completeness	$Com = \frac{1-\text{Nb of incomplete data}}{\text{Nb total of data}}$
Accuracy	Acc = $\begin{cases} 1, & \text{no quality problems} \\ 0, & \text{at least one quality problems (P3-P7)} \end{cases}$
Consistency	Con = $\begin{cases} 1, & \text{no quality problems} \\ 0, & \text{at least one quality problems (P8-P10)} \end{cases}$
Timeliness	Tim = $\begin{cases} 1, & \text{no quality problems} \\ 0, & \text{at least one quality problems P12} \end{cases}$
Uniqueness	Uni = $\begin{cases} 1, & \text{no quality problems} \\ 0, & \text{at least one quality problems P13} \end{cases}$
Understandability	Und = $\begin{cases} 1, & \text{no quality problems} \\ 0, & \text{at least one quality problems (P1-P2-P14)} \end{cases}$
Openness	Open = $\begin{cases} 1, & \text{no quality problems} \\ 0, & \text{at least one quality problems (P15-P16)} \end{cases}$

– Industry: the primary contribution of [93] is the development of a general model for objectively assessing the quality of industrial signal data. This model can give a beneficial decision foundation in data mining procedures about the effective and efficient utilization of accessible data. Additionally, Guo *et al.* [92] gave a theoretical study of data quality, they defined and supplied various techniques and technologies

to enhance data quality. Ultimately, a data quality evaluation model is given and a model calculation approach is used to compute the outcome of each dataset in this model.

$$D = \frac{\sum_{i=1}^{n} W_i * L_i}{\sum_{i=1}^{n} W_i}$$

$$R = \frac{\sum_{i=1}^{n} W_i * E_i}{\sum_{i=1}^{n} W_i}$$

With D representing the real data quality of the data set T, R representing the difference between D and the expected value, RT representing the rule set of data set T, W_i representing the weight of rule R_i in RT, Ei representing the expected value, and L_i representing the result of computation.

– General context: Simard *et al.* [114] proposed a broad framework that focuses on the measurement and categorization of data and tries to offer a measure of the amount of uncertainty, concentrating on the four primary characteristics of accuracy, completeness, consistency, and timeliness. Table 12 include the defined quality evaluation measures. According to Aljumaili *et al.* [102], model for the assessment of data quality which takes in consideration the models of [125], [126] to present a model based on metadata and content analysis. In addition, a software tool is developed to validate the proposed measures and to provide an overview of metadata quality. Table 13 presents the quality metrics chosen by [102].

Table 12. Quality metrics used in [114]

Accuracy	Completeness	Timeliness	Consistency
$1 - \lvert\frac{\frac{RV-MV}{RV+MV}}{2}\rvert$	$1 - \frac{\text{Nb of incomplete value}}{\text{Total Nb of value}}$	$1 - \lvert\text{Accuracy1 - Accuracy2}\rvert$	$1 - \frac{\text{Nb of inconsistent value}}{\text{Total Nb of Value}}$

Table 13. Quality metrics used in [102]

Accuracy	Completeness	Redundancy	Data type
$1 - \frac{\text{Nb of data in error}}{\text{Total Nb of data}}$	$1 - \frac{\text{Nb of incomplete data}}{\text{Total Nb of data}}$	$1 - \frac{\text{Nb of redundant data}}{\text{Total Nb of data}}$	$1 - \frac{\text{Nb of data violationg datatype}}{\text{Total Nb of data}}$

It remains almost the same dimensions used for the meteorological data, however, the specific metrics and criteria used to evaluate the quality may vary based on the unique characteristics and requirements of meteorological data. Tsai and Chan [115] proposed the metrics shown in Table 14. Finally, Liu *et al.* [110] provided a summary on the current situation of research on data quality, they analyzed most pertinent characteristics of quality and defined evaluation criteria based on user-defined requirements. The established DQA model includes the major dimensions of data quality, quality characteristics, and quality indicators. Finally, a dynamic data quality evaluation process is built on the basis of the model shown in Table 15.

Table 14. Quality metrics used in [115]

Accuracy	Completeness	Timeliness
$1 - \frac{\text{Nb of failed range checks}}{\text{Nb of Total Data Tested}}$	$1 - \frac{\text{Nb of missing data}}{\text{Nb of Total Data Tested}}$	$1 - \frac{\text{Nb of data not corrected}}{\text{Nb of Total Data Tested}}$

5. DISCUSSION

In this section, we provide a comprehensive overview and discussion of the findings and results obtained in response to the research questions posed in RQ1, RQ2, RQ3, and RQ4. Through mapping these findings, we aim to illustrate the current landscape of data quality research and highlighting key insights and trends.

5.1. Discussion of findings

Since the context is essential to determine which data is relevant to the user, as the data must be tailored to the user's environment and the environment determines the context, RQ1 aims to discover the application

Table 15. Assessment metrics in [110]

Dimensions	Indicators	Metrics
Completeness	Attribute completeness	All data that match the criteria/all data that participated in the evaluation of this indicator
	Record completeness	Nb of records meeting all the criteria/Nb of records participating in the evaluation of this indicator
Accuracy	Range accuracy	All data that match the criteria/all data that participated in the evaluation of this indicator
Consistency	Reference consistency	Nb of data of all eligible rows/Nb of all rows participating in the evaluation
	Format consistency	Columns eligible/all columns involved
Confidentially	Source confidentially	Average points based on survey results
Recoverability	Periodic backup	Average points based on survey results
Traceability	Access traceability	Average points based on survey results
	Value traceability	Average points based on survey results
Understandability	Data understandability	Average points based on survey results
	Data model understandability	Average points based on survey results

domains of data quality. By examining the distribution across domains, researchers and practitioners can identify areas that have received more attention and those that may require further exploration and investigation. In this research question, we have observed that certain research domains related to data quality have received more attention than others. Specifically, the domains of big data, IoT, and information systems have been extensively studied, with a significant number of papers dedicated to exploring the quality aspects in these areas. The prominence of these domains highlights their importance and the need for effective data quality management in these contexts. Additionally, the web has emerged as a significant domain of interest, reflecting its growing role as a major source of information. Healthcare data has also garnered attention, emphasizing the crucial role of data quality in improving healthcare systems. Furthermore, social media, AI, and industry have also been subjects of research, albeit to a lesser extent. The overall distribution of publications across these domains is depicted in Figure 4, providing a visual representation of the research focus in different application areas.

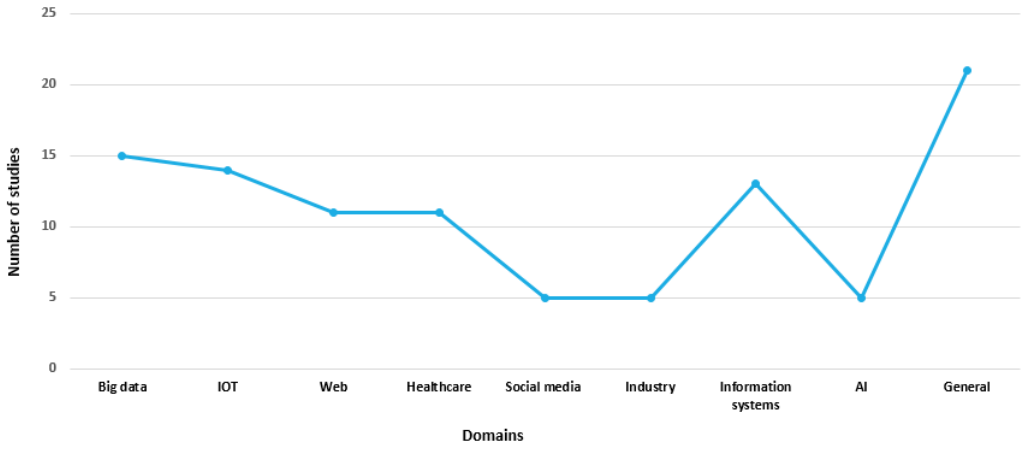


Figure 4. The overall distribution of publications across research domains

One of our objectives in this systematic review was to identify quality models to determine data quality dimensions commonly employed by researchers. Each of these data quality dimensions may establish quality metrics for evaluating data quality. As a result, the majority of research between 2016 and 2021 concentrated on the data quality model. 39 publications investigated the link between data quality dimensions, detailed current data quality management difficulties, and assessed existing data quality methodologies. As a result, RQ2 has over 54 quality dimensions. The most often utilized dimensions to measure data quality in the studies reviewed are completeness, correctness, consistency, and timeliness. Although there are some differences in their definitions due to the contextual nature of quality, they are universal for any data quality evaluation, regardless

of its significance. In order to give readers a better understanding, the plot of the percentage utilization of the dimensions employed in each research is shown in Figure 5. The most common data quality dimensions, as can be seen, are completeness, correctness, consistency, and timeliness.

Research on data quality has covered wide topics of discussion in various domains, as shown in RQ1. In addition, a number of methodologies and frameworks have been disclosed by the authors for an improved evaluation strategy and to overcome data quality issues. For this reason, RQ3 reviewed the proposed methodologies and frameworks to provide a systematic and comparative description of existing data quality methodologies. Analyzing the proposed methodologies revealed a rising need for new approaches to assessing data quality that are more effective and scalable.

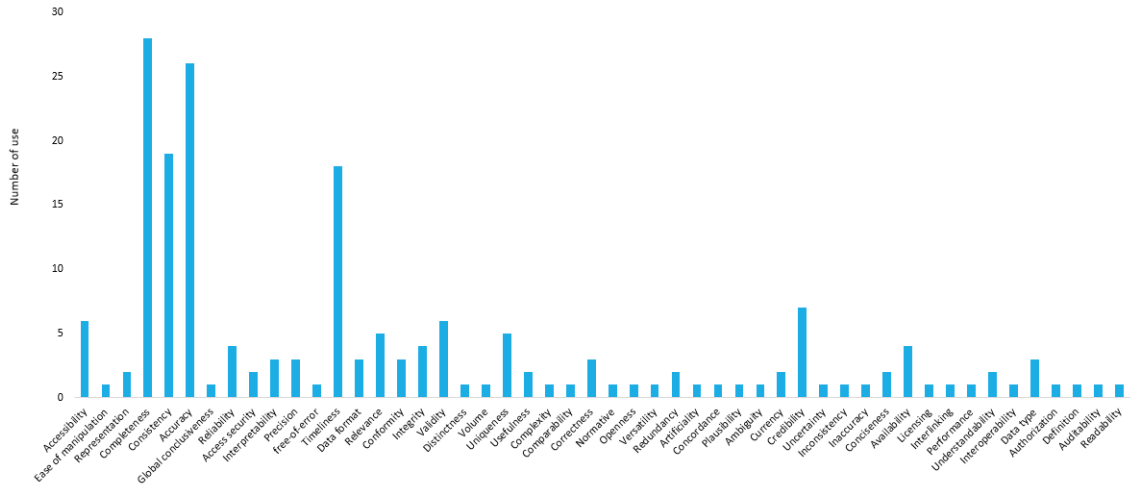


Figure 5. Representation of our findings by percentage of use of quality dimensions

Data quality metrics are used to analyze data quality by measuring accuracy, completeness, consistency, timeliness, and other aspects of data quality. There are two types of metrics: qualitative metrics and quantitative metrics. Quantitative metrics are those that can be quantified or assigned a numerical value. Qualitative measurements cannot be defined and are based on user impression. Finally, RQ4 revealed the evaluation metrics most used. We can conclude from the studies of existing metrics that there is no single formula to measure quality. It is dependent on the context and use of this facts. Indeed, only 12 research publications were identified linked to data quality measures out of the 100 final studies reviewed and there are 40 metrics offered for all 54 aspects.

6. CONCLUSION

Existing researches that focus on domain-specific requirements in relation to DQA are quite considerable, and work done in relation to data dimensionality is also widespread. A thorough literature study was used in this research to clarify the landscape of data quality evaluation. We supplemented our evaluation with 100 research publications on data quality published during 2016 and 2021. This study's findings reveal a substantial trend in data quality research publishing. Each year, the number of papers on data quality grows dramatically, this demonstrates the significance of data quality research across a variety of study disciplines, including online users, databases, web information, sensors, and big data. As a result, this study will not only help academics and practitioners, but it will also give support and insight for future research on data quality evaluation. Then, we considered that our objective was met and that we answered each of the research questions. As future work, we intend to take advantage of this SLR to contribute to the implementation of a new data quality model including all relevant quality dimensions as well as their metrics needed to perform an effective DQA.

REFERENCES

[1] R. Y. Wang and D. M. Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems*, vol. 12, no. 4, pp. 5–33, 1996, doi: 10.1080/07421222.1996.11518099.

[2] V. Siegert, "Content-and Context-Related Trust in Open Multi-agent Systems Using Linked Data," in *Web Engineering*, Cham: Springer, 2019, pp. 541–547, doi: 10.1007/978-3-030-19274-7_42.

[3] M. S. Marev, E. Compatangelo, and W. Vasconcelos, "Towards a context-dependent numerical data quality evaluation framework," *Arxiv-Computer Science*, vol. 1, pp. 1–12, 2018.

[4] A. Nikiforova, "Definition and Evaluation of Data Quality: User-Oriented Data Object-Driven Approach to Data Quality Assessment," *Baltic Journal of Modern Computing*, vol. 8, no. 3, pp. 391–432, 2020, doi: 10.22364/bjmc.2020.8.3.02.

[5] J. E. Olson, *Data Quality: The Accuracy Dimension*. San Francisco: Morgan Kaufmann, 2003.

[6] L. Ehrlinger and W. Wöß, "A Novel Data Quality Metric for Minimality," in *Data Quality and Trust in Big Data*, Cham: Springer, 2019, pp. 1–15, doi: 10.1007/978-3-030-19143-6_1.

[7] R. Silvola, J. Harkonen, O. Vilppola, H. K. Vehkapera, and H. Haapasalo, "Data quality assessment and improvement," *International Journal of Business Information Systems*, vol. 22, no. 1, pp. 62–81, 2016, doi: 10.1504/IJBIS.2016.075718.

[8] O. Reda, I. Sassi, A. Zellou, and S. Anter, "Towards a Data Quality Assessment in Big Data," in *Proceedings of the 13th International Conference on Intelligent Systems: Theories and Applications*, 2020, pp. 1–6, doi: 10.1145/3419604.3419803.

[9] O. Reda and A. Zellou, "SMDQM-Social Media Data Quality Assessment Model," in *2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, 2022, pp. 1–7, doi: 10.1109/IRASET52964.2022.9738330.

[10] C. Batini and M. Scannapieca, *Data Quality: Concepts, Methodologies and Techniques*. Heidelberg: Springer, 2006, doi: 10.1007/3-540-33173-5.

[11] M. Jarke, M. Lenzerini, Y. Vassiliou, and P. Vassiliadis, *Fundamentals of Data Warehouses*. Heidelberg: Springer, 2000, doi: 10.1007/978-3-662-04138-3.

[12] F. Naumann, "Quality-Driven Query Answering for Integrated Information Systems," in *Completeness-Driven Query Optimization*, Heidelberg: Springer, 2002, pp. 123–149, doi: 10.1007/3-540-45921-9_8.

[13] C. Bizer and R. Cyganiak, "Quality-driven information filtering using the WIQA policy framework," *Journal of Web Semantics*, vol. 7, no. 1, pp. 1–10, 2009, doi: 10.1016/j.websem.2008.02.005.

[14] G. Weikum, "Towards Guaranteed Quality and Dependability of Information Services," in *Datenbanksysteme in Büro, Technik und Wissenschaft*, Heidelberg: Springer, 1999, pp. 379–409, doi: 10.1007/978-3-642-60119-4_24.

[15] F. Radulovic, N. Mihindukulasooriya, R. G. -Castro, and A. G. -Pérez, "A comprehensive quality model for Linked Data," *Semantic Web*, vol. 9, no. 1, pp. 3–24, 2017, doi: 10.3233/SW-170267.

[16] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer, "Quality assessment for Linked Data: A Survey," *Semantic Web*, vol. 7, no. 1, pp. 63–93, 2015, doi: 10.3233/SW-150175.

[17] B. Kitchenham and P. Brereton, "A systematic review of systematic review process research in software engineering," *Information and Software Technology*, vol. 55, no. 12, pp. 2049–2075, 2013, doi: 10.1016/j.infsof.2013.07.010.

[18] L. Bertossi and F. Rizzolo, "Contexts and Data Quality Assessment," *Arxiv-Computer Science*, vol. 1, pp. 1–36, 2016.

[19] A. Ramasamy and S. Chowdhury, "Big Data Quality Dimensions: A Systematic Literature Review," *Journal of Information Systems and Technology Management*, vol. 17, pp. 1–13, 2020, doi: 10.4301/S1807-1775202017003.

[20] P. L. Benedick, J. Robert, and Y. L. Traon, "A systematic approach for evaluating artificial intelligence models in industrial settings," *Sensors*, vol. 21, no. 18, pp. 1–17, 2021, doi: 10.3390/s21186195.

[21] Y. Sun, T. Lu, and N. Gu, "A method of electronic health data quality assessment: Enabling data provenance," in *2017 IEEE 21st International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 2017, pp. 233–238, doi: 10.1109/CSCWD.2017.8066700.

[22] A. Karkouch, H. Mousannif, H. A. Moatassime, and T. Noel, "Data quality in internet of things: A state-of-the-art survey," *Journal of Network and Computer Applications*, vol. 73, pp. 57–81, 2016, doi: 10.1016/j.jnca.2016.08.002.

[23] M. I. Jaya, F. Sidi, L. S. Affendey, M. A. Jabar, and I. Ishak, "Systematic review of data quality research," *Journal of Theoretical and Applied Information Technology*, vol. 97, no. 21, pp. 3043–3068, 2019, doi: 10.5281/zenodo.5374485.

[24] O. Azeroual and M. Abuosba, "Improving the data quality in the research information systems," *International Journal of Computer Science and Information Security*, vol. 15, no. 11, pp. 82–86, 2017.

[25] R. Berlanga, I. L. -Cruz, and M. J. Aramburu, "Quality Indicators for Social Business Intelligence," in *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 2019, pp. 229–236, doi: 10.1109/SNAMS.2019.8931862.

[26] C. Cappiello, W. Samá, and M. Vitali, "Quality awareness for a Successful Big Data Exploitation," in *Proceedings of the 22nd International Database Engineering & Applications Symposium on-IDEAS 2018*, 2018, pp. 37–44, doi: 10.1145/3216122.3216124.

[27] I. Taleb, M. A. Serhani, and R. Dssouli, "Big Data Quality: A Data Quality Profiling Model," in *SERVICES 2019*, Cham: Springer, 2019, pp. 61–77, doi: 10.1007/978-3-030-23381-5_5.

[28] J. Yang, C. Zhao, and C. Xing, "Big Data Market Optimization Pricing Model Based on Data Quality," *Complexity*, vol. 2019, pp. 1–10, 2019, doi: 10.1155/2019/5964068.

[29] G. Mylavarapu, J. P. Thomas, and K. A. Viswanathan, "An Automated Big Data Accuracy Assessment Tool," in *2019 IEEE 4th International Conference on Big Data Analytics (ICBDA)*, 2019, pp. 193–197, doi: 10.1109/ICBDA.2019.8713218.

[30] D. Lee, "Big Data Quality Assurance Through Data Traceability: A Case Study of the National Standard Reference Data Program of Korea," *IEEE Access*, vol. 7, pp. 36294–36299, 2019, doi: 10.1109/ACCESS.2019.2904286.

[31] G. A. Lakshen, S. Vranes, and V. Janev, "Big data and quality: A literature review," in *2016 24th Telecommunications Forum (TELFOR)*, 2016, pp. 1–4, doi: 10.1109/TELFOR.2016.7818902.

[32] P. Zhang, F. Xiong, J. Gao, and J. Wang, "Data quality in big data processing: Issues, solutions and open problems," in *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, 2017, pp. 1–7, doi: 10.1109/UIC-ATC.2017.8397554.

[33] I. Taleb, H. T. E. Kassabi, M. A. Serhani, R. Dssouli, and C. Bouhaddioui, "Big Data Quality: A Quality Dimensions Evaluation," in *2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World*

Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), 2016, pp. 759–765, doi: 10.1109/UIC-ATC-ScalCom-CBDCom-IoP-SmartWorld.2016.0122.

[34] M. Klas, W. Putz, and T. Lutz, "Quality Evaluation for Big Data: A Scalable Assessment Approach and First Evaluation Results," in *2016 Joint Conference of the International Workshop on Software Measurement and the International Conference on Software Process and Product Measurement (IWSM-MENSURA)*, 2016, pp. 115–124, doi: 10.1109/IWSM-Mensura.2016.026.

[35] M. T. Baldassarre, I. Caballero, D. Caivano, B. R. Garcia, and M. Piattini, "From big data to smart data: a data quality perspective," in *Proceedings of the 1st ACM SIGSOFT International Workshop on Ensemble-Based Software Engineering*, 2018, pp. 19–24, doi: 10.1145/3281022.3281026.

[36] H. Liu, Z. Sang, and S. Karali, "Approximate Quality Assessment with Sampling Approaches," in *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2019, pp. 1306–1311, doi: 10.1109/CSCI49370.2019.00244.

[37] M. Ghasemaghaei and G. Calic, "Can big data improve firm decision quality? The role of data quality and data diagnosticity," *Decision Support Systems*, vol. 120, pp. 38–49, 2019, doi: 10.1016/j.dss.2019.03.008.

[38] F. Auer and M. Felderer, "Addressing Data Quality Problems with Metamorphic Data Relations," in *2019 IEEE/ACM 4th International Workshop on Metamorphic Testing (MET)*, 2019, pp. 76–83, doi: 10.1109/MET.2019.00019.

[39] I. Taleb, M. A. Serhani, C. Bouhaddioui, and R. Dssouli, "Big data quality framework: a holistic approach to continuous quality management," *Journal of Big Data*, vol. 8, no. 1, pp. 1–41, 2021, doi: 10.1186/s40537-021-00468-0.

[40] P. Sinthong, D. Patel, N. Zhou, S. Shrivastava, A. Iyengar, and A. Bhamidipaty, "DQDF: data-quality-aware dataframes," *Proceedings of the VLDB Endowment*, vol. 15, no. 4, pp. 949–957, 2021, doi: 10.14778/3503585.3503602.

[41] D. Kuemper, T. Iggena, R. Toenjes, and E. Pulvermueller, "Valid.IoT: a framework for sensor data quality analysis and interpolation," in *Proceedings of the 9th ACM Multimedia Systems Conference*, 2018, pp. 294–303, doi: 10.1145/3204949.3204972.

[42] N. Zubair, A. Niranjan, K. Hebbar, and Y. Simmhan, "Characterizing IoT Data and its Quality for Use," *Arxiv-Computer Science*, vol. 1, pp. 1–16, 2019.

[43] G. R. C. d. Aquino, C. M. d. Farias, and L. Pirmez, "Hygieia: data quality assessment for smart sensor network," in *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, 2019, pp. 889–891, doi: 10.1145/3297280.3297564.

[44] T. Luo, J. Huang, S. S. Kanhere, J. Zhang, and S. K. Das, "Improving IoT Data Quality in Mobile Crowd Sensing: A Cross Validation Approach," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 5651–5664, 2019, doi: 10.1109/JIOT.2019.2904704.

[45] J. Puentes, L. Lecornu, and B. Solaiman, "Data and Information Quality in Remote Sensing," in *Information Quality in Information Fusion and Decision Making*, Cham: Springer, 2019, pp. 401–421, doi: 10.1007/978-3-030-03643-0_17.

[46] E. Ferreira and D. Ferreira, "Towards altruistic data quality assessment for mobile sensing," in *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, 2017, pp. 464–469, doi: 10.1145/3123024.3124439.

[47] M. Ge, S. Chren, B. Rossi, and T. Pitner, "Data Quality Management Framework for Smart Grid Systems," in *Business Information Systems*, Cham: Springer, 2019, pp. 299–310, doi: 10.1007/978-3-030-20482-2_24.

[48] Z. Li, S. Wu, H. Zhou, S. Zou, and T. Dong, "Analytic Model and Assessment Framework for Data Quality Evaluation in State Grid," *Journal of Physics: Conference Series*, vol. 1302, no. 2, pp. 1–6, 2019, doi: 10.1088/1742-6596/1302/2/022083.

[49] I. Khokhlov and L. Reznik, "Knowledge Graph in Data Quality Evaluation for IoT applications," in *2020 IEEE 6th World Forum on Internet of Things (WF-IoT)*, 2020, pp. 1–6, doi: 10.1109/WF-IoT48130.2020.9221091.

[50] J. Byabazaire, G. O'. Hare, and D. Delaney, "Data Quality and Trust: Review of Challenges and Opportunities for Data Sharing in IoT," *Electronics*, vol. 9, no. 12, pp. 1–22, 2020, doi: 10.3390/electronics9122083.

[51] S. Chren, B. Rossi, B. Buhnova, and T. Pitner, "Reliability data for smart grids: Where the real data can be found," in *2018 Smart City Symposium Prague (SCSP)*, 2018, pp. 1–6, doi: 10.1109/SCSP.2018.8402648.

[52] A. G. Labouseur and C. C. Matheus, "An Introduction to Dynamic Data Quality Challenges," *Journal of Data and Information Quality*, vol. 8, no. 2, pp. 1–3, 2017, doi: 10.1145/2998575.

[53] Z. Korachi and B. Bounabat, "Data Driven Maturity Model for Assessing Smart Cities," in *Proceedings of the 2nd International Conference on Smart Digital Environment*, 2018, pp. 140–147, doi: 10.1145/3289100.3289123.

[54] O. Azeroual, G. Saake, and M. Abuosba, "Data Quality Measures and Data Cleansing for Research Information Systems," *Journal of Digital Information Management*, vol. 16, no. 1, pp. 12–21, 2018.

[55] Y. Timmerman and A. Bronselaer, "Measuring data quality in information systems research," *Decision Support Systems*, vol. 126, 2019, doi: 10.1016/j.dss.2019.113138.

[56] A. G. Labouseur and C. C. Matheus, "Dynamic Data Quality for Static Blockchains," in *2019 IEEE 35th International Conference on Data Engineering Workshops (ICDEW)*, 2019, pp. 19–21, doi: 10.1109/ICDEW.2019.00-41.

[57] H. N. Benkhaled and D. Berrabah, "Data quality management for data warehouse systems: State of the art," *CEUR Workshop Proceedings*, vol. 2351, pp. 1–10, 2019.

[58] A. Oliveira, R. Gaio, P. Baylina, C. Rebelo, and L. P. Reis, "Data Quality Mining," in *New Knowledge in Information Systems and Technologies*, Cham: Springer, 2019, pp. 361–372, doi: 10.1007/978-3-030-16181-1_34.

[59] F. Serra and A. Marotta, "Data quality in data warehouse systems: A context-based approach," in *2016 XLII Latin American Computing Conference (CLEI)*, 2016, pp. 1–12, doi: 10.1109/CLEI.2016.7833371.

[60] N. Micic, D. Neagu, F. Campean, and E. H. Zadeh, "Towards a Data Quality Framework for Heterogeneous Data," in *2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, 2017, pp. 155–162, doi: 10.1109/iThings-GreenCom-CPSCom-SmartData.2017.28.

[61] F. Basciani, J. d. Rocco, D. d. Ruscio, L. Iovino, and A. Pierantonio, "A Customizable Approach for the Automated Quality Assessment of Modelling Artifacts," in *2016 10th International Conference on the Quality of Information and Communications Technology (QUATIC)*, 2016, pp. 88–93, doi: 10.1109/QUATIC.2016.025.

[62] M. Kara, O. Lamouchi, and A. R. -Cherif, "Ontology Software Quality Model for Fuzzy Logic Evaluation Approach," *Procedia Computer Science*, vol. 83, pp. 637–641, 2016, doi: 10.1016/j.procs.2016.04.143.

[63] G. Liebchen and M. Shepperd, "Data Sets and Data Quality in Software Engineering," in *Proceedings of the The 12th International Conference on Predictive Models and Data Analytics in Software Engineering*, 2016, pp. 1–4, doi: 10.1145/2972958.2972967.

[64] M. Souibgui, F. Atigui, S. Zammali, S. Cherfi, and S. B. Yahia, "Data quality in ETL process: A preliminary study," *Procedia Computer Science*, vol. 159, pp. 676–687, 2019, doi: 10.1016/j.procs.2019.09.223.
[65] X. Lu and D. Fahland, "A conceptual framework for understanding event data quality in behavior analysis," *CEUR Workshop Proceedings*, vol. 1826, pp. 11–14, 2017.
[66] R. Singh and S. Kawaljeet, "A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing," *IJCSI International Journal of Computer Science Issues*, vol. 7, no. 2, pp. 41–50, 2010.
[67] X. -T. Li, J. Zhai, G. -F. Zheng, and C. -F. Yuan, "Quality Assessment for Open Government Data in China," in *Proceedings of the 2018 10th International Conference on Information Management and Engineering*, 2018, pp. 110–114, doi: 10.1145/3285957.3285962.
[68] M. A. O. Sanabria, F. O. A. Fernández, and M. P. G. Zabala, "Colombian Case Study for the Analysis of Open Data Government," in *Proceedings of the 11th International Conference on Theory and Practice of Electronic Governance*, 2018, pp. 389–394, doi: 10.1145/3209415.3209474.
[69] M. Yi, "Exploring the quality of government open data," *The Electronic Library*, vol. 37, no. 1, pp. 35–48, 2019, doi: 10.1108/EL-06-2018-0124.
[70] S. Sadiq and M. Indulska, "Open data: Quality over quantity," *International Journal of Information Management*, vol. 37, no. 3, pp. 150–154, 2017, doi: 10.1016/j.ijinfomgt.2017.01.003.
[71] H. H. Ahmed, "Data Quality Assessment in the Integration Process of Linked Open Data (LOD)," in *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, 2017, pp. 1–6, doi: 10.1109/AICCSA.2017.178.
[72] A. Hadhiatma, "Improving data quality in the linked open data: a survey," *Journal of Physics: Conference Series*, vol. 978, pp. 1–7, 2018, doi: 10.1088/1742-6596/978/1/012026.
[73] A. To, R. Meymandpour, J. G. Davis, G. Jourjon, and J. Chan, "A Linked Data Quality Assessment Framework for Network Data," in *Proceedings of the 2nd Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA)*, 2019, pp. 1–8, doi: 10.1145/3327964.3328493.
[74] J. Debattista, S. Auer, and C. Lange, "Luzzu—A Methodology and Framework for Linked Data Quality Assessment," *Journal of Data and Information Quality*, vol. 8, no. 1, pp. 1–32, 2016, doi: 10.1145/2992786.
[75] A. Zaveri and A. Rula, "Data Quality and Data Cleansing of Semantic Data," in *Encyclopedia of Big Data Technologies*, Cham: Springer, 2019, pp. 573–579, doi: 10.1007/978-3-319-77525-8_289.
[76] G. A. Lakshen, V. Janev, and S. Vraneš, "Challenges in Quality Assessment of Arabic DBpedia," in *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*, 2018, pp. 1–4, doi: 10.1145/3227609.3227675.
[77] N. Mihindukulasooriya, R. G. -Castro, F. Priyatna, E. Ruckhaus, and N. Saturno, "A Linked Data Profiling Service for Quality Assessment," in *The Semantic Web: ESWC 2017 Satellite Events*, Cham: Springer, 2017, pp. 335–340, doi: 10.1007/978-3-319-70407-4_42.
[78] C. Salvatore, S. Biffignandi, and A. Bianchi, "Social Media and Twitter Data Quality for New Social Indicators," *Social Indicators Research*, vol. 156, no. 2, pp. 601–630, 2021, doi: 10.1007/s11205-020-02296-w.
[79] O. Zengin and M. E. Onder, "YouTube for information about side effects of biologic therapy: A social media analysis," *International Journal of Rheumatic Diseases*, vol. 23, no. 12, pp. 1645–1650, 2020, doi: 10.1111/1756-185X.14003.
[80] J. A. Qundus, A. Paschke, S. Gupta, A. M. Alzouby, and M. Yousef, "Exploring the impact of short-text complexity and structure on its quality in social media," *Journal of Enterprise Information Management*, vol. 33, no. 6, pp. 1443–1466, 2020, doi: 10.1108/JEIM-06-2019-0156.
[81] E. Anderson, M. Koss, A. L. C. Luque, D. Garcia, E. Lopez, and K. Ernst, "WhatsApp-Based Focus Groups Among Mexican-Origin Women in Zika Risk Area: Feasibility, Acceptability, and Data Quality," *JMIR Formative Research*, vol. 5, no. 10, pp. 1–12, 2021, doi: 10.2196/20970.
[82] V. C. Pezoulas *et al.*, "Medical data quality assessment: On the development of an automated framework for medical data curation," *Computers in Biology and Medicine*, vol. 107, pp. 270–283, 2019, doi: 10.1016/j.compbiomed.2019.03.001.
[83] A. L. Terry *et al.*, "A basic model for assessing primary health care electronic medical record data quality," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, pp. 1–11, 2019, doi: 10.1186/s12911-019-0740-0.
[84] F. Fox, V. R. Aggarwal, H. Whelton, and O. Johnson, "A Data Quality Framework for Process Mining of Electronic Health Record Data," in *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, 2018, pp. 12–21, doi: 10.1109/ICHI.2018.00009.
[85] K. Lee, N. Weiskopf, and J. Pathak, "A Framework for Data Quality Assessment in Clinical Research Datasets," *AMIA-Annual Symposium proceedings*, vol. 2017, pp. 1080–1089, 2017.
[86] G. M. Zaccaria *et al.*, "Data quality improvement of a multicenter clinical trial dataset," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2017, pp. 1190–1193, doi: 10.1109/EMBC.2017.8037043.
[87] Z. Wang, S. Dagtas, J. Talburt, A. Baghal, and M. Zozus, "Rule-based data quality assessment and monitoring system in healthcare facilities," *Studies in Health Technology and Informatics*, vol. 257, pp. 460–467, 2019, doi: 10.3233/978-1-61499-951-5-460.
[88] S. Zan and X. Zhang, "Medical Data Quality Assessment Model Based on Credibility Analysis," in *2018 IEEE 4th Information Technology and Mechatronics Engineering Conference (ITOEC)*, 2018, pp. 940–944, doi: 10.1109/ITOEC.2018.8740576.
[89] L. A. Kapsner *et al.*, "Moving towards an EHR data quality framework: The miracum approach," *Studies in Health Technology and Informatics*, vol. 267, pp. 247–253, 2019, doi: 10.3233/SHTI190834.
[90] Z. Wang, J. R. Talburt, N. Wu, S. Dagtas, and M. N. Zozus, "A Rule-Based Data Quality Assessment System for Electronic Health Record Data," *Applied Clinical Informatics*, vol. 11, no. 4, pp. 622–634, 2020, doi: 10.1055/s-0040-1715567.
[91] J. -P. Stoldt and J. H. Weber, "Provenance-based Trust Model for Assessing Data Quality during Clinical Decision Making," in *2021 IEEE/ACM 3rd International Workshop on Software Engineering for Healthcare (SEH)*, 2021, pp. 24–31, doi: 10.1109/SEH52539.2021.00012.
[92] A. Guo, X. Liu, and T. Sun, "Research on Key Problems of Data Quality in Large Industrial Data Environment," in *Proceedings of the 3rd International Conference on Robotics, Control and Automation*, 2018, pp. 245–248, doi: 10.1145/3265639.3265680.
[93] I. Kirchen, D. Schutz, J. Folmer, and B. V. -Heuser, "Metrics for the evaluation of data quality of signal data in industrial processes," in *2017 IEEE 15th International Conference on Industrial Informatics (INDIN)*, 2017, pp. 819–826, doi: 10.1109/INDIN.2017.8104878.

[94] Q. Xiao, M. Shan, X. Xiao, and C. Rao, "Evaluation Model of Industrial Operation Quality Under Multi-source Heterogeneous Data Information," *International Journal of Fuzzy Systems*, vol. 22, no. 2, pp. 522–547, 2020, doi: 10.1007/s40815-019-00776-x.
[95] D. Król and T. Czarnecki, "Testing for Data Quality Assessment: ACase Study from the Industry 4.0 Perspective," in *Advances in Computational Collective Intelligence*, Cham: Springer, 2021, pp. 73–85, doi: 10.1007/978-3-030-88113-9_6.
[96] L. B. Iantovics and C. Enăchescu, "Method for Data Quality Assessment of Synthetic Industrial Data," *Sensors*, vol. 22, no. 4, pp. 1–21, 2022, doi: 10.3390/s22041608.
[97] T. He, S. Yu, Z. Wang, J. Li, and Z. Chen, "From Data Quality to Model Quality: An Exploratory Study on Deep Learning," in *Proceedings of the 11th Asia-Pacific Symposium on Internetware*, 2019, pp. 1–6, doi: 10.1145/3361242.3361260.
[98] C. Arbesser, F. Spechtenhauser, T. Muhlbacher, and H. Piringer, "Visplause: Visual Data Quality Assessment of Many Time Series Using Plausibility Checks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 641–650, 2017, doi: 10.1109/TVCG.2016.2598592.
[99] L. Bertossi and F. Geerts, "Data Quality and Explainable AI," *Journal of Data and Information Quality*, vol. 12, no. 2, pp. 1–9, 2020, doi: 10.1145/3386687.
[100] S. Shrivastava, D. Patel, N. Zhou, A. Iyengar, and A. Bhamidipaty, "DQLearn: A Toolkit for Structured Data Quality Learning," in *2020 IEEE International Conference on Big Data (Big Data)*, 2020, pp. 1644–1653, doi: 10.1109/BigData50022.2020.9378296.
[101] H. Zhang, S. Wang, and X. Wang, "Rule-based Data Quality Intelligent Monitoring System," *Journal of Physics: Conference Series*, vol. 1670, no. 1, pp. 1–6, 2020, doi: 10.1088/1742-6596/1670/1/012031.
[102] M. Aljumaili, R. Karim, and P. Tretten, "Metadata-based data quality assessment," *VINE Journal of Information and Knowledge Management Systems*, vol. 46, no. 2, pp. 232–250, 2016, doi: 10.1108/VJIKMS-11-2015-0059.
[103] J. Bicevskis, Z. Bicevska, A. Nikiforova, and I. Oditis, "An Approach to Data Quality Evaluation," in *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 2018, pp. 196–201, doi: 10.1109/SNAMS.2018.8554915.
[104] C. Cappiello, C. Cerletti, C. Fratto, and B. Pernici, "Validating Data Quality Actions in Scoring Processes," *Journal of Data and Information Quality*, vol. 9, no. 2, pp. 1–27, 2017, doi: 10.1145/3141248.
[105] P. Ceravolo and E. Bellini, "Towards Configurable Composite Data Quality Assessment," in *2019 IEEE 21st Conference on Business Informatics (CBI)*, 2019, pp. 249–257, doi: 10.1109/CBI.2019.00035.
[106] B. Heinrich, D. Hristova, M. Klier, A. Schiller, and M. Szubartowicz, "Requirements for Data Quality Metrics," *Journal of Data and Information Quality*, vol. 9, no. 2, pp. 1–32, 2017, doi: 10.1145/3148238.
[107] G. Grispos, W. Glisson, and T. Storer, "How Good is Your Data? Investigating the Quality of Data Generated During Security Incident Response Investigations," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019, pp. 7156–7165, doi: 10.24251/HICSS.2019.859.
[108] M. Kara, O. Lamouchi, and A. R. -Cherif, "Semantically equivalent model for quality evaluation," in *Proceedings of the Second International Conference on Internet of things, Data and Cloud Computing*, 2017, pp. 1–5, doi: 10.1145/3018896.3056776.
[109] A. Li, L. Zhang, J. Qian, X. Xiao, X. -Y. Li, and Y. Xie, "TODQA: Efficient Task-Oriented Data Quality Assessment," in *2019 15th International Conference on Mobile Ad-Hoc and Sensor Networks (MSN)*, 2019, pp. 81–88, doi: 10.1109/MSN48538.2019.00028.
[110] Z. Liu, Q. Chen, and L. Cai, "Application of Requirement-oriented Data Quality Evaluation Method," in *2018 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, 2018, pp. 407–412, doi: 10.1109/SNPD.2018.8441103.
[111] S. Schelter, D. Lange, P. Schmidt, M. Celikel, F. Biessmann, and A. Grafberger, "Automating large-scale data quality verification," in *Proceedings of the VLDB Endowment*, 2018, vol. 11, no. 12, pp. 1781–1794, doi: 10.14778/3229863.3229867.
[112] S. Sadiq *et al.*, "Data Quality: The Role of Empiricism," *ACM SIGMOD Record*, vol. 46, no. 4, pp. 35–43, 2018, doi: 10.1145/3186549.3186559.
[113] S. Schelter *et al.*, "Differential Data Quality Verification on Partitioned Data," in *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, 2019, pp. 1940–1945, doi: 10.1109/ICDE.2019.00210.
[114] V. Simard, M. Rönnqvist, L. Lebel, and N. Lehoux, "A General Framework for Data Uncertainty and Quality Classification," *IFAC-PapersOnLine*, vol. 52, no. 13, pp. 277–282, 2019, doi: 10.1016/j.ifacol.2019.11.181.
[115] W. -L. Tsai and Y. -C. Chan, "Designing a Framework for Data Quality Validation of Meteorological Data System," *IEICE Transactions on Information and Systems*, vol. 102, no. 4, pp. 800–809, 2019, doi: 10.1587/transinf.2018DAP0021.
[116] S. V. d. Berghe and K. V. Gaeveren, "Data Quality Assessment and Improvement: A Vrije Universiteit Brussel Case Study," *Procedia Computer Science*, vol. 106, pp. 32–38, 2017, doi: 10.1016/j.procs.2017.03.006.
[117] Y. Yang, Y. Yuan, and B. Li, "Data Quality Evaluation: Methodology and Key Factors," in *Smart Computing and Communication*, Cham: Springer, 2018, pp. 222–230, doi: 10.1007/978-3-319-73830-7_22.
[118] W. Wijayanti, A. N. Hidayanto, N. Wilantika, I. R. Adawati, and S. B. Yudhoatmojo, "Data Quality Assessment on Higher Education: A Case Study of Institute of Statistics," in *2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 2018, pp. 231–236, doi: 10.1109/ISRITI.2018.8864476.
[119] M. Smith *et al.*, "Assessing the quality of administrative data for research: a framework from the Manitoba Centre for Health Policy," *Journal of the American Medical Informatics Association*, vol. 25, no. 3, pp. 224–229, 2018, doi: 10.1093/jamia/ocx078.
[120] A. Jungbluth, J. L. Yeng, and L. Vives, "Quality data extraction methodology based on the labeling of coffee leaves with nutritional deficiencies," in *Proceedings of the 2nd International Conference on Information System and Data Mining*, 2018, pp. 59–64, doi: 10.1145/3206098.3206102.
[121] Á. V. -Parra, L. Parody, Á. J. V. -Vaca, I. Caballero, and M. T. G. -López, "DMN4DQ: When data quality meets DMN," *Decision Support Systems*, vol. 141, 2021, doi: 10.1016/j.dss.2020.113450.
[122] M. Mohammed, J. R. Talburt, S. Dagtas, and M. Hollingsworth, "A Zero Trust Model Based Framework For Data Quality Assessment," in *2021 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2021, pp. 305–307, doi: 10.1109/CSCI54926.2021.00123.
[123] H. -U. Prokosch *et al.*, "MIRACUM: Medical Informatics in Research and Care in University Medicine," *Methods of Information in Medicine*, vol. 57, no. 1, pp. 82–91, 2018, doi: 10.3414/ME17-02-0025.
[124] G. K. Tayi and D. P. Ballou, "Examining data quality," *Communications of the ACM*, vol. 41, no. 2, pp. 54–57, 1998, doi:

10.1145/269012.269021.
[125] Y. W. Lee, L. L. Pipino, J. D. Funk, and R. Y. Wang, *Journey to Data Quality*. Massachusetts: The MIT Press, 2006, doi: 10.7551/mitpress/4037.001.0001.
[126] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–52, 2009, doi: 10.1145/1541880.1541883.

BIOGRAPHIES OF AUTHORS

Oumaima Reda is pursuing Ph.D. in Software Project Management from National School of Computer Science and System Analysis (ENSIAS), Mohammed V University in Rabat, Morocco. She obtained a master degree in internet of things and services mobile (IOSM) from National School of Computer Science and System Analysis (ENSIAS), in 2019. Her research interest includes data science, data quality, and social media. She can be contacted at email: oumaima_reda@um5.ac.ma and redam7687@gmail.com.

Naoual Chaouni Benabdellah is an assistant professor of Computer Science since 2018 at National School of Computer Science and System Analysis (ENSIAS) one of the Enginering Schools at the Mohammed V University of Rabat. She is member Department of Web and Mobile Engineering and the Software Project Management research team. Her research interests are e-learning, education, and data science. She was guest speaker at many events. She participated as a reviewer and as session's president at international conferences. She can be contacted at email: naoual.chaouni_benabdellah@um5.ac.ma.

Ahmed Zellou received his Ph.D. in Applied Sciences at the Mohammedia School of Engineers, Mohammed V University, Rabat, Morocco 2008. His habilitation to supervise research work in 2014. He becomes full professor in 2021. His research interests include interoperability, mediation systems, distributed computing, data quality, and semantic web. He is the author/co-author of over 100 research publications. He can be contacted at email: ahmed.zellou@um5.ac.ma.