Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 234 (2024) 341–348

Procedia Computer Science

www.elsevier.com/locate/procedia

Seventh Information Systems International Conference (ISICO 2023)

A Review on Data Quality Dimensions for Big Data

Fakhitah Ridzuan[a,b]*, Wan Mohd Nazmee Wan Zainon[c]

[a]Faculty of Data Science and Computing, Universiti Malaysia Kelantan, Pengkalan Chepa, 16100 Kota Bharu, Kelantan
[b]Institute for Artificial Intelligence & Big Data, Universiti Malaysia Kelantan, Pengkalan Chepa, 16100 Kota Bharu, Kelantan
[c]School of Computer Sciences, Universiti Sains Malaysia, 11800 Penang.

Abstract

Big Data wave has led to a rapid increase in the amount of data being collected by organizations. While the accuracy and reliability of prediction models are often prioritized, the quality of the collected data is frequently overlooked. Poor data quality can result in the common problem of 'garbage in, garbage out'. Traditional measures of data quality, such as accuracy, consistency, completeness, and timeliness, are no longer adequate in the era of Big Data. Therefore, this paper proposes a taxonomy of data quality dimensions specifically for Big Data, addressing emerging challenges by formulating 20 dimensions and categorizing them into four distinct categories.

Keywords: Big data; Data quality; Data quality dimension; Data management

1. Introduction

Big Data revolution has brought a significant change in the way organizations handle and analyze data, making data a crucial component in decision-making processes. However, poor data quality, resulting from errors, inconsistencies, and inaccuracies, commonly known as "dirty data," is a major threat to data quality, leading to incorrect decisions with significant consequences. Data quality has become a crucial issue in data management as data is available from various sources and in various formats [1].

With the rapid growth of big data, deriving high-quality data from it is critical for making informed decisions, optimizing operations, and delivering superior customer service [2]. Data quality refers to the degree to which data meets the needs and requirements of a business or organization. It is determined by factors such as the nature and scale of the organization, its processes and procedures, and the specific ways in which data is used. Effective data quality management is crucial as even minor errors can lead to revenue loss, inefficiencies in processes, and non-compliance

* Corresponding author. Tel.: +609 - 771 7000
E-mail address: fakhitah.r@umk.edu.my

10.1016/j.procs.2024.03.008

with industry and government regulations [3]. Data quality represents the standard to which the information is fit for usage in the required business processes. It can be defined, measured, and managed through several data quality metrics, such as completeness, conformity, consistency, and accuracy. Data quality is not about how cleansed the data is, but it has become a critical issue because it involves operational processes and is perceived as the greatest challenge in data management [4].

In the context of Big Data, the complexity of the data quality algorithm increases due to the five V's of Big Data (Volume, Velocity, Variety, Value and Veracity) [5]. Additionally, the assessment of data quality in Big Data environment requires a distinct approach, as traditional evaluation models are not suitable [6]. Data quality has gained considerable attention in the Big Data era due to its complexity and the need for well-defined, lightweight measurement processes that can operate in tandem with each phase [7]. A recent report by O'Reilly has shown that data quality may deteriorate before it improves, highlighting the importance of managing data quality issues for organizations to make informed decisions using data analytics[8].

Due to the critical role of data in organizational decision-making processes, it is imperative that the data used is of the highest quality [9]. Therefore, evaluating data quality and ensuring high veracity of data is the utmost importance. To address this issue, a review was conducted with the research question: "What are the best data quality dimensions for Big Data?" The objective was to identify existing data quality dimensions specific to Big Data, as they serve as a foundation for assessing data veracity. The paper is organized as follows; The challenges related to data quality are presented in Section 2, followed by an explanation of the existing data quality dimensions in Section 3. Section 4 presents the results of the study on the importance of accuracy, completeness, consistency, and timeliness as factors for data quality assessment. In Section 5, a taxonomy of data quality dimensions for Big Data is proposed. Lastly, Section 6 provides the conclusion and outlines the future direction of this research.

2. Data Quality Challenges

As data becomes more voluminous and varied in the Big Data era, resolving one data quality issue can lead to the emergence of a new one [10]. Evaluating data quality in this context can be challenging due to the multiple characteristics of Big Data, including its variability, velocity, and volatility. To properly assess and manage the quality of Big Data, it is essential to consider the five V's of Big Data, which are the massive volume of data, the fast velocity of arriving data, and the variety of heterogeneous data [11]. These characteristics can have a significant impact on the quality of the dataset, making it essential to employ effective quality management techniques.

The use of poor quality data can lead to biases and misuse of predictive models. Data quality plays a critical role in predictive models, and multisite distributed research networks face unique data quality challenges that can negatively impact machine learning and predictive modeling [12]. Machine learning research has often prioritized algorithm development over data collection and quality, leading to complaints that research institutions allocate a disproportionate amount of effort towards algorithms [13]. However, it is important to allocate sufficient resources to data preparation and quality to avoid biases and ensure robustness and accuracy in inference pipelines.

Besides, the enormous volume of data makes it difficult to assess data quality within a reasonable period [14,15], and collecting data from various resources increases the challenge of achieving high quality data [16]. Inconsistent, incomplete, and noisy data have a negative impact on variety of data types that need to be addressed when dealing with data quality [17,18]. The process to identify and filter low-quality data is challenging due to the variety of data types and format [15]. Furthermore, the presence of semi-structured or unstructured data makes correlating unstructured data difficult [14].

One of the most significant problems in data quality is its rapid degeneration over time [19,20]. Experts indicate that customer files typically experience a two percent of record obsolescence rate in a month due to various factor such as death, divorce, marriage, and relocation [20,21]. Therefore, timeliness is critical in applications where real-time responses are required, and the processing technology need to meet higher requirements to handle rapidly changing data [14,15].

In addition, the use of IoT devices and blockchain technology in food supply chains aims to ensure product authenticity, but challenges related to data integrity, quality, and interoperability arise due to the dynamic nature of the physical environment [22]. However, the integration of these technologies does not fully address the issue of data quality, as data stored on a blockchain may not be reliable. This has led to concerns about food fraud in the case of Australia's beef export to China, which could negatively impact the country's brand reputation and market value. Thus, there is a need for effective solutions to ensure data accuracy and reliability in the supply chain.

To ensure high quality and veracity of the data, the organisations need to develop a taxonomy or guidelines for measuring Big Data. These guidelines provide a framework for evaluating Big Data quality, enabling organizations to make informed decisions based on accurate data.

3. Existing Data Quality Dimensions

Data quality dimension is important to measure the quality of the data. Four main critical dimensions are commonly used to measure data quality, which are accuracy, consistency, completeness, and timeliness.

Accuracy

Accuracy is one of the essential criteria that identifies whether the data is free from common errors, such as misspelling and typos. The data is considered accurate when it corresponds to the real world [23], it reflects the true state of information and the information collected does not create uncertainty [15]. Accuracy is related to the context of the data. Sometimes it can be easily evaluated if it just has two distinct values: male and female (e.g: for gender). However, if there is no reference value, it might be difficult to assess the accuracy.

Consistency

Meanwhile, consistency refers to the violation of semantic rules defined over a set of rules defined in the system [5,24,25]. It is a measure of information equivalency where after the data have been processed, the concept, value domains, and formats still match even before data processing [15]. Consistency can be divided into two dimensions: structural and semantic [26]. Structural consistency describes the structure value in the data, while semantic consistency refers to the rule that explains the relationship between the fields in the dataset.

Completeness

Another important dimension to measure data quality is completeness. Completeness refers to the degree of information to describe a significant state of the represented real world [25]. It is usually expressed as a percentage of the raw data volume relative to the required data volume [5,27]. It is a quantitative metric for evaluating how many reliable analytical data can be derived from the dataset [11] and it is achieved when all the components that influence data accuracy and integrity are present [15].

Timeliness

Data is being processed in real-time and near real-time. Since data is collected every second and the changes happen quickly, thus timeliness is crucial. Timeliness consists of currency and volatility [28,29]; currency is a measure of its validity in the system or to the real world, and volatility is related to the frequency of data change. It refers to whether data is delivered on time, is modified regularly, and the timeframe between data collection and processing meets the criteria [15]. In contrast, some researchers argued that timeliness should be measured as part of the data quality dimension because it does not expose the relevancy of the data [30].

Data quality dimensions need to be revised to accommodate the five V's of Big Data [31]. It should be able to deal with a large amount of data (volume) coming from various resources (variety) and accommodate the dynamic aspects of data (velocity). Besides, the data should be assessable and accurate (veracity), so that it can be used for decision-making and monetary aspect (value). Moreover, the data quality dimensions are dependable to the context of use, thus no agreed standard set of dimensions that can contribute to high data quality can be defined [32].

4. Is measuring data quality in Big Data using accuracy, consistency, completeness, and timeliness still relevant?

Effective management of data quality requires a combination of people, technology, and processes to ensure compliance and consistency. The quality of data can be described, measured, and managed through various attributes such as accuracy, consistency, completeness and timeliness. Table 1 presents data quality dimension discussed by other researchers.

Table 1: Main data quality dimension discussed by other researchers.

Author	Accuracy	Consistency	Completeness	Timeliness	Other DQ Dimension
Onyeabor et al. [33]	✓	✓	✓	✓	Latency, Response, Throughput, Capacity, Scalability
Abdellaoui et al. [34]	✓	✓	✓		Freshness, Reputation, Credibility, Value Added
Kulkarni [35]	✓	✓	✓	✓	Believability, Relevance, Validity, Verifiability,
Schelter et al. [29]	✓	✓	✓	✓	Currency, Volatility, Interpretability, Accessibility, Maintainability, Uniqueness, Reliability, Applicability
Li et al. [36]	✓	✓	✓	✓	Integrity, Validity, Scientifically, Stability, Sharing, Standardisation,
Ehrlinger and Wöß [37]		✓		✓	Correctness, Pertinence, Normalization, Readability, Minimalist
Taleb et al. [23]	✓	✓	✓		
Zhou et al. [38]			✓		Correctness, Compatibility
Ardagna et al. [39]	✓	✓	✓	✓	Distinctness, Precision, Volume
Franco et al. [24]	✓	✓	✓		Redundancy, Readability, Accessibility, Trust, Usefulness
Xiaojiang et al. [40]					Currency, Trustworthiness
Radhakrishnan and Das [41]	✓	✓	✓	✓	Currency, measurement rate, appropriate amount of data, interpretability, representational consistency, availability, security, licensing
Marotta and Vaisman [42]	✓	✓			
Ya et al. [43]	✓	✓		✓	Authenticity, Integrity, Standardization, Understandability, Traceability, Accessibility, Security
Kugler et al. [44]	✓	✓	✓	✓	Resolution, Redundancy
Talha et al. [45]	✓		✓	✓	Trust, Reputation

From Table 1, most of the papers agreed that the four critical dimensions are needed to measure data quality. However, three out of 16 papers excluded their discussion of data quality dimension. Since Ehrlinger and Wöß [37] focused on the schema quality, only correctness is included as part of the data quality dimension. On the other hand, Xiaojiang et al. [40] only focused on the currency and trustworthiness dimension. However, the authors agreed that accuracy, completeness, consistency, and timeliness are important dimensions to measure data quality. Talha et al. [45] outlined four dimensions under accuracy criteria which are consistency, timeliness, and reputation. These four criteria are used to indicate the accuracy of the reference data.

Out of 16 papers, seven papers excluded timeliness as their data quality dimension. Abdellaoui et al. [34] excluded timeliness as the data quality dimension but added freshness as part of the main data quality dimension whereby freshness refers to how old the data is. Franco et al. [24] and Marotta and Vaisman [43] also referred to timeliness as freshness in the paper, but the authors did not discuss the details on this data quality dimension.

Taleb et al. [23] stated that velocity is an additional quality characteristic, which refers to the high-speed generation of large volumes of data. Thus, there is a need to consider additional quality parameters, including timeliness, in evaluating the quality of data. Since the authors only included some popular data quality dimensions commonly cited, thus only accuracy, consistency, and completeness were included.

Despite arguments on the data quality dimension, four main data quality dimensions: accuracy, completeness, consistency, and timeliness are important in order to measure the data quality. However, these four dimensions are no longer enough as most of the existing research proposed new data quality dimensions to measure the data quality. Therefore, these data quality dimensions will be analyzed to see the needs and importance of it to be included as data quality dimension for Big Data.

5. Data Quality Dimensions for Big Data

Table 2: Taxonomy of data quality dimension.

Data Quality Categories	Data Quality Dimension	Definition
Accessibility	Accessibility	Ease of access and retrieve the data
	Availability	Data is accessible and ready to use
	Security	Protect data from unauthorize access
Contextual	Appropriate Amount of Data	Quantity of data aligned with intended purpose
	Currency	Freshness of the data
	Relevance	Data is directly applicable and meaningful
	Validity	Accuracy of data in representing the intended concept
	Value Added	Additional insight derived from data
	Volatility	Frequency of data being updated
	Completeness	Data contains all necessary elements
	Timeliness	Data is usable within desired timeframe
Intrinsic	Correctness	Correspond to the real-world
	Redundancy	No duplicated data within the dataset
	Reputation	Reflecting level of confidence of the sources
	Trust	Degree of confidence on data
	Accuracy	Correctness in representing real-world
Representational	Interpretability	Ease of understanding the meaning of data
	Readability	Readability of data presentation
	Understandability	User can understand the meaning of the data representation
	Consistent	Uniform data across different resources

From the review, 20 data quality dimensions will be grouped into four categories, which are intrinsic, contextual, accessibility, and representational. These categories are important to provide the guideline to an organisation in choosing which dimension is needed to measure the quality of the data. The categorisation of the data quality dimension is done based on the existing literature. Table 2 shows the taxonomy of data quality dimensions.

The intrinsic category expresses the natural quality of data and can be measured with some reference [46]. It is an objective view of data quality [26] and denotes that the data have quality in their own right [47]. The data quality dimensions that consider as intrinsic are accuracy, reputation, redundancy, correctness, believability, and trust. Reputation is a subjective metric that reflects public opinion regarding the data reliability of the information source. Meanwhile, redundancy refers to minimal use of the information source and every part of the data is represented only once. Data is considered correct if it corresponds to the real world concerning the domain. Since believability is referring to the trustworthiness of the dataset, this data quality dimension is excluded and only trust is being considered. Trust reflects the reliability and trustworthiness of the dataset.

The contextual category is associated with the data value, while the intrinsic dimension refers to the objective and native data attributes [14,29]. It considers the quality of the data within the context and is based on the preference of the user [46,48,49]. This category illustrates the concept of data quality is a factor of the task at hand [47,50]. It consists of seven quality dimensions which are: timeliness, relevance, currency, volatility, value added, validity, and appropriate amount of data.

Next, accessibility, availability and security are grouped as part of accessibility categories. Accessibility category refers to dimensions that define how conveniently data can be accessed by the users [46,47,50]. The information must be readily available and editable with the organisation's information system [51]. Accessibility and availability are intended to measure the ability to access the information within a specific timeframe.

The final category is the representational category. It refers to dimensions that are related to the format and context of the data [46–48]. It evaluates the representational capabilities of the system [52] and emphasises the role of the system [49]. It can be measured using four dimensions which are consistent, understandability, readability, and interpretability. Interpretability refers to the representation of the data in appropriate languages, symbols and units and the definitions are clear. Readability means that the information presented must also be concise, such as the use of a word from the dictionary.

Data quality is essential to determine the reliability and veracity of the data. The concept of data quality dimension lies at the core of data quality management. It can be conceived as a method for evaluating data quality. Ardagna et al. [39] emphasized that data quality dimensions are tightly coupled with the specific application or domain, implying that different contexts may require different dimensions to address their unique requirements. For example, Xiaojiang et al. [4] claimed that dimensions like currency and trustworthiness are crucial to maintain the relevance of information and prevent potential financial loss or misinformation. The data quality requirements specified or required by the various stakeholders involved in the execution of business processes are represented by these data quality dimensions. The importance of specific dimensions can vary based on the intended use of the data, with accuracy and integrity being more critical for social media sentiment analysis and accuracy and completeness being important for power consumption and sensor-based data.

Overall, the proposed taxonomy of data quality dimensions provides a framework for assessing data quality that goes beyond the traditional metrics of accuracy, completeness, consistency, and timeliness. Besides, it is important for organizations to consider the specific use cases of their data and choose appropriate data quality dimensions to ensure high-quality data.

6. CONCLUSION

In conclusion, the emergence of Big Data has revolutionized the way organizations manage and analyze data, with heavy reliance on data for informed decision-making. However, the issue of data quality remains a significant challenge for organizations, with dirty data leading to incorrect decisions and negative consequences. This paper presented a taxonomy of data quality dimension Big Data. This review identifies whether critical dimensions for data quality; accuracy, consistency, completeness, and timeliness are still relevant as data quality measure. In order to develop a standard or guideline for Big Data quality, the criteria of Big Data must be considered, especially the veracity of the data. Future research should focus on identifying suitable metrics for each dimension to guide organizations in evaluating and improving their data quality. By addressing data quality issues, organizations can enhance their decision-making processes, improve business outcomes, and gain a competitive advantage in today's data-driven world.

References

[1] Côrte-Real, Nadine, Pedro Ruivo, and Tiago Oliveira. 2020. "Leveraging Internet of Things and Big Data Analytics Initiatives in European and American Firms: Is Data Quality A Way to Extract Business Value?" *Information and Management* 57 (1).

[2] Verma, Pratima, Vimal Kumar, Ankesh Mittal, Bhawana Rathore, Ajay Jha, and Muhammad Sabbir Rahman. 2023. "The Role of 3S in Big Data Quality: A Perspective on Operational Performance Indicators Using an Integrated Approach." *TQM Journal* 35 (1): 153–82.

[3] Saha, Barna, and Divesh Srivastava. 2014. "Data Quality: The Other Face of Big Data." In *2014 IEEE 30th International Conference on Data Engineering*, 1294–97. Chicago, IL, USA: IEEE.

[4] Ehrlinger, Lisa, Elisa Rusz, and Wolfram Wöß. 2019. "A Survey of Data Quality Measurement and Monitoring Tools." *CoRR* abs/1907.0 (July).

[5] Cappiello, Cinzia, Walter Samá, and Monica Vitali. 2018. "Quality Awareness for a Successful Big Data Exploitation." In *Proceedings of the 22nd International Database Engineering & Applications Symposium*, 37–44. IDEAS 2018. New York, NY, USA: Association for Computing Machinery.

[6] Merino, Jorge, Ismael Caballero, Bibiano Rivas, Manuel Serrano, and Mario Piattini. 2016. "A Data Quality in Use Model for Big Data." *Future Generation Computer Systems* 63 (October): 123–30. https://doi.org/10.1016/j.future.2015.11.024.

[7] Taleb, Ikbal, Rachida Dssouli, and Mohamed Adel Serhani. 2015. "Big Data Pre-Processing: A Quality Framework." In *Proceedings - 2015 IEEE International Congress on Big Data, BigData Congress 2015*, 191–98. IEEE. https://doi.org/10.1109/BigDataCongress.2015.35.

[8] Magoulas, Roger, and Steve Swoyer. 2020. "The State of Data Quality In 2020." O'Reilly. 2020. https://www.oreilly.com/radar/the-state-of-data-quality-in-2020/.

[9] Cichy, Corinna, and Stefan Rass. 2019. "An Overview of Data Quality Frameworks." *IEEE Access* 7: 24634–48.

[10] Guo, Aizhang, Xiuyuan Liu, and Tao Sun. 2018. "Research on Key Problems of Data Quality in Large Industrial Data Environment." In *Proceedings of the 3rd International Conference on Robotics, Control and Automation - ICRCA '18*, 245–48. New York, US: Association for Computing Machinery.

[11] Gao, Jerry, Chunli Xie, and Chuanqi Tao. 2016. "Big Data Validation and Quality Assurance - Issuses, Challenges, and Needs." In *2016 IEEE*

Symposium on Service-Oriented System Engineering (SOSE), 433–41. Oxford, UK: IEEE. https://doi.org/10.1109/SOSE.2016.63.
[12] Sáez, Carlos, Nekane Romero, J. Alberto Conejero, and Juan M. García-Gómez. 2021. "Potential Limitations in COVID-19 Machine Learning Due to Data Source Variability: A Case Study in the NCov2019 Dataset." *Journal of the American Medical Informatics Association* 28 (2): 360–64.
[13] Whang, Steven Euijong, Yuji Roh, Hwanjun Song, Jae-Gil Lee, S E Whang, Y Roh Kaist, H Song, and J Lee. 2021. "Data Collection and Quality Challenges in Deep Learning: A Data-Centric AI Perspective." *VLDB Journal*, December.
[14] Talha, Mohamed, A. Abou El Kalam, Nabil Elmarzouqi, Anas Abou El Kalam, A. Abou El Kalam, and Nabil Elmarzouqi. 2019. "Big Data: Trade-off between Data Quality and Data Security." *Procedia Computer Science* 151: 916–22.
[15] Cai, Li, and Yangyong Zhu. 2015. "The Challenges of Data Quality and Data Quality Assessment in The Big Data Era." *Data Science Journal* 14 (2): 1–10.
[16] Hermans, K., W. Waegeman, G. Opsomer, B. Van Ranst, J. De Koster, M. Van Eetvelde, and M. Hostens. 2017. "Novel Approaches to Assess the Quality of Fertility Data Stored in Dairy Herd Management Software." *Journal of Dairy Science* 100 (5): 4078–89.
[17] Abdallah, Mohammad. 2019. "Big Data Quality Challenges." In *Proceedings of the 2019 International Conference on Big Data and Computational Intelligence, ICBDCI 2019*, 1–3. Pointe aux Piments, Mauritius.: IEEE.
[18] Hariri, Reihaneh H., Erik M. Fredericks, and Kate M. Bowers. 2019. "Uncertainty in Big Data Analytics: Survey, Opportunities, and Challenges." *Journal of Big Data* 6 (1): 1–16.
[19] Amini, Mohammadhossein, and Shing Chang. 2017. "Assessing Data Veracity for Data-Rich Manufacturing." In *Proceedings of the 2017 Industrial and Systems Engineering Research Conference*, 1661–66. Pittsburgh, PA.
[20] Noraini, Abdullah, Saiful Adli Ismail, Siti Sophiayati, and Suriani Mohd Sam. 2015. "Data Quality in Big Data: A Review." *International Journal of Advances in Soft Computing and Its Applications* 7 (Special Issue 3): 16–27.
[21] Eckerson, Wayne W. 2002. "Achieving Business Success through a Commitment to High Quality Data." www.dw-institute.com.
[22] Powell, Warwick, Marcus Foth, Shoufeng Cao, and Valéri Natanelov. 2022. "Garbage in Garbage out: The Precarious Link between IoT and Blockchain in Food Supply Chains." *Journal of Industrial Information Integration* 25 (January): 100261.
[23] Taleb, Ikbal, Hadeel T.El E Kassabi, Mohamed Adel Serhani, Rachida Dssouli, and Chafik Bouhaddioui. 2017. "Big Data Quality: A Quality Dimensions Evaluation." *Proceedings - 13th IEEE International Conference on Ubiquitous Intelligence and Computing, 13th IEEE International Conference on Advanced and Trusted Computing, 16th IEEE International Conference on Scalable Computing and Communications, IEEE International*, no. February 2018: 759–65.
[24] Arolfo, Franco, and Alejandro Vaisman. 2018. "Data Quality in a Big Data Context." In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, edited by András Benczúr, Bernhard Thalheim, and Tomáš Horváth, 11019 LNCS:159–72. Lecture Notes in Computer Science. Cham: Springer International Publishing.
[25] Schelter, Sebastian, Dustin Lange, Philipp Schmidt, Meltem Celikel, Felix Biessmann, and Andreas Grafberger. 2018. "Automating Large-Scale Data Quality Verification." In *Proceedings of the VLDB Endowment*, 11:1781–94. Rio de Janeiro, Brazil.
[26] Micic, Natasha, Daniel Neagu, Felician Campean, and Esmaeil Habib Zadeh. 2017. "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)*, 2018-Janua:155–62. Exeter, UK.
[27] Cheng, Hongju, Danyang Feng, Xiaobin Shi, and Chongcheng Chen. 2018. "Data Quality Analysis and Cleaning Strategy for Wireless Sensor Networks." *EURASIP Journal on Wireless Communications and Networking* 2018 (1): 61.
[28] Taleb, Ikbal, Mohamed Adel Serhani, and Rachida Dssouli. 2018. "Big Data Quality: A Survey." In *2018 IEEE International Congress on Big Data (BigData Congress)*, 166–73. San Francisco, California, USA.
[29] Serhani, Mohamed Adel, Hadeel T. El Kassabi, Ikbal Taleb, and Alramzana Nujum. 2016. "An Hybrid Approach to Quality Evaluation across Big Data Value Chain." In *2016 IEEE International Congress on Big Data (BigData Congress)*, 418–25. San Francisco, CA, USA: IEEE.
[30] Jaya, M. Izham, Fatimah Sidi, Lilly Suriani Affendey, Marzanah A Jabar, and Iskandar Ishak. 2019. "Systematic Review of Data Quality Research." *Journal of Theoretical and Applied Information Technology* 97 (21): 3069–83.
[31] Geerts, Floris, Paolo Missier, and Norman Paton. 2018. "Editorial: Special Issue on Improving the Veracity and Value of Big Data." *Journal of Data and Information Quality*.
[32] Jaya, M. Izham, Fatimah Sidi, Iskandar Ishak, Lilly Suriani Affendey, and Marzanah A. Jabar. 2017. "A Review of Data Quality Research in Achieving High Data Quality Within Organization." *Journal of Theoretical and Applied Information Technology* 95 (12): 2647–57.
[33] Onyeabor, Grace Amina, and Azman Ta'a. 2019. "A Model for Addressing Quality Issues in Big Data." In *Recent Trends in Data Science and Soft Computing*, edited by Faisal Saeed, Nadhmi Gazem, Fathey Mohammed, and Abdelsalam Busalim, 843:65–73. Kuala Lumpur, Malaysia: Springer.
[34] Abdellaoui, Sabrina, Ladjel Bellatreche, and Fahima Nader. 2016. "A Quality-Driven Approach for Building Heterogeneous Distributed Databases: The Case of Data Warehouses." In *Proceedings - 2016 16th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing, CCGrid 2016*, 631–38. CCGRID '16. Cartagena, Columbia: IEEE.
[35] Kulkarni, Anuja. 2016. "A Study on Metadata Management and Quality Evaluation in Big Data Management." *International Journal for Research in Applied Science & Engineering Technology (IJRASET)* 4 (VII): 455–59.
[36] Li, Zhe, Sai Wu, Hongwei Zhou, Sheng Zou, and Tingting Dong. 2019. "An Overview on Quality Evaluation Constitution in Context of Big Data Application." *Journal of Physics: Conference Series* 1302 (2): 1–6.
[37] Ehrlinger, Lisa, and Wolfram Wöß. 2017. "Automated Data Quality Monitoring." In *22nd MIT International Conference on Information Quality*, 1–9. Little Rock, Arkansas, USA
[38] Zhou, Ningning, Guofang Huang, and Suyang Zhong. 2018. "Big Data Validity Evaluation Based on MMTD." *Mathematical Problems in Engineering* 2018 (June): 1–6.
[39] Ardagna, Danilo, Cinzia Cappiello, Walter Samá, and Monica Vitali. 2018. "Context-Aware Data Quality Assessment For Big Data." *Future Generation Computer Systems* 89 (December): 548–62.
[40] Xiaojiang, Fan, Zheng Liwei, and Liu Jianbin. 2017. "Measurement for Social Network Data Currency and Trustworthiness." In *2017 2nd IEEE International Conference on Cloud Computing and Big Data Analysis, ICCCBDA 2017*, 1–5. Chengdu, China: IEEE.
[41] Radhakrishnan, Asha, and Sarasij Das. 2018. "Quality Assessment of Smart Grid Data." In *2018 20th National Power Systems Conference*

(NPSC), 1–6. Tiruchirappalli, India: IEEE.

[42] Marotta, Adriana, and Alejandro Vaisman. 2016. "Rule-Based Multidimensional Data Quality Assessment Using Contexts." Edited by Sanjay Madria and Takahiro Hara. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Lecture Notes in Computer Science, 9829 LNCS: 299–313.

[43] Ya, Li, Song Heliang, and Xu Yingcheng. 2020. "Studies on Data Quality Evaluation Index System for Internet Plus Government Services in Big Data Era." *Journal of Physics: Conference Series* 1584 (1): 012014.

[44] Kugler, Zs., Gy. Szabó, H. M. Abdulmuttalib, C. Batini, H. Shen, A. Barsi, and G. Huang. 2018. "Time-Related Quality Dimensions of Urban Remotely Sensed Big Data." *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives* 42 (4): 383–87.

[45] Talha, Mohamed, Nabil Elmarzouqi, and Anas Abou El Kalam. 2020. "Towards A Powerful Solution for Data Accuracy Assessment in The Big Data Context." *International Journal of Advanced Computer Science and Applications* 11 (2): 419–29.

[46] Laranjeiro, Nuno, Seyma Nur Soydemir, and Jorge Bernardino. 2016. "A Survey on Data Quality: Classifying Poor Data." In *Proceedings - 2015 IEEE 21st Pacific Rim International Symposium on Dependable Computing, PRDC 2015*, 179–88. Zhangjiajie, China: IEEE.

[47] Juddoo, Suraj, Carlisle George, Penny Duquenoy, and David Windridge. 2018. "Data Governance in the Health Industry: Investigating Data Quality Dimensions within a Big Data Context." *Applied System Innovation* 1 (4): 43.

[48] Immonen, Anne, Pekka Pääkkönen, and Eila Ovaska. 2015. "Evaluating the Quality of Social Media Data in Big Data Architecture." *IEEE Access* 3: 2028–43.

[49] Maqboul, Jaouad, and Bouchaib Bounabat. 2018. "An Approach of Data-Driven Framework Alignment to Knowledge Base." In *LOPAL '18: Proceedings of the International Conference on Learning and Optimization Algorithms: Theory and Applications*, 1–5. Rabat, Morocco.

[50] Weichselbraun, Albert, and Philipp Kuntschik. 2017. "Mitigating Linked Data Quality Issues in Knowledge-Intense Information Extraction Methods." In *WIMS '17: Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics*. Vol. Part F1294. Amantea, Italy.

[51] Wanner, Jonas, and Christian Janiesch. 2019. "Big Data Analytics in Sustainability Reports: An Analysis Based on The Perceived Credibility of Corporate Published Information." *Business Research* 12 (1): 143–73.

[52] Nemeth, Tanja, Fazel Ansari, and Wilfried Sihn. 2019. "A Maturity Assessment Procedure Model for Realizing Knowledge-Based Maintenance Strategies in Smart Manufacturing Enterprises." In *Procedia Manufacturing*, 39:645–54. Chicago, Illinois, USA: Elsevier B.V.