Future Computing and Informatics Journal

Volume 6 | Issue 1 (2021) Article 3

2021

Data Quality Dimensions, Metrics, and Improvement Techniques

Menna Ibrahim Gabr lecturer Assistant
Helwan University, mennatallah.gabr@gmail.com

Yehia mostafa helmy prof
helwan university, ymhelmy@commerce.helwan.edu.eg

Doaa Saad Elzanfaly assoc. prof
helwan university, doaa.saad@fci.helwan.edu.eg

Recommended Citation

Gabr, Menna Ibrahim lecturer Assistant; helmy, Yehia mostafa prof; and Elzanfaly, Doaa Saad assoc. prof (2021) "Data Quality Dimensions, Metrics, and Improvement Techniques," *Future Computing and Informatics Journal*: Vol. 6: Iss. 1, Article 3.
DOI: http://doi.org/10.54623/fue.fcij.6.1.3
Available at: https://digitalcommons.aaru.edu.jo/fcij/vol6/iss1/3

Future Computing and Informatics Journal

Volume 6 | Issue 1 (2021) Article 3

2021

Data Quality Dimensions, Metrics, and Improvement Techniques

Menna Ibrahim Gabr lecturer Assistant
Helwan University, mennatallah.gabr@gmail.com

Yehia mostafa helmy prof
helwan university, ymhelmy@commerce.helwan.edu.eg

Doaa Saad Elzanfaly assoc. prof
helwan university, doaa.saad@fci.helwan.edu.eg

Recommended Citation

Gabr, Menna Ibrahim lecturer Assistant; helmy, Yehia mostafa prof; and Elzanfaly, Doaa Saad assoc. prof (2021) "Data Quality Dimensions, Metrics, and Improvement Techniques," *Future Computing and Informatics Journal*: Vol. 6 : Iss. 1 , Article 3.
DOI: http://doi.org/10.54623/fue.fcij.6.1.3
Available at: https://digitalcommons.aaru.edu.jo/fcij/vol6/iss1/3

Future Computing and Informatics Journal

Homepage: https://digitalcommons.aaru.edu.jo/fcij/

doi: http://Doi.org/10.54623/fue.fcij.6.1.3

DATA QUALITY DIMENSIONS, METRICS, AND IMPROVEMENT TECHNIQUES

[1, a] Menna Ibrahim Gabr, Yehia M. Helmy[1, b], Doaa Saad Elzanfaly[1,c]
[1] Helwan University
[a] mennatallah.gabr@gmail.com,[b] ymhelmy@commerce.helwan.edu.eg,
[c]Doaa.saad@fci.helwan.edu.eg

ABSTRACT

Achieving high level of data quality is considered one of the most important assets for any small, medium and large size organizations. Data quality is the main hype for both practitioners and researchers who deal with traditional or big data. The level of data quality is measured through several quality dimensions. High percentage of the current studies focus on assessing and applying data quality on traditional data. As we are in the era of big data, the attention should be paid to the tremendous volume of generated and processed data in which 80% of all the generated data is unstructured. However, the initiatives for creating big data quality evaluation models are still under development. This paper investigates the data quality dimensions that are mostly used in both traditional and big data to figure out the metrics and techniques that are used to measure and handle each dimension. A complete definition for each traditional and big data quality dimension, metrics and handling techniques are presented in this paper. Many data quality dimensions can be applied to both traditional and big data, while few number of quality dimensions are either applied to traditional data or big data. Few number of data quality metrics and barely handling techniques are presented in the current works.

KEYWORD: Data Quality, Data Quality Dimensions, Big Data, Traditional Data.

1. INTRODUCTION

Data are explosively increasing, they become more complicated and diversified. These data come from various sources like sensors, meters, GPSs, and at least 80 percent of new data are unstructured, such as Web contents, Web logs, email, image, videos etc. Traditionally, it has been well known that problems related to data quality, such as, incomplete, redundant, inconsistent...etc. pose a major challenge making the whole process of using and processing this data useless.

Data Quality (DQ) refers to how relevant, precise, useful, in context, understandable and timely data is. Data is of high quality if it satisfies the requirements stated in a particular specification that reflects the implied needs of the user. In other words, data quality is often defined as 'fitness for use', i.e. an evaluation of to which extent data serve the purposes of the user[1].

Data Quality Dimensions (DQDs) have a great role in data quality assessment. Data can be clearly described and measured through its dimensions. There are numbers of data quality dimensions in the literature,

https://digitalcommons.aaru.edu.jo/fcij/vol6/iss1/3
DOI: http://doi.org/10.54623/fue.fcij.6.1.3

however, they are broadly classified into four groups: Intrinsic, Contextual, Representational and Accessibility According to Wang & Strong [2], the Intrinsic DQD denotes that data have quality in their own right, this means that the data itself should have a high level of believability, Objectivity and reputation. The user should trust the data before working on it. Contextual DQD highlights that data quality must be considered within the context of the task at hand; that is, data must be relevant, timely, complete, and appropriate in terms of the add value. Representational DQD is related to the presentation of data in terms of its format that communicates its meaning. Where Accessibility DQD is more related to accessing and protecting data[2]. These dimensions are assessed to evaluate by how much the data is qualified for a specific use. Dimensions are measured either objectively or subjectively. Subjective measurements are based on measuring how far the data is fit for use by the consumers. In most cases, measures are based on scaled-response questionnaires that weight the value of each dimension from four different views: definition, synonym, direct, and reverse. Whereas Objective measurements are used to evaluate to which extent data conforms to specifications. A simple ratio between the undesirable outcomes and the total outcomes is usually used as an objective measure [3]. To objectively assess different dimensions' percentage in datasets, a simple ratio metric is used. For example, to know the percentage of missing data in your dataset we can use (number of missing values/Total number of records).

Nowadays, **Big Data** considered to be one of the dominant research areas. Generally speaking, Big Data is huge volumes of data generated with high speed, and has varying degree of complexity and ambiguity. Big data cannot be processed, stored, and managed with traditional methods and algorithms. It needs new platforms and architectures that enable high-velocity capture, discovery, and analysis to extract values. Big data can be also defined through its 3Vs, volume, variety, and velocity at which the data is generated, collected, and processed. More Vs are added by time like value, veracity, complexity, and others[4].

To develop insights from the Big Data, a variety of methods from statistics, machine learning, data mining, visualization, and databases are used. However, ensuring the quality of data is a necessity for getting more beneficial insights from big data. That is why, checking Data Quality is consider an important integral part of any process in both Traditional and Big Data.

The rest of this paper is organized in the following sections: Sections 2 and 3 discuss different data quality dimensions along with their assessment metrics and quality improvement techniques in traditional data and big data respectively. Section 4 summarizes the literature and Section 5 concludes the paper.

2. DATA QUALITY DIMENSIONS ON TRADITIONAL DATA

This section discusses the traditional data quality from three aspects: Data quality dimensions, the metrics and improvement techniques for each dimension. We present different studies from the literature that deals with data quality dimensions in traditional data in the first subsection, where the metrics used to measure the ratio of each dimension are presented in section 2.2. The techniques that are used to handle the data quality of each dimension are presented in section 2.3.

Published by Arab Journals Platform, 2021

2.1 Traditional Data Quality Dimensions

For traditional data, there are around 28 data quality dimensions that have been studied in different researches. These dimensions have been applied either combined or individually on traditional data. Among these dimensions are believability, accuracy, objectivity, reputation, value-added, relevancy, timeliness, completeness, Interpretability, ease of understanding and representational. A brief definition for each dimension is presented in Table 1. DQ dimensions can be ranked according to their significance based on how extensively they have been studied and presented in the literature. The two data quality dimensions that come in the first rank are Completeness [5]–[11] and Relevance [5], [10]–[15]. In [6], the Completeness dimension is divided into completeness of case ascertainment and completeness of the items. Accessibility comes in the second rank as it has been also considered in different studies [5]–[7], [10], [12], but less extensively than Completeness and Relevancy. Other dimensions like Timeliness[5], [6], [9], [10], Accuracy [3], [5], [7], [10], Consistency [5], [8]–[10], Reputation [5], [8], [10], [12] and Objectivity [5], [7], [10], [12] come in the third rank. In [8], the Consistency Dimension is divided into two other dimensions: Semantic Consistency and Structural Consistency.

Following the same ranking, Understandability [5], [10], [12], Representational Format [5], [7], [10]and Interpretability [7], [10], [12] came in the fourth class. While Duplication [6], [11], Believability and Value-Added [10], [12] have less presence in the literature. The least or barely mentioned data quality dimensions are Usefulness, Validity, Comparability [6], Coherence, Actuality, Statistical Disclosure Control, Optimal use of Resources, Utility, Informative [7], Correctness [9], Appropriate Amount of Information and Security [10].

2.2 Traditional Data Quality Metrics

Among the 28 data quality dimensions mentioned above, only 16 of them have assessment metrics.

It has been notices that most of these metrics are more related to the dome of the data itself. For instance, the metrics that are proposed for the health data differ from those used in social media data. Moreover, the same dimension may be divided into two or three different sub-dimensions, each is measured differently. However, there are more generic metrics that can be used regardless of the domain as shown in Table 1. Different metrics may be used for the same data quality dimension. In this section, these metrics are presented.

In [5], seven different metrics have been used to measure different data quality dimensions. **Respondent Opinion** is used to measure Completeness, Timeliness, Accuracy, Relevancy, Consistency, Understandability, Representational Format, Security, Accessibility, Reputation and Objectivity. **Element Presence** is used to measure Completeness and Accessibility. **Gold Standard** is used to measure Completeness, Accuracy and Consistency. Furthermore, **Data Source Agreement** measure both the Completeness dimension and Accuracy dimension. **Log Review** is used to measures the Timeliness dimension. **Data Element Agreement** is used as a metrics for Accuracy, Consistency, Relevancy and Understandability. Finally, **Validity Check** is used to check the Accuracy dimension and Representational Format dimension.

Deterministic and Probabilistic Matches are used to find the Duplicate Records, Deterministic matching is used to determine a match or an exact comparison between fields, while in probabilistic or fuzzy, several field values are compared between two records and each field is assigned a weight that indicates how closely the two field

https://digitalcommons.aaru.edu.jo/fcij/vol6/iss1/3
DOI: http://doi.org/10.54623/fue.fcij.6.1.3

values match. While Kappa Coefficient checks the Validity and reliability of diagnostic tests as they work on health domain [16][17]. For certain purpose, the Completeness dimension is divided into Completeness of case ascertainment which is measured by pooling method [18], which combine data of interest from two or more sources, and screening method [18] which isolates and identifies a group of components in a sample with the minimum number of steps and the least manipulation. More basically, a screening method is a simple measurement providing a "yes/no" response. And Completeness of the items which is measured by missing values. Standardization of definitions, use of standard clinical vocabularies, terminologies, classifications and ontology are used to check Comparability [19]. Furthermore Accessibility can be measured by the availability of data [20]. While adaptability or the capacity to include new data items is used to figure out the Usefulness of the data. In this case, Timeliness refers to the rapidity at which a registry can collect, process and report sufficiently reliable and complete data to take actions, so Timeliness dimension is measured by four criteria through the following steps [21], Step (1) Time until receipt: time from the clinical event to the record in the registry. Step (2) Processing: the time from the presence of the record to its availability for research. Step (3) Availability. Step (4) Reporting: Number of patients or data recorded in the registry after the database was 'frozen' to produce an annual report [6].

2.3 Traditional Data Quality Improvement Techniques

Throughout this section the techniques used to improve traditional data quality problems are presented. Out of the 28 dimensions mentioned before, only 3 of them are on the focus of researchers with consideration to propose improvement techniques, as shown in Table 1.

The Completeness improvement techniques are KNN imputation, mean / median/mode imputation, lit wise deletion[22][23]. While Duplication handling techniques are standard duplicate elimination algorithm, Duplicate Count Strategy (DCS), Duplicate Count Strategy (DCS++) [24], sorted neighborhood algorithm [25] and sorted blocks [11].

Many improvement techniques for Relevancy (Feature Selection) dimension are proposed in the literature [11], [13]–[15]. Among these techniques are Filter Approach [15], Wrapper Approach, Embedded Approach [26] and Hybrid Feature Selection Approach [27][28]. In filter approach the attribute selection method is independent of the machine learning (ML) algorithm used and it assess the relevance of features by looking only at the intrinsic properties of the data. Where in Wrapper approach, the attribute selection method uses the result of the ML algorithm to determine how good a given attribute subset is. Moreover, the hybrid approach combines both wrapper and filter technique to gain the advantages of both methods. In addition to the previously mention techniques, Fast Correlation-Based Filter (FCBF) [29], Best First Search Algorithm (BFS) [30], CfsSubset Evaluator (CSER), Chi-Squared Attribute Evaluator (CSAER) [31], Information Gain Attribute Evaluator (IGAER) and Relief Attribute Evaluator (RAER) [31] are all used techniques to handle relevance dimensions. The BFS searches the attribute subsets space via a method of Greedy Hill Climbing improved with a backtracking aptitude. Where CSER evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. On the other side CSAER evaluates an attribute

Published by Arab Journals Platform, 2021

by computing the value of the chi-squared statistic with respect to the class. The IGAER evaluates an attribute by measuring the information gain with respect to the class Info Gain. Finally, RAER evaluates the worth of an attribute by repeatedly sampling an instance and considering the value of the given attribute for the nearest instance of the same and different class. In [15] they confess that Filter methods are the best choice for the high dimensional data. Also ranking-based methods are the best choice for selecting the relevant features.

We can point out from these studies that a wide range of data quality dimensions have been introduced and can be applied on traditional data. A few number of the studies present how to measure the dimensions they introduced, the objective measurements (ratio %) are their main choice. Furthermore, few papers mentioned how to handle and improve the quality dimensions. Papers are either presenting the measurements or the improvement techniques for most common data quality dimensions.

3. DATA QUALITY ON BIG DATA

Section 3.1covers different works that apply data quality dimensions on big data, there is a little difference in big data quality dimensions compared to traditional data quality dimensions. The big data quality metrics are presented in section 3.2, and in section 3.3 the techniques that are used to overcome data quality problems in big data are depicted. All mentioned big data quality dimensions, metrics and techniques are presented in Table 1.

3.1 Big Data Quality Dimensions

On the track of traditional data quality dimensions, the researchers in big data area follow the same traditional quality dimensions. Such as, Accuracy, Consistency, and Completeness [32]–[38]. In addition to the aforementioned data quality dimensions, in [36] they mentioned the Precision, Distinctness, Timeliness and Volume as big data related dimensions. In [37], the Availability has been added, in addition to the aforementioned data quality dimensions.

In [33], they handle the previous dimensions through a discovery model for the Big Data Quality Rules (DQRs). The DQRs consist of: (a) Big Data sampling and profiling, (b) Big data quality mapping and evaluation, (c) Big data quality rules discovery (e) DQR validation and (f) DQR optimization. Generally speaking, they measure the level of data quality (in ratio e.g. 50%), then compare it with the quality requirements (e.g. 90%). After this a rule (e.g. Technique) is generated to improve the level of data quality. After applying the rule, the level of data quality is measured again to validate the rule, if it satisfies the requirements, then the rule will be optimized.

The authors of [39] introduced a new concept for big data quality dimensions by presenting Data quality-in-Use model. Based on the interpretation provided by ISO/IEC 25010 the Quality-in-Use: is the sort of quality perceived by the final user, or the extent of fulfillment of the goals set for data. The main Data Quality concern when assessing the level of Quality-in-Use in Big Data projects is the Adequacy. So they identify three critical Data Quality characteristics: Contextual Adequacy, Temporal Adequacy and Operational Adequacy. Each one of these categories contains a number of dimensions that is somehow related to the big data 3Vs (Volume, Velocity and Variety).
Contextual Adequacy refers to the capability of different datasets to be used, for analysis within the same domain, independently of any format, any size or velocity of the flow. It contains (Relevancy, Completeness,

https://digitalcommons.aaru.edu.jo/fcij/vol6/iss1/3
DOI: http://doi.org/10.54623/fue.fcij.6.1.3

Uniqueness, Semantically Interoperable, Semantically Accurate, Credibility, Confidentiality and Compliance). Temporal Adequacy refers to the fact that data is within an appropriate time slot for the analysis. It focuses on the temporal aspects of the data itself. Temporal adequacy includes (Time-Concurrent, Currentness, Timely Updated, Frequency and Time-Consistent). While Operational Adequacy refers to the extent to which data can be processed in the intended analysis by an adequate set of technologies without leaving any piece of data outside the analysis. It contains (Accessibility, Compliance, Confidentiality, Efficiency, Precision, Traceability, Availability, Portability and Recoverability) [39].

The main quality dimensions that are used in most of the Big Data quality measurement are categorized based on four perspectives [40]. (a) Data perspective, which are similar to the traditional data quality dimensions mentioned in section 2.1, Plus some other new dimensions such as Currency, Cohesion, Usability, Privacy, Accountability, Complexity, Minimality, Compactness, Conciseness and Scalability. (b) Management perspective that includes: Organization management, big data management, Data quality assurance, Integrity constraints, Data edits, Business rules and Reputation. (c) Processing and Service perspective that has Data collection issues, Data conversion issues, Data service scalability issues and Data transform issues). Finally, (d) The User perspective that covers Data Visualization, Trust, Pertinence, Readability, Comprehensibility, Clarity, Simplicity, Relevance, Completeness, Accessibility, Availability, Technologically Available, Believability and Reliability.

Data quality dimensions in [41] are quite the same as quality dimensions mentioned in this section. Availability, usability, reliability, relevance, and presentation are the big data quality dimensions used to assess the level of data quality. Differently, each dimension has sub elements and each element has an indicator to correctly assess the level of quality. The elements of Availability are (accessibility, authorization, and timeliness). While Usability has (data definition/documentation, Credibility, and metadata). Reliability consists of (accuracy, consistency, completeness, integrity, and auditability). Relevance is measured by Fitness of data. Furthermore, Presentation quality has (readability and structure).

Readability and Trust are big data quality dimensions which are used besides the traditional data quality dimensions (Accuracy, Completeness, Accessibility and Consistency) to examine the relationship between data quality and big data [42].

Noise, Heterogeneity, Commercial Sensitivity, provenance (trust), Incompleteness, Inconsistency, Redundancy, Amount of data, Timeliness, and Accessibility [43], are used to figure out the level of quality in 13 datasets that have been extensively used in the research.

After investigating the current studies we can point that Precision, Availability, Semantically Interoperable, Semantically Accurate, Credibility, Confidentiality, Compliance, Time-Concurrent, Currentness, Timely Updated, Frequency, Time-Consistent, Efficiency, Traceability, Portability, Recoverability, Cohesion, Usability, Privacy, Accountability, Complexity, Minimality, Compactness, Conciseness, Scalability, Readability, Pertinence, Comprehensibility, Clarity, Simplicity, Technologically Available, Believability, Reliability, Auditability, Noise, Heterogeneity, and Commercial

Sensitivity are considered as big data related quality dimensions.

3.2 Big Data Quality Metrics

Many of big data studies use the same traditional metrics for measuring the level of data quality which is (1 – (Nic/Tn), where Nic represent the number of incorrect values and Tn represent the total number of records) [32]–[34]. However few papers figured out that these used formulas are more related to traditional data, but they didn't mention any big data related metrics [35]. In Table 1, big data quality metrics are presented.

Differently from the traditional used metrics, ISO/IEC 25024 was used as a metric for the data quality dimensions [39]. ISO/IEC 25024 contains number of concepts and relationships between these concepts to measure data quality dimensions. Like quality measure, quality measure elements, property, etc.

In [36], they used traditional metrics to measure Accuracy, Completeness, and Consistency. However, different metrics are used such as; **Distinctness** to measures the percentage of unique values in a dataset, **Precision** is used to measure the degree to which the values of an attribute are close to each other. In particular, precision is derived by considering the mean, and the standard deviation of all the values of the considered attribute. The **Timeliness** is evaluated by the timestamp that is related to the last update (currency) and the average validity of the data (volatility) [44]. The **Volume** is the used to measure the percentage of values contained in the analyzed Data Object. C4.5 algorithm in [40] is used to measure **Noise** by classifying if data is noisy or not. While **Commercial Sensitivity** is measured by searching data item anonymization or transformation. Commercial sensitivity information could also be indicated as part of the metadata. **Heterogeneity** was difficult to determine, so they derive this information from publications that had used these datasets previously.

3.3 Big Data Quality Improvement Techniques

Unlike traditional data quality handling techniques, few studies [32]–[34], in the area of big data quality, are presented to improve data quality. Among these, the techniques that are using the traditional data techniques such as handling data completeness problem by discarding the missing values, filling the missing values with the mean value, normal value, filled with zeros, or combination of them. While clearing the inconsistency of decimal to integer by rounding the value to the nearest integer is considered as a solving technique to handle data Consistency issues. However, up to our knowledge, there are no specific techniques for handling big data quality issues. The available handling techniques are presented in Table 1.

4. SUMMARIZATION

To summarize the above sections, we can point out that most of big data studies are using the traditional data quality dimensions in the big data cases. Furthermore, they are using the same metrics to measure each dimension and the same techniques to handle data quality problems. Many of these studies present different data quality dimensions that are more related to big data. However, few number of big data related metrics and barely big data solving techniques are presented. While other study admitted that data quality dimensions are case specific, based on the big data situation, the data quality dimensions should be carefully selected. In this section a summary of 63 data quality dimensions, their metrics and handling techniques that are used for both big data and traditional data are presented in Table 1.

https://digitalcommons.aaru.edu.jo/fcij/vol6/iss1/3
DOI: http://doi.org/10.54623/fue.fcij.6.1.3

Table 1. Data Quality Dimensions on Big Data and Traditional Data

Dimension	Definition	TD	BD	Metrics (Measurements)	Solving Technique	
					SD	BD
Accessibility	The extent to which data is available, or easily and quickly retrievable. [5]–[7], [9], [10], [12], [41], [45]	✓	✓	Availability of data. [6]	Not mentioned in the studies.	
Accountability	Refers to the ability to know when someone performs an action on data and to hold them responsible for that action. [46]	X	✓	Not mentioned in the studies.	Not mentioned in the studies.	
Accuracy	The closeness of measurement to the true value of the quality being measured. [5], [7], [9], [10], [32], [36], [41], [45]	✓	✓	$1 - \frac{\text{Number of incorrect data units}}{\text{Total Rows}}$ [3], [32], [34]	Not mentioned in the studies.	
Actuality	Refers to data collection and processing speed and frequency of renewal.[7]	✓	X	Not mentioned in the studies.	Not mentioned in the studies.	
Amount of data (Volume)	The extent to which the volume of data is appropriate for the task at hand. [7], [10], [36], [45]	✓	✓	Total number of records.[43] percentage of values contained in the analyzed data object. [36]	Not mentioned in the studies.	
Auditability	Means that auditors can fairly evaluate data accuracy and integrity within rational time and manpower limits during the data use phase. [41]	X	✓	Not mentioned in the studies.	Not mentioned in the studies.	
Availability	The extent to which data is accessible, easily made public or easily purchased, and timely updated. [45]	X	✓	Not mentioned in the studies.	Not mentioned in the studies.	
Believability	The extent to which data is credible and true. [7], [10], [12], [45]	✓	✓	Not mentioned in the studies.	Not mentioned in the studies.	

Table 1. (Continue)

Dimension	Definition	TD	BD	Metrics (Measurements)	Solving Technique	
					SD	BD
Clarity	Refer to ease of understanding of data by users. [42]	X	✓	Not mentioned in the studies.	Not mentioned in the studies.	
Coherence	Reflects the degree to which data can be successfully brought together, it covers the internal consistency of data collection as well as its comparability. [6], [7]	✓	✓	Not mentioned in the studies.	Not mentioned in the studies.	
Cohesion	Refer to the capability of data to comply without contradictions to all integrity constraints, data edits, business rules and other formalisms. [42]	X	✓	Not mentioned in the studies.	Not mentioned in the studies.	
Commercial Sensitivity	Is considered as one of the factors that restrains the use of provenance.[43]	X	✓	Search data item anonymization or transformation. Commercial sensitivity information could also be indicated as part of the metadata.[43]	Not mentioned in the studies.	
Compactness	Refer to the capability of representing the reality of interest with the minimal use of informative resources. [42]	X	✓	Not mentioned in the studies.	Not mentioned in the studies.	
Comparability	The extent to which the data can be analyzed to make a comparison with other registries over time. This is very important in the analysis of geographical and temporal distribution. [6]	✓	NA	Not mentioned in the studies.	Standardization of definitions, use of standard clinical vocabularies, terminologies, classifications and ontology, is the only sure way to achieve the international comparability. [6]	

https://digitalcommons.aaru.edu.jo/fcij/vol6/iss1/3
DOI: http://doi.org/10.54623/fue.fcij.6.1.3

Table 1. (Continue)

Dimension	Definition	TD	BD	Metrics (Measurements)	Solving Technique	
					SD	BD
Completeness	The extent to which data is not missing and has a sufficient breadth and depth for the task at hand. [5], [7]–[11], [32], [36], [39], [41], [45]	✓	✓	$1 - \frac{Number\ of\ incomplete\ data\ rows}{Total\ number\ of\ rows}$ [8], [9], [32], [34] ISO/IEC 25024[39]	KNN imputation, mean / median/mode imputation, lit wise deletion [11].	
Complexity	Refers to an attribute on which it is difficult to define an ordered relationship which can be objectively assessed. [48]	NA	✓	Not mentioned in the studies.	Not mentioned in the studies.	
Compliance	The extent to which data is compliant to the stated regulations and requirements. [39]	X	✓	ISO/IEC 25024 [39]	Not mentioned in the studies.	
Comprehensibility	Refer to ease of understanding of data by users. [42]	X	✓	Not mentioned in the studies.	Not mentioned in the studies.	
Conciseness	Refer to the capability of representing the reality of interest with the minimal use of informative resources. [42]	X	✓	Not mentioned in the studies.	Not mentioned in the studies.	
Confidentiality	Data must be used and accessed by authorized group of people. [39]	X	✓	ISO/IEC 25024 [39]	Not mentioned in the studies.	
Consistency	Describes the logical coherence of the data with respect to logical rules and constraints. [5], [7], [9], [10], [32], [36], [41], [45]	✓	✓	$1 - \frac{Number\ of\ \text{inconsistent units}}{Total\ \text{Number of consistency checks performed}}$ [9], [32], [34]	Rounding the value to the nearest integer in the case of inconsistent decimal or integers.[32]	

Table 1. (Continue)

Dimension	Definition	TD	BD	Metrics (Measurements)	Solving Technique	
					SD	BD
Correctness	The extent to which data is correct and reliable. [9]	✓	✓	$1 - \frac{\text{Number of incorrect data units}}{Total\ number\ of\ rows}$ [9]	Not mentioned in the studies.	
Credibility	The extent to which data/ source of data has high level of believability and trust. [39], [41]	X	✓	ISO/IEC 25024 [39]	Not mentioned in the studies.	
Currentness	The extent to which data must be similar in age (Timing). The degree to which a datum is up-to-date. [39], [45]	X	✓	current Time – update Time[32] ISO/IEC 25024 [39]	Not mentioned in the studies.	
Duplication/ Uniqueness	A measure of unwanted duplication existing within a particular field, record or dataset. [6], [8], [11], [36], [39], [45]	✓	✓	$1 - \frac{\text{Total Unique Rows}}{Total\ rows}$ [8] ISO/IEC 25024 [39]	Deterministic/ Probabilistic matches [6]. Sorted neighborhood algorithm, standard duplicate elimination algorithm, and sorted blocks [11].	
Efficiency	Extent to which data are able to quickly meet the information needs for the task at hand. [39], [45]	X	✓	ISO/IEC 25024 [39]	Not mentioned in the studies.	
Frequency	Refers to data used for producing results related to future time slots (required frequencies). [39]	X	✓	ISO/IEC 25024 [39]	Not mentioned in the studies.	
Heterogeneity	Means whether the source of the data used in model development is from a single organization or multiple organizations. [43]	NA	✓	Not mentioned in the studies.	Not mentioned in the studies.	
Informative	Refers to Data presentation form that will enable data users to capture data quickly and easily navigate the data range. [7]	✓	X	Not mentioned in the studies.	Not mentioned in the studies.	

https://digitalcommons.aaru.edu.jo/fcij/vol6/iss1/3
DOI: http://doi.org/10.54623/fue.fcij.6.1.3

Table 1. (Continue)

Dimension	Definition	TD	BD	Metrics (Measurements)	Solving Technique	
					SD	BD
Integrity	To believe it free from defects. [7], [41]	✓	NA	Not mentioned in the studies.	Not mentioned in the studies.	
Interpretability	The extent to which data is in appropriate languages, symbols and units, and the definitions are clear. [7], [10], [12], [45]	✓	X	Not mentioned in the studies.	Not mentioned in the studies.	
Minimality	Refer to the capability of representing the reality of interest with the minimal use of informative resources. [42]	X	✓	Not mentioned in the studies.	Not mentioned in the studies.	
Noise	Means erroneous data or incorrect data. [43]	X	✓	By classification algorithms like c4.5 [43]	Not mentioned in the studies.	
Objectivity	The extent to which data is unbiased, unprejudiced and impartial. [7], [10], [12], [45]	✓	NA	Not mentioned in the studies.	Not mentioned in the studies.	
Optimal Use of Resources	Refers to Efficient use of existing resources for data collection and processing. [7]	✓	NA	Not mentioned in the studies.	Not mentioned in the studies.	
Pertinence	Refers to the capability of representing all and only the relevant aspects of the reality of interest. [42]	X	✓	Not mentioned in the studies.	Not mentioned in the studies.	
Portability	The extent to which data can be expressed using similar data types and with the same amount of precision that allow data to be portable and can be moved. [39]	X	✓	ISO/IEC 25024 [39]	Not mentioned in the studies.	

Table 1. (Continue)

Dimension	Definition	TD	BD	Metrics (Measurements)	Solving Technique	
					SD	BD
Precision	The degree to which the values of an attribute are close to each other. [36]	X	✓	Calculated by considering the mean and the standard deviation of all the values of the considered attribute. [36]	Not mentioned in the studies.	
Readability	Refer to ease of understanding of data by users. [41], [42]	X	✓	Not mentioned in the studies.	Not mentioned in the studies.	
Recoverability	The extent to which data is easily recoverable. [39]	X	✓	ISO/IEC 25024 [39]	Not mentioned in the studies.	
Relevance	The extent to which data is applicable and relevant for the task at hand. [5], [7], [9]–[12], [39], [45]	✓	✓	ISO/IEC 25024 [39]	Filter, Hybrid, Embedded and Wrapper Approach [11][15] Fast Correlation-Based Filter algorithm [13]. Best First Search Algorithm (BFS), CfsSubset Evaluator (CSER), Chi-Squared Attribute Evaluator (CSAER), Information Gain Attribute Evaluator (IGAER) and Relief Attribute Evaluator (RAER) [14].	
Reliability	The extent to which data is sufficiently complete and error free to be convincing for its purpose and context. In addition to being reliable, data must also meet other tests for evidence. [45]	✓	✓	Not mentioned in the studies.	Not mentioned in the studies.	

https://digitalcommons.aaru.edu.jo/fcij/vol6/iss1/3
DOI: http://doi.org/10.54623/fue.fcij.6.1.3

Table 1. (Continue)

Dimension	Definition	TD	BD	Metrics (Measurements)	Solving Technique	
					SD	BD
Representational Format	The extent to which data is compactly represented in the same format.[5], [9], [45]	✓	✓	Not mentioned in the studies.	Not mentioned in the studies.	
Reputation	The extent to which data is highly regarded in terms of its source or content. [5], [10], [12], [45]	✓	X	Not mentioned in the studies.	Not mentioned in the studies.	
Scalability	Refers to how well big data are structured, designed, collected, generated, stored, and managed to support large-scale services in data achieving, access, transport, migration, and analytics. [47]	NA	✓	Not mentioned in the studies.	Not mentioned in the studies.	
Security & Privacy	The extent to which access to data is restricted appropriately to maintain its security. [5], [10], [45]	✓	✓	Not mentioned in the studies.	Not mentioned in the studies.	
Semantically Accurate	The extent to which data represent real entities in the context of big data. [39]	X	✓	ISO/IEC 25024 [39]	Not mentioned in the studies.	
Semantically Consistent	Describes rules that explain mandatory relationships between fields. [8]	✓	NA	$\frac{\text{Total Semantically Consistent Rows}}{Total\ number\ of\ rows}$ [8]	Not mentioned in the studies.	
Semantically Interoperable	The extent to which data is understandable and free of inconsistencies. [39]	X	✓	ISO/IEC 25024 [39]	Not mentioned in the studies.	
Simplicity	Refer to ease of understanding of data by users. [42]	X	✓	Not mentioned in the studies.	Not mentioned in the studies.	
Statistical Disclosure Control	Refers to Confidentiality of the information provided by respondents. [7]	✓	X	Not mentioned in the studies.	Not mentioned in the studies.	

Table 1. (Continue)

Dimension	Definition	TD	BD	Metrics (Measurements)	Solving Technique	
					SD	BD
Structurally Consistent	The extent to which a value falling within the expressed ranges stated in the accompanying static metadata file. It describes the structure of the values in the data. [8]	✓	X	$\frac{\text{Total Structurally Consistent Values}}{\text{Total Values}}$ [8]	Not mentioned in the studies.	
Time-Concurrent	Refers to the facts happened in similar or appropriate time slot. [39]	X	✓	ISO/IEC 25024 [39]	Not mentioned in the studies.	
Time-Consistent	Refers to data that shouldn't include any incoherence related to the represented time (e.g. impossible dates, disordered events). [39]	X	✓	ISO/IEC 25024 [39]	Not mentioned in the studies.	
Timeliness	The extent to which data is sufficiently up-to-date for the task at hand. [5]–[7], [9], [10], [32], [36], [41], [45]	✓	✓	Evaluated by timestamp related to the last update (currency) and the average validity of the data (volatility) [36] Qtimeliness (w,A) = e (-decline (A).age(w,A)) [9] $1 - \frac{\text{Currency}}{\text{Volatility}}$ [32]	Not mentioned in the studies.	
Timely Updated	Data must be properly updated for the task at hand. So data will has a convenient age for analysis. [39]	X	✓	ISO/IEC 25024 [39]	Not mentioned in the studies.	
Traceability	The extent to which data provide an audit trail that allow to trace the access and changes. [39]	X	✓	ISO/IEC 25024 [39]	Not mentioned in the studies.	
Understandability	The extent to which data is easily comprehended. [5], [10], [12], [45]	✓	✓	Not mentioned in the studies.	Not mentioned in the studies.	

https://digitalcommons.aaru.edu.jo/fcij/vol6/iss1/3
DOI: http://doi.org/10.54623/fue.fcij.6.1.3

Table 1. (Continue)

Dimension	Definition	TD	BD	Metrics (Measurements)	Solving Technique	
					SD	BD
Usability	To extent to which data is clear and easily used. [45], [46]	X	✓	Not mentioned in the studies.	Not mentioned in the studies.	
Usefulness	The extent to which the data provides any benefit or value. [6], [45]	✓	X	Data adaptability or its capacity to include new data items. [6]	Not mentioned in the studies.	
Utility	Refers to Data users demand to the data. [7]	✓	X	Not mentioned in the studies.	Not mentioned in the studies.	
Validity	Refers to the proportion of cases in a dataset with a given characteristic, which truly have the attribute. Lack of validity is referred to a bias or systematic error. [6]	✓	X	kappa coefficient. [6]	Not mentioned in the studies.	
Value-Added	The extent to which information is beneficial and provides advantages from its use. [10], [12], [45]	✓	X	Not mentioned in the studies.	Not mentioned in the studies.	

Note: The right mark means that the dimension can be applied and mentioned in the current studies, the wrong mark means that the dimension can be applied, however no one mentions it. While NA means that the dimension is not applicable. We refer to Traditional Data as (TD) and Big Data as (BD)

Many data quality dimensions can be applied to both traditional data and big data. While few number of these dimensions can't be applied, such as Heterogeneity, Scalability, and Complexity can't be applied on structured data due to its nature. While on the other side Objectivity, Comparability and optimal use of resources aren't applicable on big data. Also we can point out that data quality dimensions are complementary which means that two or more dimensions can support the same idea. Like Accessibility and Availability, also Coherence, Cohesion and Pertinence. Few number of metrics and improvement techniques are presented during current studies.

5. CONCLUSION

The importance of achieving and maintaining a high standard of data quality is highly needed in any business. Traditional data quality dimensions have gained tremendous attraction in the past few years. A wide range of data quality dimensions can be easily applied on traditional structured data due to its simple nature. But the studies for big data quality dimensions are still initiatives. Only few number of the current works studied how to enhance the big data quality, and this enhancement process was applied on a very common dimensions like completeness, while the rest of the dimensions are still underground. Furthermore, other papers only introduced a conceptual data quality frameworks or a framework to measure the level of data quality in big data projects. More attention should be paid to data quality dimensions that are applicable and matched with big data characteristics. Also, the big data quality metrics and their solving techniques should be under focus as well. Through our paper we investigated the current works and introduced a complete definition for data quality dimensions, its related metrics and handling techniques that can be applied to traditional data or big data or both. We conclude that many data quality dimensions are complementary and can be applied to traditional and big data. Furthermore, the metrics and techniques can be also used for both types of data. While data quality metrics that are domain specific (e.g. health domain) may differ from the common used ones.

REFERENCES

[1] A. Haug, F. Zachariassen, and D. Van Liempd, "The costs of poor data quality," J. Ind. Eng. Manag., vol. 4, no. 2, pp. 168–193, 2013.

[2] R. W. Wang and D. M. Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers.," J. Manag. Inf. Syst., vol. 12, no. 4, p. 5, 1996.

[3] R. Vaziri, M. Mohsenzadeh, and J. Habibi, "TBDQ: A pragmatic task-based method to data quality assessment and improvement," PLoS One, vol. 11, pp. 1–30, 2016.

[4] I. Lee, "Big data: Dimensions, evolution, impacts, and challenges," Bus. Horiz., vol. 60, no. 2017, pp. 293–303, 2017.

[5] J. Alipour and M. Ahmadi, "Dimensions and assessment methods of data quality in health information

https://digitalcommons.aaru.edu.jo/fcij/vol6/iss1/3
DOI: http://doi.org/10.54623/fue.fcij.6.1.3

systems," Acta Medica Mediterr., vol. 33, no. March, pp. 313–320, 2017.

[6] Y. Kodra, M. Posada, D. Paz, A. Coi, M. Santoro, F. Bianchi, F. Ahmed, Y. R. Rubinstein, and J. Weinbach, "Data Quality in Rare Diseases Registries," in Rare Diseases Epidemiology: Update and Overview, Springer Cham, 2017, pp. 149–164.

[7] T. Romanian and E. Journal, "Data Quality Dimensions to Ensure Optimal Data Quality," Rom. Econ. J., no. 63, pp. 89–103, 2017.

[8] N. Micic, D. Neagu, F. Campean, and E. H. Zadeh, "Towards a Data Quality Framework for Heterogeneous Data," in Proceedings - 2017 IEEE International Conference on Internet of Things, IEEE Green Computing and Communications, IEEE Cyber, Physical and Social Computing, IEEE Smart Data, iThings-GreenCom-CPSCom-SmartData 2017, 2018, vol. 2018-Janua, pp. 155–162.

[9] O. Azeroual, G. Saake, and J. Wastl, "Data measurement in research information systems: metrics for the evaluation of data quality," Scientometrics, vol. 115, pp. 1271–1290, 2018.

[10] N. Laranjeiro, S. N. Soydemir, and J. Bernardino, "A Survey on Data Quality : Classifying Poor Data," IEEE 21st Pacific Rim Int. Symp. Dependable Comput., 2015.

[11] M. Nasr, E. Shaaban, and M. I. Gabr, "Data Quality Dimensions," in Internet of Things—Applications and Future Conference, Springer, 2020, pp. 201–218.

[12] M. Z. Abdullah and R. A. Arshah, "Development of Data Quality Dimensions from User ' s Perspective Model," 2016.

[13] M. Rehman, "Role of FCBF Feature Selection in Educational Data Mining," Mehran Univ. Res. J. Eng. Technol., vol. 39, no. 4, pp. 772–778, 2020.

[14] C. Anuradha and T. Velmurugan, "Performance Evaluation of Feature Selection Algorithms in Educational Data Mining," Int. J. Data Min. Tech. Appl., vol. 5, no. 131, pp. 131–139, 2016.

[15] D. A. A. Gnana, "Literature Review on Feature Selection Methods for High-Dimensional Data," Int. J. Comput. Appl., vol. 136, no. 1, pp. 9–17, 2016.

[16] J. E. Hollander, A. J. Singer, S. Valentine, and M. C. Henry, "Wound Registry: Development and Validation," Ann. Emerg. Med., vol. 25, no. MAY, pp. 675–684, 1995.

[17] M. Topp, J. Langhoff-roos, and P. Uldall, "of a Cerebral Palsy Register," J. Clin. Epidemiol., vol. 50, no. 9, pp. 1017–1023, 1997.

[18] S. Paediatric, "A Method for the Validation of Data in a Register," Public Heal. Servier, vol. 100, no. 5, pp. 316–324, 1986.

[19] R. G. Skeet, "Cancer registration: principles and methods. Quality and quality control.," IARC Sci. Publ. 95, pp. 101–107., 1991.

[20] F. X. Liu, P. Rutherford, K. Smoyer-tomic, S. Prichard, and S. Laplante, "A global overview of renal registries : a systematic review," BMC Nephrol 1631, pp. 1–10, 2015.

[21] C. Couchoud, M. Lassalle, R. Cornet, and K. J. Jager, "Renal replacement therapy registries-time for a structured data quality evaluation programme," Nephrol. Dial. Transplant., vol. 28, no. February, pp. 2215–2220, 2013.

[22] Minaksh, D. R. Vohra, and Gimpy, "Missing Value Imputation in Multi Attribute Data Set," Int. J. Comput.

Published by Arab Journals Platform, 2021

anf Inf. Technol., vol. 5, no. 4, pp. 5315–5321, 2014.

[23] E. G. Armitage, "Missing value imputation strategies for metabolomics data," Etectrophoresis, vol. 36, pp. 3050–3060, 2015.

[24] A. Skandar, M. Rehman, and M. Anjum, "An Efficient Duplication Record Detection Algorithm for Data Cleansing," Int. J. Comuter Appl., vol. 127, no. 6, pp. 28–37, 2015.

[25] G. V. Dhivyabharathi and S. Kumaresan, "A survey on duplicate record detection in real world data," ICACCS 2016 - 3rd Int. Conf. Adv. Comput. Commun. Syst. IEEE, vol. 1, pp. 1–5, 2016.

[26] H. Zhao, J. Chen, Y. Liu, Q. Shi, Y. Yang, and C. Zheng, "Procedia Engineering The use of feature selection based data mining methods in biomarkers identification of disease," 2011.

[27] Z. Hu, Y. Bao, T. Xiong, and R. Chiong, "Hybrid filter – wrapper feature selection for short-term load forecasting," Eng. Appl. Artif. Intell., vol. 40, pp. 17–27, 2015.

[28] D. Tomar and S. Agarwal, "Twin Support Vector Machine Approach for Diagnosing Breast Cancer , Hepatitis , and Diabetes," vol. 2015, 2015.

[29] A. Gopalakrishnan, C. Graterol, and M. B. Love, "A Multifaceted Data Mining Approach to Understanding what Factors Lead College Students to Persist and Graduate," in Computing Conference ,IEEE, 2017, no. July, pp. 372–381.

[30] K. S. Sahedani and P. B. S. Reddy, "A Review : Mining Educational Data to Forecast Failure of Engineering Students," Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 3, no. April, pp. 628–635, 2013.

[31] A. Acharya and D. Sinha, "Application of Feature Selection Methods in Educational Data Mining," Int. J. Comput. Appl., vol. 103, no. 2, pp. 34–38, 2014.

[32] M. A. Serhani, H. T. El Kassabi, I. Taleb, and A. Nujum, "An Hybrid Approach to Quality Evaluation Across Big Data Value Chain," in IEEE International Congress on Big Data, 2016.

[33] I. Taleb and M. A. Serhani, "Big Data Pre-Processing: Closing the Data Quality Enforcement Loop," Proc. - 2017 IEEE 6th Int. Congr. Big Data, BigData Congr. 2017, no. 1, pp. 498–501, 2017.

[34] I. Taleb, H. T. El 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, 2016.

[35] S. Juddoo, "Overview of data quality challenges in the context of Big Data," in 2015 International Conference on Computing, Communication and Security, ICCCS 2015, 2015.

[36] D. Ardagna, C. Cappiello, W. Samá, and M. Vitali, "Context-aware data quality assessment for big data," Futur. Gener. Comput. Syst., vol. 89, pp. 548–562, 2018.

[37] R. L. VolkerSchwieger, "ModelingDataQualityUsingArtificial Neural Networks," Int. Assoc. Geod. Symp. Springer Int. Publ. Switz., vol. 140, pp. 3–8, 2015.

[38] S. Juddoo and C. George, "Discovering Most Important Data Quality Dimensions Using Latent

https://digitalcommons.aaru.edu.jo/fcij/vol6/iss1/3
DOI: http://doi.org/10.54623/fue.fcij.6.1.3

Semantic Analysis," in 2018 International Conference on Advances in Big Data, Computing and Data Communication Systems, icABCD 2018, 2018, pp. 1–6.

[39] J. Merino, I. Caballero, B. Rivas, M. Serrano, and M. Piattini, "A Data Quality in Use model for Big Data," Futur. Gener. Comput. Syst., vol. 63, pp. 123–130, 2016.

[40] M. Abdallah, "Big Data Quality Challenges," Proc. 2019 Int. Conf. Big Data Comput. Intell. ICBDCI 2019, pp. 1–3, 2019.

[41] L. Cai and Y. Zhu, "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era," data Sci. J., vol. 14, no. 2, pp. 1–10, 2015.

[42] C. Batini, A. Rula, M. Scannapieco, and G. Viscusi, "From data quality to big data quality," J. Database Manag., vol. 26, pp. 60–82, 2016.

[43] M. F. Bosu and S. G. Macdonell, "Experience: Quality benchmarking of datasets used in software effort estimation," J. Data Inf. Qual., vol. 11, no. 4, pp. 1–38, 2019.

[44] M. Bovee, R. P. Srivastava, and B. Mak, "A Conceptual Framework and Belief- Function Approach to Assessing Overall Information Quality," Int. J. Intell. Syst., vol. 18, pp. 51–74, 2003.

[45] F. Sidi, P. Hassany, S. Panahy, L. S. Affendey, M. A. Jabar, H. Ibrahim, and A. Mustapha, "Data Quality : A Survey of Data Quality Dimensions," in Information Retrieval & Knowledge Management (CAMP), 2012 International Conference ,IEEE, 2012, pp. 300–304.

[46] J. Gao, C. Xie, and C. Tao, "Big Data Validation and Quality Assurance – Issuses , Challenges , and Needs," IEEE Symp. Serv. Syst. Eng., pp. 433–441, 2016.

[47] S. G. Stockman, A. R. Todd, and G. A. Robinson, "A Framework for Software Quality Measurement," IEEE J. Sel. Areas Commun., vol. 8, no. 2, pp. 224–233, 1990.

[48] F. Sidi, P. H. Shariat Panahy, L. S. Affendey, M. a. Jabar, H. Ibrahim, and A. Mustapha, "Data quality: A survey of data quality dimensions," Proc. - 2012 Int. Conf. Inf. Retr. Knowl. Manag. CAMP'12, no. August, pp. 300–304, 2012.

Published by Arab Journals Platform, 2021