A survey on data quality: principles, taxonomies and comparison of approaches.

Mehdi YALAOUI
dept. of computer science lab. LSI
Univercity of Science and Technology
Houari Boumedien (USTHB)
Algiers, Algeria
m.yalaoui@usthb.dz

Saida BOUKHEDOUMA
dept. of computer science lab. LSI
Univercity of Science and Technology
Houari Boumedien (USTHB)
Algiers, Algeria
s.boukhedouma@usthb.dz

2021 International Conference on Information Systems and Advanced Technologies (ICISAT) | 978-1-6654-7824-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICISAT54145.2021.9678209

***Abstract*—Nowadays, data generation keeps increasing exponentially due to the emergence of the Internet of Things (IoT) and Big data technologies. The manipulation of such Big amount of data becomes more and more difficult because of its size and its variety. For better governance of organizations (decision making, data analysis, earnings increase …), data quality and data governance at present of Big data are two major pillars for the design of any system handling data within the organization. This explains the number of researches conducted as it constitutes a research subject with several gaps and opportunities. Many works were conducted to define and standardize Data Quality (DQ) and its dimensions, others were directed to design and propose data quality assessment and improvement models or frameworks. This work aims to recall the data quality principles starting by the needed background knowledge, then identify and compare the relevant taxonomies existing in the literature, next surveys and compares the available Data quality assessment and improvement approaches. After that, we propose a metamodel highlighting the main concepts of DQ assessment and we describe a generic process for DQ assessment and improvement. Finally, we evoke the main challenges in the field of DQ before and after the emergence of Big Data.**

Keywords*— *Data Quality, Big Data, Quality Dimensions, Quality Metrics, Metamodel, Assessment process, Improvement.

I. Introduction

The number of data in the world must reach 150 zettabytes (150 trillion gigabytes) in 2024 and was estimated about 59 zettabytes in 2020, which represents a huge volume of data but only 37% of this data has the potential to be analyzed, whereas according to studies in the IT field, 97.2% of organizations invest in Big data and artificial intelligence, which shows the importance of ensuring the quality of data used within the organization and find the necessary methods to govern them. Thus, the poor quality (or non-quality) of the data can cause considerable damage within the organization for bad management (bad decision), ignorance of customers and collaborators, and a loss of income, according to a study conducted by IBM in 2019, the U.S. government is losing approximately $ 1.3 trillion annually due to poor data quality [33].

A problem within data quality certainly creates great difficulties in the conduct and management of data-driven projects and a huge loss of income *[53]*. Nowadays large organizations are mainly based on the data collected in their decision making (proposal of new products or offers, acquisition of new markets ...). Therefore, a defect in the data constituting the statistics and the analysis reports, carried out towards making bad and poor financial, operational, and/or technical decisions.

Analyzing Big data is very beneficial for the organization and allows leaders to make the best decision, but manipulating such an amount of data becomes more and more difficult and shows every time a set of challenges that researchers are trying to solve and deal with. Ensuring the quality of this Big data is one of the most important challenges. Thus, adapting the data quality process to Big data means dealing with the particularities of Big data which are mainly concretized in the famous **3V:** *Variety, Velocity* and *Volume [10]*.

Another important challenge is the measurement of DQ in Big data context, as mentioned previously the metric used to measure the DQ of a system for the majority of dimensions uses the number of correct and incorrect data to get the quality rate. But, in case of Big data, this quantification is almost difficult when having such amount and diversity of data that's why researchers need to look for other quantifying ways for such metrics.

This paper aims to present the Data Quality principles and to parse and compare the existing literature dimensions' taxonomies; it provides also a review of Data Quality assessment/improvement frameworks, models, and methods. A comparison conducted is based on quality dimensions used and whatever the proposed framework (or model or method) is related to Data Quality assessment or improvement before or after the emergence of Big Data. Finally, some open questions are identified to give research orientations and interesting axes to be explored.

The rest of the paper is structured like follows: Section II presents the background and the definitions of the used concepts (data, big data, data quality …). In Section III, we expose the most relevant dimension's taxonomies existing in the literature and we categorize them under two main classes: *Data-relevant* or *System-relevant*. Section IV surveys and compares existing models/methods and frameworks for Data Quality assessment and improvement, before and after the appearance of Big Data. In Section V, we describe our metamodel for DQ assessment and the main steps of the process of DQ assessment and improvement. Section VI exposes Data Quality's major problems and challenges. Finally, we conclude this paper and talk about our future perspectives.

II. Background and Definitions

A. General Background

Before delving into the topic and conduct this literature survey, let us start with a set of concepts definitions to be in the same wavelength for the following sections. The related concepts are *Data Quality* **(DQ)** which the main subject of this paper, *dimensions*, and *metrics* that will be used to

978-1-6654-7824-3/21/$31.00 ©2021 IEEE

evaluate the quality of a system, then we will continue by defining the *Big data* and *Cloud* concepts where Data Quality is one of the most important parts to care about.

But first of all, let's talk about ***data*, *information* and *knowledge*** from the Information System **(IS)** point of view.

Data in IS are the basic elements for designing or representing information. This data can be either internal (information about the company, product, service ...) or external (outside from the company's IS or concerns people/objects external to the organization) [*18*]. Data comes from different sources (internal or external source*)*; it is represented in different forms: *unstructured*: digital/ flat data (files, photos, videos, etc.) which requires an FMS (File Management System), *semi-structured* (xml, JSON …) or *structured*. For this latter, we may have two categories: primitive data and non-primitive data. To be able to understand and read this data, additional information is provided describing this data, their type and signification, it is called meta-data which means "*data about data*".

Thus, **information** is constituted from a set of data processed, organized, structured, and presented in a well-defined *context;* it is usefull to respond to a given question asked *(request, decision-making, report…)*. For example, a price of an object in an order list is considered as data while the total of this order or the average price of all objects is a piece of information.

On the other hand, the association of relevant information in a defined domain constitutes **knowledge** in this domain which must always be *true* if all information constituting it are met *[19]*.

B. Big Data

Big data designates the specialty or the work which consists in processing, analyzing, extracting, or even manipulating very large data sets such that their manipulation using traditional methods (computer software, database management tools) is become very complex *[52]*.

According to the Gartner Group, Big Data is a set of tools that responds to a triple problem known as the 3V rule: *Volume*, *Variety* and *Velocity*. Volume represents the huge volume of data to be processed, Variety refers to data for different domains and different types, and Velocity which represents the frequency of creation, collection, and sharing of data. [47], these Big data characteristics are related to the multi-sourcing collection and data types variety.

In other studies, such as [8], the authors talk about 4V by adding a *Veracity* rule which means compliance, consistency, and reliability of data with the context and needs of the organization.

But [8] in their work succeeded in extending the initial set of rules composing Big Data (3V then 4V) to a set of 7V by adding three other rules which are: The *Validity* which for them seems similar to Veracity but it covers the aspect of accuracy and precision of the data concerning the intended use. Another rule added is *Volatility* to cover the retention side (storage time) of data. The last rule added qualified as a special V is the *Value*, in fact, unlike the other Vs, this V (Value) is the desired result of Big data processing.

Fig. 1. Represents our proposed data meta-model in context of Big data, this meta-model summarizes the main concepts and highlights the dependencies/relations between them.

C. Data Quality

1) Definitions

Several definitions have been assigned for data quality which converges to describe the same concept, the following are the most relevant ones:

"*data quality is the mix of six core dimensions: accuracy, completeness, consistency, currency, security, and reliability*" *[3]*.

"data quality is the degree to which a set of inherent characteristics of an object fulfills requirements" [1] [2].

"data quality means the conformance to requirements and the fitness for use" which means that the data quality can differ from one context to another and must be dependent on the actual use case [20] [21].

2) Dimensions and metrics

Dimensions are the criteria that allow DQ evaluation, *[17]* defined the data quality dimensions as a set of data quality attributes that represent a single aspect or construct of data quality, these dimensions are defined based on the client subjective opinion "I need the data to be right, and I need to be able to cross-reference systems A and B" which is translated into objective criteria *[3]*.

The evaluation of these dimensions passes through **metrics** defined to have values between 0 and 1 to represent the lowest to highest data quality *[52] [4]*. The community defined a very important number of dimensions that can be used to evaluate a DQ system, most of them are defined and shortly explained, bellow.

The common metric used to measure most of these dimensions like accuracy, completeness and, consistency is:

$$\Delta = 1 - (N_\iota / N_\tau)$$

When **D** represents the metric associated with the dimension, $\mathbf{N_i}$ number of incorrect values and $\mathbf{N_t}$ the total number of values *[7]*.

The following table illustrates the way accuracy, completeness, and consistency are calculated:

DQ Dimensions	**Metric functions**
Accuracy	Acc = 1 - (Ncv / N)
Completeness	Comp = 1 - (Nmv / N)
Consistency	Cons = 1 - (Nvrc / N)

TABLE I. DIMENSIONS METRICS

Where **Ncv** is the number of correct values, **Nmv** represents the number of missing values, **Nvrc** Number of values that respects the contains and **N** concerns the total number of values.

Accuracy: is the degree to which the data or information describes the object or event in question. In other words, it is the degree of proximity to which α value has described a β value considered as correct for an attribute of an entity E; if α = β we say that the information / the data (α) is correct or accurate. The accuracy can be reflected in the use of metrics or classifications used to specify an attribute domain for example to measure the length the use of a centimeter as a unit of measurement is more accurate than the use of a kilometer *[5]*.

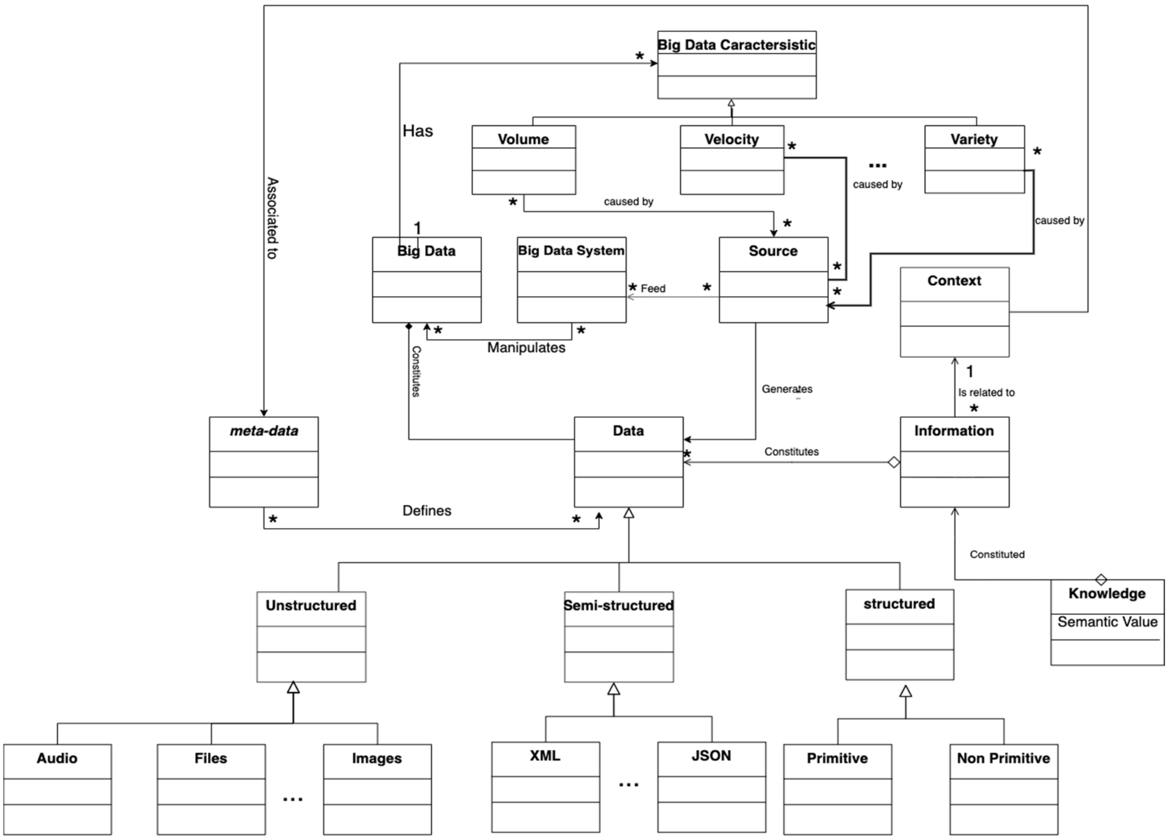


Fig. 1. Data meta-model in context of Big Data.

Completeness: Information is considered complete when it meets the expectations of all attributes and entities. This makes it possible to verify the existence of missing values that negatively affect the correct understanding/use of the information *[13]*

Consistency: It is when the same information must be present in several places inside an IS, if this information matches then it is consistent. This can be illustrated by the information of an employee who exists in two different databases (HR department and payroll department) in the same company IS, if the employee resigns (deleted or deactivated in the service DB HR) but he continues to receive his salary then the information is inconsistence *[6]*.

Currency: Means that the information must be available exactly at the moment when it is needed.

Security: The information needs to be available only for the users who are allowed to access it [13].

Reliability: it is one of the most important dimensions for measuring the DQ, some researchers like Hoare (1975) consider it a synonym of the overall quality of the data, others like Chapple (1976) consider it as synonymous with accuracy, ANSYASQC Standard A3-1978 (1978) qualifies it as a measure of the probability that a datum is used to perform a certain function under certain conditions in a period time [22].

Uniqueness: this is to ensure that the information is unique in the system or the database to avoid duplication (for example, in a telecommunications company, Network elements can generate multiple call details for a single transaction so the billing system must make sure to remove any duplication).

Validity: This dimension refers to the compliance of the information the business rules that describe it (the age of a person must be Integer ...).

III. Data Dimensions Taxonomies

In the DQ field, the researchers always tried to categorize and classify DQ's dimensions in order to simplify the DQ evaluation and choose the right dimensions for each system from different points of view, that's why several classifications were proposed.

A. Classification of Wand et al. with Laranjeiro et al. adaptation

It was one of the first DQ classifications proposed in 1996, which was based on the definition of internal and external views, where **internal view** represents the organization internal generated data but from the other side the **external view** refers to the information that is given as a black box to be used inside the organization *(see Table II),* and for each view was indicated whether a dimension is related to the *data* or a *system* perspective (Fig. 2.).

The same authors proposed another classification of DQ dimensions into 4 categories *[15]* which are: The **intrinsic category** that includes dimensions expressing the natural quality of the data; the dimensions under the **contextual category** express the fact that data quality must be considered within a specific context; the **representational category** refers to dimensions that are related with the format and meaning of the data; and finally, the

accessibility category, refers to dimensions that express how data is accessible to users.

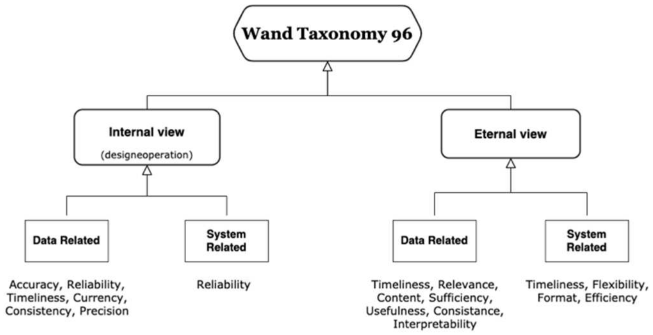

Wand Taxonomy 96
Internal view
(designeoperation)
Eternal view
Data Related
System Related
Data Related
System Related
Accuracy, Reliability, Timeliness, Currency, Consistency, Precision
Reliability
Timeliness, Relevance, Content, Sufficiency, Usefulness, Consistance, Interpretability
Timeliness, Flexibility, Format, Efficiency

Fig. 2. Wand et al. categorization schema

B. Classification of Merinoet al.

Authors of [16] proposed a classification of DQ dimensions based on **inherent** and **system dependent** point of view, where **Inherent** refers to the degree to which quality characteristics of data have the intrinsic potential to satisfy stated and implied needs when data is used under specified conditions. **System dependent** refers to the degree to which Data Quality is reached and preserved within a computer system when data is used under specified conditions.

After the emergence of Big data, multiple classifications were proposed to match with its particularities (3Vs).

C. Classification of Caballero et al.

Caballero et al. *[10]* established a classification according to the **Adequacy** of the data to the purpose of the analysis to define three important DQ characteristics for the data within Big Data context which are:

Contextual Adequacy refers to the capability of datasets to be used within the same domain of interest of the analysis independently of any format (e.g., structured vs. unstructured), any size or the velocity of inflow.

Temporary Adequacy refers to data within an appropriate time slot for the analysis (e.g., similar age, or throughout a specific duration for historical data, or coetaneous data.

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.

Then, the authors adapted this classification to the specifications of Big Data (Volume, Variety, and Velocity).

According to *[7]* the mapping of these dimensions was only based on hypothesis, thus there is a need to conduct more researchers in the area of Big Data Quality to investigate the dimensions that are more relevant for the Big Data context.

TABLE. II is our adapted version of classification with adding the Big Data 5Vs specifications, as in many works the researchers talk about the famous 5Vs instead of the usual 3Vs or 7Vs.

	Velocity	Volume	Variety	Veracity	Value
Contextual Adequacy	***Completeness***	***Completeness*** ***Consistency*** ***Confidentiality***	***Accuracy*** ***Consistency*** ***Credibility*** ***Compliance*** ***Understandability*** ***Confidentiality***	***Accuracy*** ***Credibility*** ***Reliability*** ***Integrity***	***Credibility*** ***Understandability*** ***Integrity***
Temporal Adequacy	***Accuracy*** ***Correctness***	***Correctness***	***Consistency*** ***Correctness***	***Consistency***	***Accuracy***
Operational Adequacy	***Confidentiality*** ***Efficiency***	***Efficiency*** ***Existence***	***Accessibility*** ***Confidentiality*** ***Efficiency***	***Efficiency*** ***Interpretability***	***Precision***

TABLE II. ADAPTED CABALLERO'S TAXONOMY INCLUDING ADDITIONAL VS

Fig. 3 summarizes the major classifications existing in the literature with the related categories. The proposed categories inside each taxonomy are based on the data utilization context. However, in our vision, all categories are belonging to one of two major views: **Data-view** (which can decline into data-related, Intrinsic, Contextual, Temporal or Inherent) or **System-view** (which can decline in System-related, Accessibility, System dependent or Operational).

TABLE. III represents our classification adaptation of the famous 15 dimensions [30] inside each of the previous taxonomies. We completed the attributes of this table based on analyzing each taxonomy and using all inputs and selection criteria.

IV. METHODS, MODELS, AND FRAMEWORKS FOR DATA QUALITY

All researchers agree that having a reliable data system, generating reports based on data with good quality, and making strategic or operational decisions based on these reports is essential for organization management. They also agree that one of the major steps or tasks to ensure this data quality is the assessment and validation of the organization's data quality *[25]*. On the other hand, the assessment is not the only step to ensure the quality as it ensures only the Quality quantification, the other important task for ensuring DQ is DQ improvement *[32; 36]*. In fact, after assessing the quality and detecting all dimensions with poor quality, it is necessary to launch a DQ improvement process *[48; 49]*.

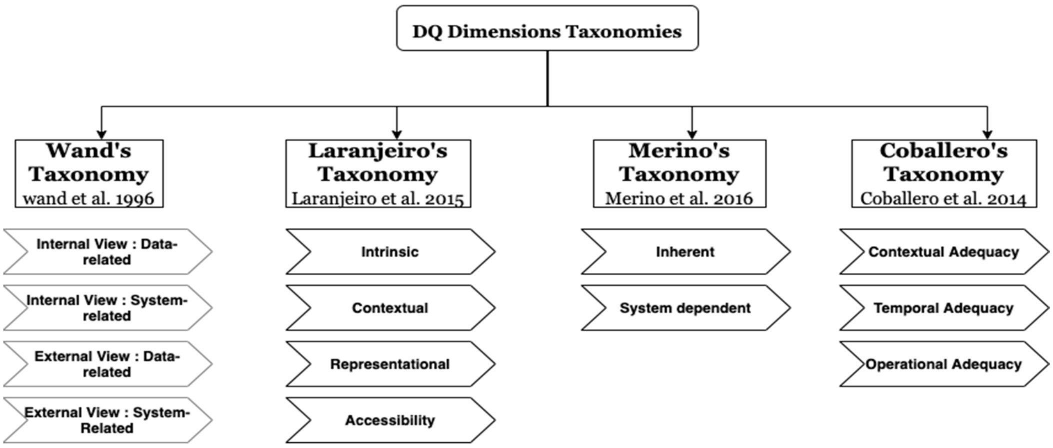
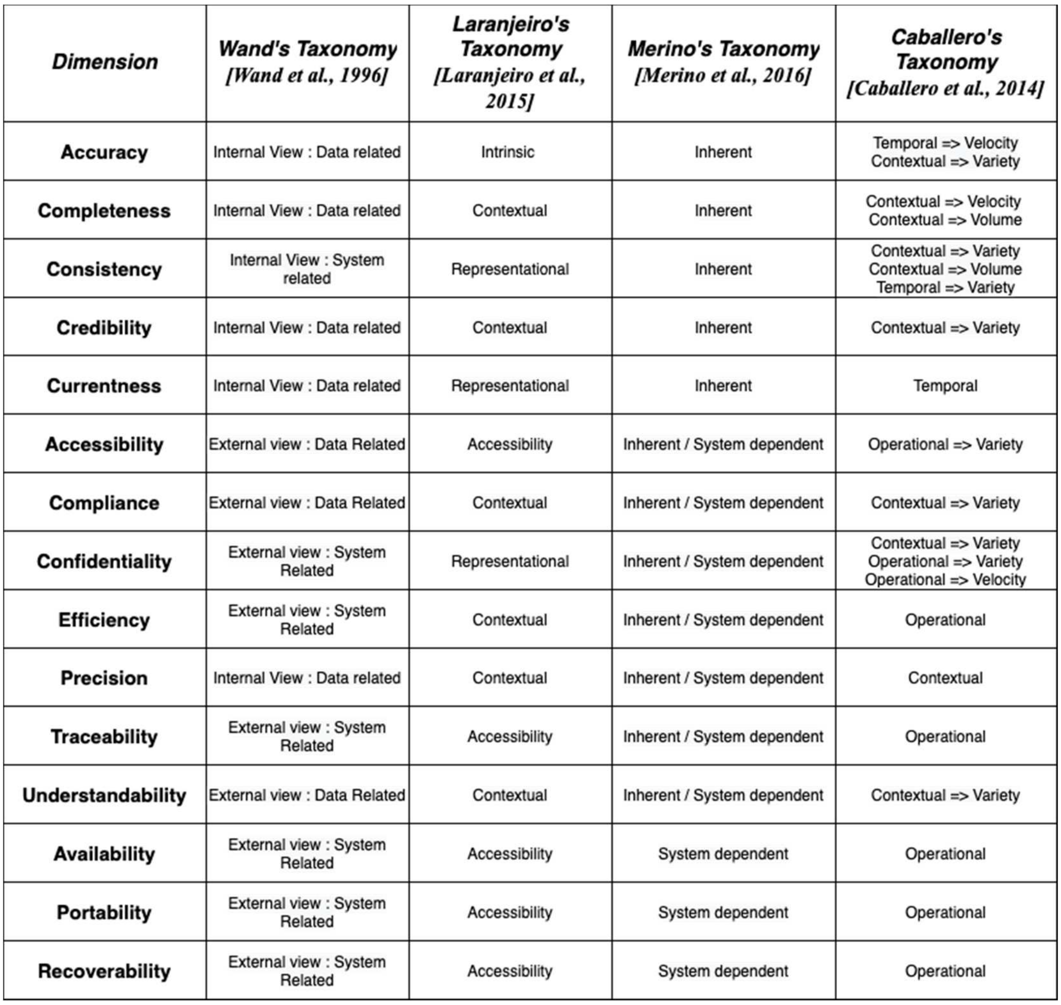


Fig. 3. Global view of relevant taxonomies of Data Quality



Dimension	*Wand's Taxonomy [Wand et al., 1996]*	*Laranjeiro's Taxonomy [Laranjeiro et al., 2015]*	*Merino's Taxonomy [Merino et al., 2016]*	*Caballero's Taxonomy [Caballero et al., 2014]*

Accuracy	Internal View : Data related	Intrinsic	Inherent	Temporal => Velocity
				Contextual => Variety

Completeness	Internal View : Data related	Contextual	Inherent	Contextual => Velocity
				Contextual => Volume

Consistency	Internal View : System related	Representational	Inherent	Contextual => Variety
				Contextual => Volume
				Temporal => Variety

Credibility	Internal View : Data related	Contextual	Inherent	Contextual => Variety
Currentness	Internal View : Data related	Representational	Inherent	Temporal
Accessibility	External view : Data Related	Accessibility	Inherent / System dependent	Operational => Variety
Compliance	External view : Data Related	Contextual	Inherent / System dependent	Contextual => Variety

Confidentiality	External view : System Related	Representational	Inherent / System dependent	Contextual => Variety
				Operational => Variety
				Operational => Velocity

Efficiency	External view : System Related	Contextual	Inherent / System dependent	Operational
Precision	Internal View : Data related	Contextual	Inherent / System dependent	Contextual
Traceability	External view : System Related	Accessibility	Inherent / System dependent	Operational
Understandability	External view : Data Related	Contextual	Inherent / System dependent	Contextual => Variety
Availability	External view : System Related	Accessibility	System dependent	Operational
Portability	External view : System Related	Accessibility	System dependent	Operational
Recoverability	External view : System Related	Accessibility	System dependent	Operational

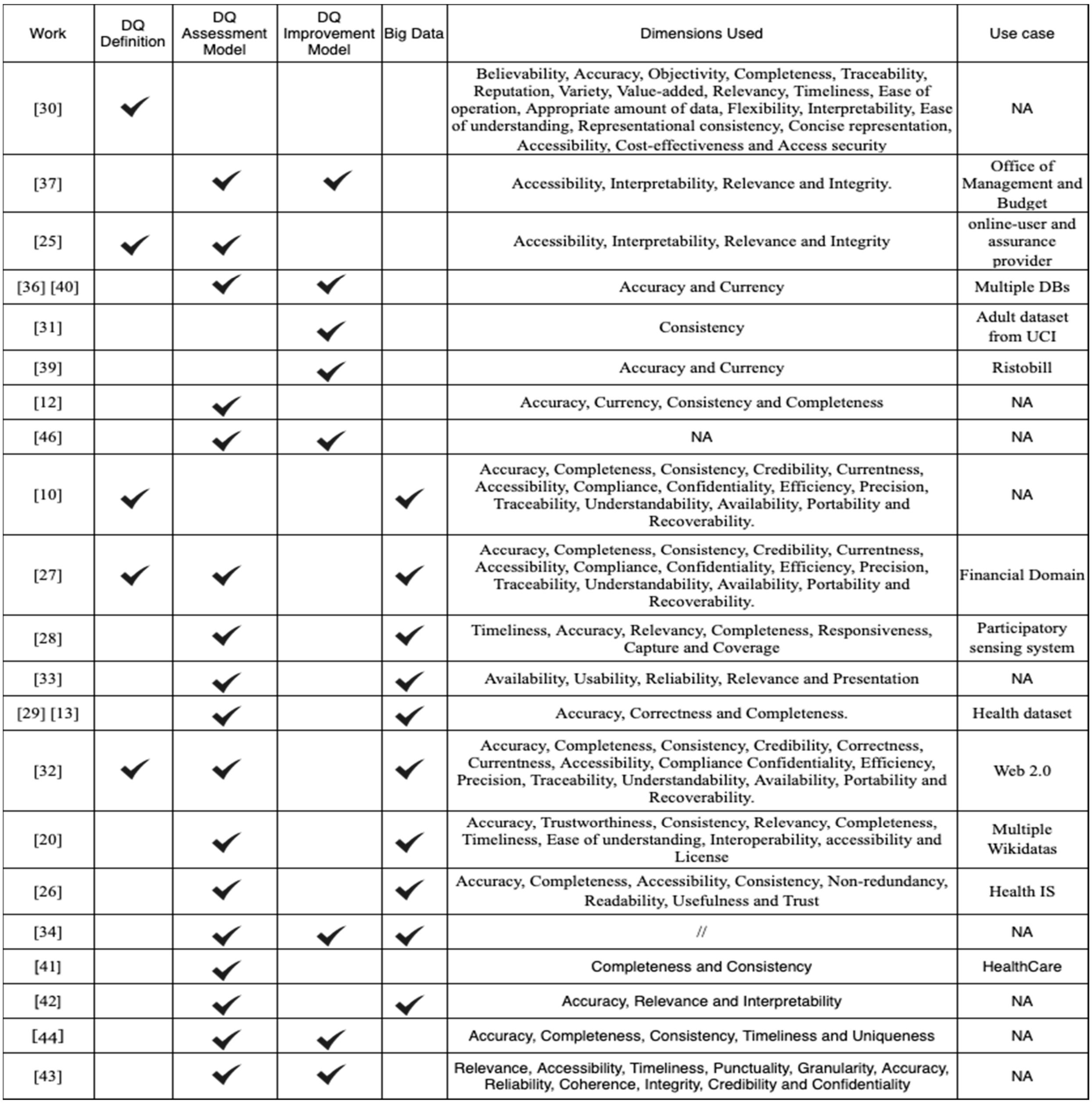
TABLE III. DIMENSIONS' CATEGORIZATION ACCORDING TO THE DIFFERENT TAXONOMIES

Thus, number of works were conducted in order to provide the community with a set of methods, models, and frameworks simplifying, defining, and ensuring the data quality assessment, evaluation, and improvement inside an organization's system *[50]*. With the appearance of Big data, this assessment becomes more and more complex and ensuring the system's data quality becomes harder, this explains all the new researches conducted for the DQ assessment in Big data era. However, the DQ improvement methods were not as much proposed as the main issue remains the Quality assessment and the previous proposed DQ improvement methods can remain the same in Big data.

Also, many works were provided to summarize and survey the existing models and frameworks. *[24]* surveyed and compared the defined twelve general-purpose applicable DQ frameworks that contain DQ definition, assessment, and improvement process, the twelve selected and studied frameworks were judged generally applicable in the most circumstance in practice. Authors of *[23]* described the quality factors used in evaluating the quality in Big data application, then they listed the commonly used framework for the Big data quality assessment and the challenges related to the Big data specifications (the famous Vs). Authors of *[22]* from their side, tried to analyze the literature and find out the list of DQ dimensions that are still valid in Big data context.

From our side, we reviewed a larger set of 23 works that dealt with DQ assessment before and after the appearance of Big Data. The related works are summarized and listed in the **TABLE IV** bellow.

The proposed approaches are compared according to six main criteria we have established: *DQ Definition* to show whatever the research gives a definition and principles of DQ, *DQ Assessment* for the works that proposed a DQ evaluation model or framework, *DQ Improvement* for approaches that including DQ improvement process, the *Big Data* criteria to distinguish the works that are adapted especially in case of Big Data, the list of used *dimensions* and finally the *use case* criteria used for validation.



Work	DQ Definition	DQ Assessment Model	DQ Improvement Model	Big Data	Dimensions Used	Use case
[30]	✓				Believability, Accuracy, Objectivity, Completeness, Traceability, Reputation, Variety, Value-added, Relevancy, Timeliness, Ease of operation, Appropriate amount of data, Flexibility, Interpretability, Ease of understanding, Representational consistency, Concise representation, Accessibility, Cost-effectiveness and Access security	NA
[37]		✓	✓		Accessibility, Interpretability, Relevance and Integrity.	Office of Management and Budget
[25]	✓	✓			Accessibility, Interpretability, Relevance and Integrity	online-user and assurance provider
[36] [40]		✓	✓		Accuracy and Currency	Multiple DBs
[31]			✓		Consistency	Adult dataset from UCI
[39]			✓		Accuracy and Currency	Ristobill
[12]		✓			Accuracy, Currency, Consistency and Completeness	NA
[46]		✓	✓		NA	NA
[10]	✓			✓	Accuracy, Completeness, Consistency, Credibility, Currentness, Accessibility, Compliance, Confidentiality, Efficiency, Precision, Traceability, Understandability, Availability, Portability and Recoverability.	NA
[27]	✓	✓		✓	Accuracy, Completeness, Consistency, Credibility, Currentness, Accessibility, Compliance, Confidentiality, Efficiency, Precision, Traceability, Understandability, Availability, Portability and Recoverability.	Financial Domain
[28]		✓		✓	Timeliness, Accuracy, Relevancy, Completeness, Responsiveness, Capture and Coverage	Participatory sensing system
[33]		✓		✓	Availability, Usability, Reliability, Relevance and Presentation	NA
[29] [13]		✓		✓	Accuracy, Correctness and Completeness.	Health dataset
[32]	✓	✓		✓	Accuracy, Completeness, Consistency, Credibility, Correctness, Currentness, Accessibility, Compliance Confidentiality, Efficiency, Precision, Traceability, Understandability, Availability, Portability and Recoverability.	Web 2.0
[20]		✓		✓	Accuracy, Trustworthiness, Consistency, Relevancy, Completeness, Timeliness, Ease of understanding, Interoperability, accessibility and License	Multiple Wikidatas
[26]		✓		✓	Accuracy, Completeness, Accessibility, Consistency, Non-redundancy, Readability, Usefulness and Trust	Health IS
[34]		✓	✓	✓	//	NA
[41]		✓			Completeness and Consistency	HealthCare
[42]		✓		✓	Accuracy, Relevance and Interpretability	NA
[44]		✓	✓		Accuracy, Completeness, Consistency, Timeliness and Uniqueness	NA
[43]		✓	✓		Relevance, Accessibility, Timeliness, Punctuality, Granularity, Accuracy, Reliability, Coherence, Integrity, Credibility and Confidentiality	NA

TABLE IV. COMPARISON OF THE MOST RELEVANT WORKS ON DATA QUALITY.

According to the above comparison, we can find that most of the works are split between two Data Quality subdomains, the Data Quality Assessment (DQA) and the Data Quality Improvement (DQI). However, after the appearance of Big data the works related to DQI were limited as the most important challenge is to deal with the DQA as the data volume and variety became uncontrollable. From the other side, we can extract a set of three relevant dimensions that we may find in most of the works and proposed methods. In fact, accuracy, completeness, and consistency are the dimensions that come back in each work and each proposed method or framework. Also, most of the existing use cases for approach validation are in the health care field due to its importance, variety and sensibility of the data quality. Although according to our vision, other fields can be considered relative to biological and geographical data, for example.

V. Data Quality Metamodel and Assessment/Improvement Process

A. DQ Assessment metamodel

We propose a data quality assessment meta-model (see ***Fig. 4***) based on comparing and grouping all assessment methods, models, and frameworks existing in the literature (***TABLE. III***) and adapting the different concepts figuring inside ISO / IEC 25025 meta-model [1]. This meta-model remains valid for the Big Data specifications.

This meta-model puts data at the heart of any relationship because it is the basic element that constitutes any system and ensuring the quality of these data is the object of this work. Business rules represent the major criteria that allow determining whatever the received data has a good quality or not according to the usage context, these rules can be divided into two important classes the semantic and syntactic rules. Dimensions definition represents the key for each DQ assessment model, thus, measuring the dimensions passes by evaluating the related metrics based on the defined business rules.

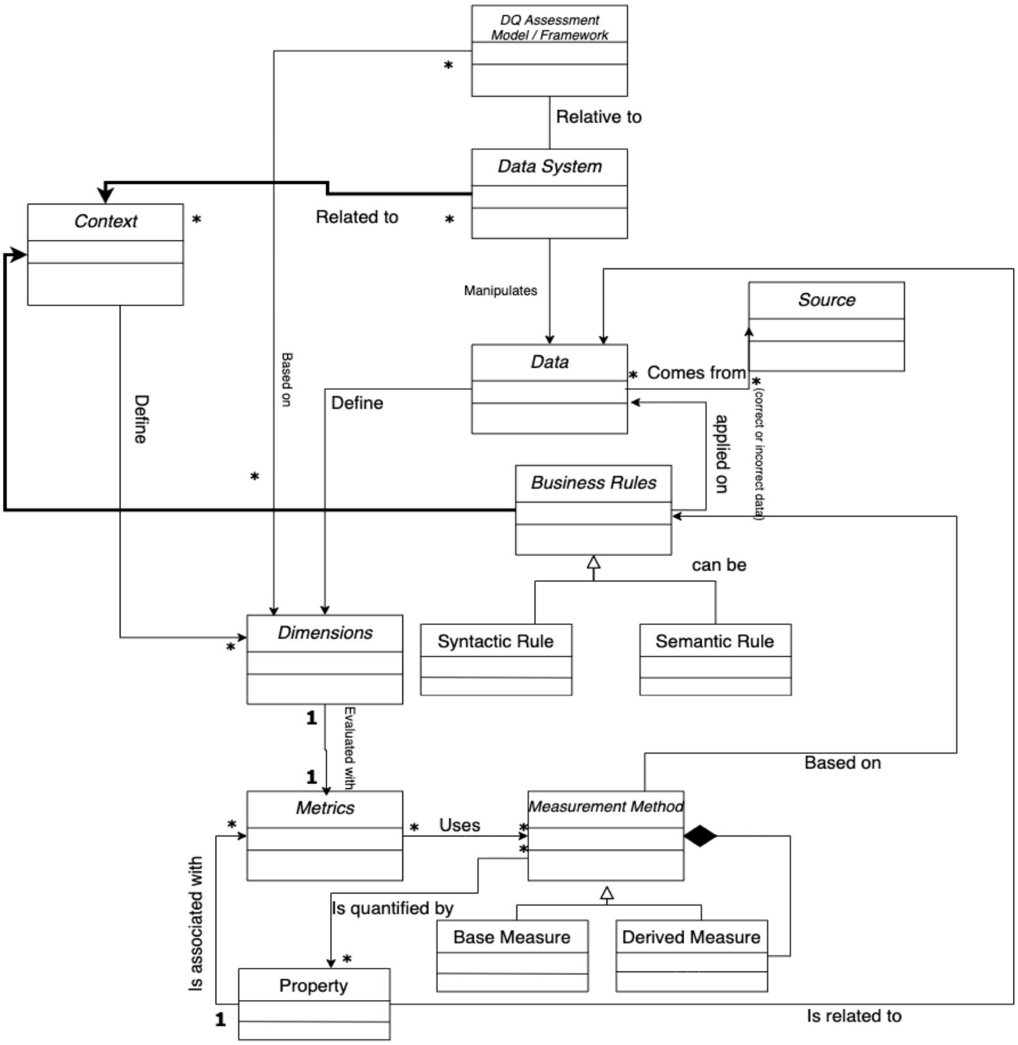
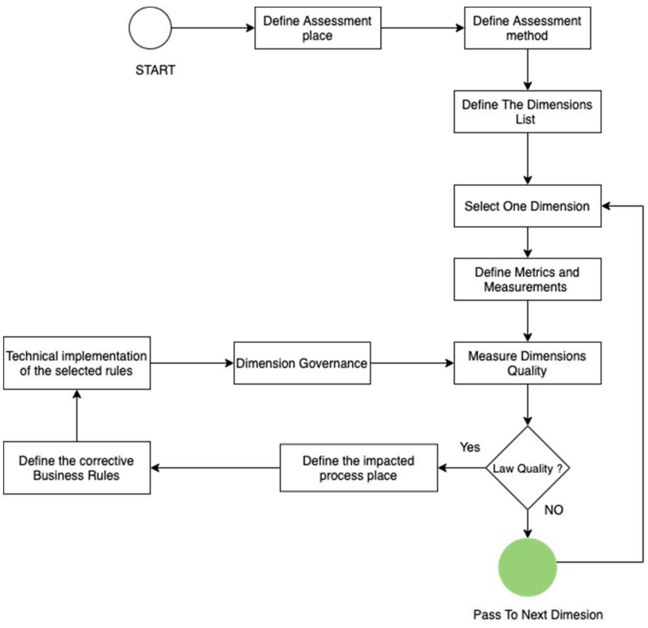


Fig. 4. Data Quality Assessment meta-model.

B. DQ Assessment and Improvement process

From another hand, we propose a generic DQ Assessment and Improvement process (see ***Fig. 5***). The process is composed of five steps described like follows:

- **DQ Assessment** (evaluating each single defined dimension) where the organization need to have a valid and reliable assessment system or model as the DQ needs to be measured many times during a single DQ Improvement process, during this step the assessment place needs to be defined inside the data flow, which means the place in which the users and analyzers expect a lack of quality.
- **Define Dimensions with poor Quality** as the assessment is based on data dimensions it is necessary to define those with low quality and at what level this quality is going down, once the issued dimensions are defined the
- define the **corrective Business rules** which will allow improving the dimension quality by decreasing the number of incorrect values (filtering, normalization …),
- **Rules implementation** on the system is required to deal with this incorrect values and finally,
- **Data Quality Dimension Governance** process is required to ensure that the income data is correct and all inconsistency was resolved.



START
Define Assessment place
Define Assessment method
Define The Dimensions List
Select One Dimension
Define Metrics and Measurements
Technical implementation of the selected rules
Dimension Governance
Measure Dimensions Quality
Yes
Define the corrective Business Rules
Define the impacted process place
Law Quality ?
NO
Pass To Next Dimesion

Fig. 5. Proposed Data Quality Assessment and Improvement Process.

VI. Data Quality Problems and Challenges

Several works in the literature, addressed the problems related to data quality. These works tried to categorize the problems and find out the dimensions that allow fixing them. *[13]* stated that data quality problems can be divided into two classes: **single-source** and **multi-source** problems, and each of them is split into two categories (Schema level and Instance level).

According to *[14]*, several factors or processes generate bad data: human data entry, sensors devices readings, social media, unstructured data, and missing values, they enumerate many reasons for poor data which affect its quality elements and its related dimensions (Accuracy, Completeness, and Consistency).

[15] from there side, proposed a larger list of issues vs DQ Dimensions with the source particularity when they categorized them into single and multiple sources, which is a mix between the two previous works *[14];[13].*

However, in context of Big Data, most of the exposed issues are systematically included under Big Data characteristics, for example, all multi-sources problems are included under the *variety* characteristic also duplicated data and the outdated data issues are included under the *veracity* and *volume* characteristics; these issues are related to the major steps of data lifecycle inside a Big data system (data multi-sourcing collection and data profiling and processing).

Thus, the main challenges of DQ in case of Big Data are the assessment, evaluation methods, and DQ improvement processes, as the data is collected from multiple sources (variety) with a speed increasing volume (velocity and volume). Some techniques are used for the DQ assessment such as the sampling method or the use the artificial intelligence (smart data validators …) *[14],* but researchers still conducting works for the assessment complications and decrease the processing costs [51]; [9].

Handling this challenges and resulting with the best assessment and improvement methods needs a deep analyze and understanding of each step of Big Data lifecycle and determine the relevant and affected quality dimensions inside each step of this lifecycle, then define the sensible DQ step inside a Big Data system.

VII. Conclusions and Future Works

One of the main keys to the organization's growth is the ability to making good decisions. Thus, making such decisions needs to be based on very valid, solid, and reliable Data and statistics visualization systems can cause considerable damage within the organization for lack of bad management, which requires a very high Data Quality. Ensuring the Quality of data inside a system passes by two main processes: the *DQ Assessment* to measure and evaluate the quality of data inside this system and *DQ Improvement* to fix and repair the lack of quality.

The emergence of Big Data transformed Data Quality management into a very complex task inside the organizations because of its characteristics (3Vs or more).

In this paper, we first discussed and introduced the main concepts that are related directly to the DQ. Then, we exposed, summarized, and compared the relevant Data dimensions' taxonomies existing in the literature that allowed thereafter many researchers to propose their DQ models, we noticed also that all taxonomies turn around two major classes: *data dependent* and *system dependent*. Afterward, we surveyed and compared more than twenty methods, frameworks and models proposed for the DQ assessment and/or improvement before and after the emergence of Big Data where we showed that the main challenge was and still the correct assessment of organization's DQ in the era of Big Data. The number of works and proposed approaches for the DQ assessment has surprisingly increased due to the quantification issue for the huge amount of data, even if a multiple of them proposed a very valid assessment method (such as the Quality assessment based on data sampling) but the Quality evaluation remains a considerable challenge. We proposed then our assessment meta-model and DQ assessment and improvement process based on analyzed models and approaches. Finally, we exposed the existing and remaining Data Quality problems and challenges before and after the emergence of Big Data.

As further perspectives, we plan to analyze and adapt the existing models or frameworks to match with the specific characteristics of Big Data by analyzing the different Big data lifecycle steps and criteria to determine the relevant assessment and improvement methods.

VIII. References

[1] International Organization for Standardization (September 2015). "ISO 9000:2015(en) Quality management systems — Fundamentals and vocabulary". International Organization for Standardization

[2] Fürber, C. (2015). "3. Data Quality". Data Quality Management with Semantic Technologies. Springer. pp. 20–55. ISBN 9783658122249. Archived from the original on 31 July 2020. Last access : April 2020.

[3] Huh, Y., Keller, F., Redman, T., & Watkins, A. (1990). Data quality. Information and Software Technology, 32(8), 559–565.

[4] LEE, Y. W., PIPINO, L. L., FUNK, J. D., AND WANG, R. Y. 2006. Journey to Data Quality. The MIT Press. LEE, Y. W., STRONG, D. M., KAHN, B. K., AND WANG, R. Y. 2002. AIMQ: A methodology for information quality assessment. Inf. Manag. 40, 133–146

[5] [Fox et al., 1994] Fox, C., Levitin, A., & Redman, T. (1994). The notion of data and its quality dimensions. Information Processing & Management, 30(1), 9–19.

[6] https://blog.syncsort.com/2019/08/data-quality/data-quality-dimensions-measure/ . Last access : April 2020.

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

[8] M. Ali-ud-din Khan, M. F. Uddin, and N. Gupta, "Seven V's of Big Data understanding Big Data to extract value," in American Society for Engineering Education (ASEE Zone 1), 2014 Zone 1 Conference of the, 2014, pp. 1–5.

[9] Panahy PHS, Sidi F, Affendey LS, Jabar MA, Ibrahim H, Mustapha A. A framework to construct data quality dimensions relationships. Indian J SciTechnol. 2013;6(5):4422–31.

[10] Caballero, I., Serrano, M., &amp;Piattini, M. (2014, October). A data quality in use model for big data. In International Conference on Conceptual Modeling (pp. 65-74). Springer, Cham.

[11] Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. Communications of the ACM, 39(11)

[12] foundations," Communications of the ACM, vol. 39, pp. 86-95, 1996.

[13] Sidi, F., ShariatPanahy, P. H., Affendey, L. S., Jabar, M. A., Ibrahim, H., & Mustapha, A. (2012). Data quality: A survey of data quality dimensions. 2012 International Conference on Information Retrieval & Knowledge Management. doi:10.1109/infrkm.2012.6204995

[14] Taleb, I., Kassabi, H. T. E., Serhani, M. A., Dssouli, R., &Bouhaddioui, C. (2016). Big Data Quality: A Quality Dimensions Evaluation. 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.

[15] Laranjeiro, N., Soydemir, S. N., & Bernardino, J. (2015). A Survey on Data Quality: Classifying Poor Data. 2015 IEEE 21st Pacific Rim International Symposium on Dependable Computing (PRDC).

[16] Merino, J., Caballero, I., Rivas, B., Serrano, M., &Piattini, M. (2016). A Data Quality in Use model for Big Data. Future Generation Computer Systems, 63, 123–130. doi:10.1016/j.future.2015.11.024

[17] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," Journal of Management Information Systems, pp. 5–33, 1996.

[18] https://baripedia.org/wiki/Introduction_et_typologie_des_syst%C3%A8mes_d%27information. Last access : may 2021

[19] Dammann, O. (2019). Data, Information, Evidence, and Knowledge: A Proposal for Health Informatics and Data Science. Online Journal of Public Health Informatics, 10(3). doi:10.5210/ojphi.v10i3.9631

[20] Juran's Quality Handbook: The Complete Guide to Performance Excellence, Sixth Edition. 2010.

[21] Färber, M., Bartscherer, F., Menne, C., &Rettinger, A. (2017). Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. Semantic Web, 9(1), 77–129. doi:10.3233/sw-170275.

[22] Ramasamy, Anandhi& Chowdhury, Soumitra. (2020). BIG DATA QUALITY DIMENSIONS: A SYSTEMATIC LITERATURE REVIEW. 10.4301/S1807-1775202017003.

[23] Abdallah, Mohammad &Muhairat, Mohammad &Thunibat, Ahmad &Abdalla, Ayman. (2020). Big Data Quality: Factors, Frameworks, and Challenges. 9. 3785.

[24] Cichy, C., &Rass, S. (2019). An Overview of Data Quality Frameworks. IEEE Access, 1–1.

[25] Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. Communications of the ACM, 45(4).

[26] Bovee, M., Srivastava, R. P., &Mak, B. (2003). A conceptual framework and belief-function approach to assessing overall information quality. International Journal of Intelligent Systems, 18(1), 51–74. doi:10.1002/int.10074

[27] Bai, L., Meredith, R., & Burstein, F. (2018). A data quality framework, method and tools for managing data quality in a health care setting: an action case study. Journal of Decision Systems, 27(sup1), 144–154. doi:10.1080/12460125.2018.1460161

[28] Merino, J., Caballero, I., Rivas, B., Serrano, M., &Piattini, M. (2016). A Data Quality in Use model for Big Data. Future Generation Computer Systems, 63, 123–130. doi:10.1016/j.future.2015.11.024

[29] Pratiwi, Andita&Anawar, Syarulnaziah. (2015). A theoretical framework of data quality in participatory sensing: A case of mHealth. JurnalTeknologi.

[30] Serhani, M. A., El Kassabi, H. T., Taleb, I., &Nujum, A. (2016). An Hybrid Approach to Quality Evaluation across Big Data Value Chain. 2016 IEEE International Congress on Big Data (BigDataCongress).

[31] Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems, 12(4), 5–33

[32] Alizamini, F. G., Pedram, M. M., Alishahi, M., &Badie, K. (2010). Data quality improvement using fuzzy association rules. 2010 International Conf. on Electronics and Information Engineering.

[33] Thomas C. Redman, Harvard Business Review (2016) https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year.

[34] Han, W.-M. (2017). Evaluating perceived and estimated DQ for Web 2.0 applications: a gap analysis. Soft. Quality Journal, (2), 367–383.

[35] Cai, Li & Zhu, Yangyong. (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal. 14. 10.5334/dsj-2015-002.

[36] Taleb, I., Serhani, M. A., &Dssouli, R. (2018). Big Data Quality: A Survey. 2018 IEEE International Congress on Big Data (BigData Congress). doi:10.1109/bigdatacongress.2018.00029

[37] Catarci, T., Scannapieco, M., Console, M., &Demetrescu, C. (2017). My (fair) big data. 2017 IEEE International Conference on Big Data (Big Data). doi:10.1109/bigdata.2017.8258267

[38] Batini, C., Cabitza, F., Cappiello, C., Francalanci, C., & di Milano, P. (2007). A Comprehensive Data Quality Methodology for Web and Structured Data. 2006 1st International Conference on Digital Information Management. doi:10.1109/icdim.2007.369236

[39] Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). AIMQ: a methodology for information quality assessment. Information & Management, 40(2), 133–146.

[40] Eppler, Martin &Helfert, Markus. (2004). A classification and analysis of data quality costs. International Conference on Information Quality.

[41] Carlo, B., Daniele, B., Federico, C., & Simone, G. (2011). A Data Quality Methodology for Heterogeneous Data. International Journal of Database Management Systems, 3(1), 60–79.

[42] Cappiello, C., Ficiaro, P., &Pernici, B. (2006). HIQM: A Methodology for Information Quality Monitoring, Measurement, and Improvement. Lecture Notes in Computer Science, 339–351.

[43] Johnson SG, Pruinelli L, Hoff A, Kumar V, Simon GJ, Steinbach M, et al. A framework for visualizing data quality for predictive models and clinical quality measures.

[44] https://www.eckerson.com/articles/a-data-quality-framework-for-big-data. Last access: May 2021

[45] Federal Committee on Statistical Methodology. 2020. A Framework for Data Quality. FCSM 20-04. Federal Committee on Statistical Methodology. September 2020.

[46] https://towardsdatascience.com/a-comprehensive-framework-for-data-quality-management-b110a0465e83. Last access: May 2021

[47] Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137–144. doi:10.1016/j.ijinfomgt.2014.10.007

[48] Woodall, P., Borek, A., & Parlikad, A. K. (2013). Data quality assessment: The Hybrid Approach.J. Infor & Manag, 50(7), 369–382.

[49] Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. ACM Computing Surveys, 41(3), 1–52. doi:10.1145/1541880.1541883

[50] Schmidt, C.O., Struckmann, S., Enzenbach, C. et al. Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R. BMC Med Res Methodol 21, 63 (2021).

[51] Wook, M., Hasbullah, N.A., Zainudin, N.M. et al. Exploring big data traits and data quality dimensions for big data analytics application using partial least squares structural equation modelling. J Big Data 8, 49 (2021). https://doi.org/10.1186/s40537-021-00439-5

[52] Blake, R. & Mangiameli, P., 2011. The effects and interactions of Data Quality and Problem Complexity on Classification. ACM Journal of Data and Information Quality, 2(2).

[53] Quinto, B. (2018). Big Data Governance and Management. Next-Generation Big Data, 495–506