# Towards Reliable Collaborative Data Processing Ecosystems: Survey on Data Quality Criteria

Louis Sahi, Romain Laborde, Mohamed Ali Kandi, Michelle Sibilla, Afonso Ferreira, Giorgia Macilotti, Abdelmalek Benzekri

# Towards Reliable Collaborative Data Processing Ecosystems: Survey on Data Quality Criteria

Louis Sahi[1], Romain Laborde[1], Mohamed-Ali Kandi[1], Michelle Sibilla[1],
Giorgia Macilotti[2], Benzekri Abdelmalek[1], Afonso Ferreira[2]
*Institut de Recherche en Informatique de Toulouse*
[1]*Université Toulouse III - Paul Sabatier*
[2]*CNRS*
Toulouse, France
{firstname.lastname@irit.fr}

***Abstract*—Data quality plays a crucial role in the data governance of organizations, as it is essential to ensure that data are fit for the purpose for which they are intended, whether for operational activities, decision-making processes, or strategic planning. As data silos begin to be integrated to form data spaces, guaranteeing data quality becomes a necessity to achieve a reliable collaborative ecosystem. Nevertheless, the concept of data quality remains ambiguous, with various definitions and interpretations offered in the literature, despite its importance. This lack of consensus has led to the need for a thorough review of the different data quality criteria used in scientific work. Therefore, this paper serves as a systematic survey aimed at exploring and consolidating diverse perspectives on data quality. By thoroughly analyzing existing literature, this study compiles a comprehensive set of 30 agreed-upon data quality criteria, with their respective names and definitions. These criteria act as a valuable resource for organizations seeking to establish effective data quality monitoring practices. Then, we expose challenges raised by collaborative data processing and highlight possible research directions where data quality plays a major role.**

***Index Terms*—Data quality, data quality criteria, trust, reliability, collaborative data processing, decentralized data governance.**

## I. Introduction

Data science and AI-based techniques are now widely used in various sectors, including business, politics, healthcare, transportation, research, etc. The new applications resulting from (big) data analytics technologies and processes are impacting our daily lives and will do so even more in the future, as reported by Forbes [1]. In addition, companies and public organizations have produced and/or collected various types of data which today are stored in data silos that need to be integrated to build a data economy that drives innovation [2]. Such data spaces should engage different stakeholders in collaborative and distributed data processing as well as decentralized data governance.

In this context, data quality (DQ) plays a critical role in data governance, as it is essential to ensure that data is fit for purpose, whether for operational activities, decision making or strategic planning. In fact, ensuring an appropriate level of DQ throughout the data lifecycle is fundamental to the production of valuable and reliable results. [3], [4]. Ted Friedman, vice president and distinguished analyst at Gartner, explained at the Gartner Data & Analytics Summit 2018 that, "As organizations accelerate their digital business efforts, poor data quality is a major contributor to a crisis in information trust and business value, negatively impacting financial performance" [5]. DQ aims to measure the suitability of the data to produce meaningful information and the ease with which it can be processed. [6]. It also refers to the ability to meet the needs and expectations of data consumers [7], [8]. The measurement of DQ is a combination of a set of parameters that characterise the value of the data or the process that produced or modified it [8]. These parameters are called Data Quality Criteria (DQC).

Several research articles have reviewed existing data quality criteria. As quality is domain-related, defined by a set of attributes, and based on measurement and evaluation methods [4], these reviews provide different sets of DQC. More importantly, they do not provide comprehensive and agreed criteria for data quality. As a result, there are many discrepancies between DQC names and their meanings. Indeed, some criteria may be defined differently from one article to another. For example, Cichy et al. [9] defined *timeliness* as *"the extent to which the age of the data is appropriate for the task at hand"*, while Wand et al. [10] defined this same criterion as *"the delay between a change of the real-world state and the resulting modification of the information system state"*. At the same time, some articles provided similar definitions for criteria with different names. For example, Tejay et al. [11] defined *appropriateness* as follows *"data must be appropriate to the task at hand"*, while for Pipino et al. [12], it is *relevancy* which refers to *"the extent to which data is applicable and useful for the task at hand"*.

Therefore, this paper aims to address the existing inconsistencies by conducting a systematic literature review to identify the most pertinent DQCs. By thoroughly analysing the existing literature, this study compiles a comprehensive set of 30 agreed data quality criteria with their respective names and definitions. These serve as a valuable resource for organisations seeking to establish effective data quality monitoring practices. Additionally, collaborative data processing and data

spaces open new issues related to data quality. We expose challenges that need to considered by future research.

The remainder of the paper is structured as follows: Section 2 corresponds to the state of the art in data quality assessment and its limitations. Section 3 presents our research methodology for collecting relevant research papers and the DQCs they contain. It also explains our approach to name and define them. Section 4 analyses and details the results of this survey. Section 5 highlights challenges raised by collaborative data processing. Finally, section 6 concludes the study and highlights its limitations. It then announces the future directions of our research activities.

## II. Related work

Many surveys and research articles related to data management/processing have proposed different DQCs. We found that authors focused on specific contexts, such as health information systems [13]–[15], information system management [16], information security management [11], [17], business performance [8], [18]–[20], database [21], [22], open data [23], linked data application [24], [25], genomic data [26], [27], big data analytics [28]–[30], and citizen science [31]. All authors aimed to assess data quality but looked at different contexts, purposes, data lifecycles and outcomes. As a result, they proposed distinct lists of DQCs. These differences concern the numbers, names and meanings of the DQCs.

Some surveys or research articles provided different significations to the same DQC. me surveys or research articles provided different significations to the same DQC. For instance, Vetro et al. [23] stated that understandability is *"constituted by percentage of columns with metadata and percentage of columns in comprehensible format. Percentage of columns with metadata indicates the percentage of columns in a dataset that has associated descriptive metadata"*. But, Tejay et al. [11] advanced that understandability *"is concerned with whether data is clear, readable, unambiguous and easily comprehensible"*. The difference between the later definitions is related to the divergence of their contexts. Indeed, Vetro et al. presented a metrics-driven assessment framework for evaluating the quality of open government data. Furthermore, they examine the influence of both decentralized and centralized data sharing on data quality. In contrast, Tejay et al. examined how data quality affects information systems security in companies. They claimed that the effectiveness of information systems security in a company largely depends on the quality of data used to manage it. Consequently, they identified various aspects of data quality that impact the information system security efforts of companies.

This shows that DQC definitions are not consistent or standardized. We couldn't find a complete set of quality standards for data that cover all data types and fields.

Furthermore, the surveys have insufficient information about their methods, leading to a lack of transparency in the survey process. Consequently, reproducing their analysis is challenging.

As a result, our goal was to find the general DQCs that can apply to all kinds of fields through a methodical review of the literature. In the following section, we will describe our research approach to accomplish this goal.

## III. Research methodology

In this section, we describe our methodology to conduct a systematic literature survey. We explain how we selected and redefined the resulting DQC collected from the survey. Our methodology is similar to those of Shah et al. [32], [33] and Bowling Ann et al. [34]. The survey process consists of four main steps:

1) Formulation of the research questions.
2) Identification of the pertinent research works.
3) Selection of best articles through inclusion and exclusion criteria.
4) Analysis and verification.

### A. Formulation of the research question and Identifying pertinent research works

Our goal can be summarized by the following research questions:

- What are the existing data quality criteria proposed by the literature?
- What are the most relevant and generic criteria for assessing data quality?

Starting from these questions, we realized a comprehensive analysis of the current surveys and research works on DQCs. To discover appropriate material for our objective, we studied recent research in the topic field. This part clarifies the actions we executed in our organized assessment.

We chose IEEE Xplore, Science Direct, ACM, Springer Link, Web of Science and Google Scholar. Next, in each research library, we carried out the following actions:

- Formulate research queries corresponding to the research questions mentioned above.
- Use operators like "OR" and "AND" to expand the requests with other words and synonyms.
- All the queries included keywords like "data", "information", "quality", "management", and "assessment"...

We conducted the study from 18th July 2022 until 14th April 2023. Throughout this survey, we utilized the subsequent queries: 'data quality' OR 'information quality' AND ('dimensions' OR 'evaluation' OR 'assessment' OR 'criteria' OR 'management', 'categories' OR 'characteristics' OR 'elements', OR 'assurance'). We searched these keywords in the title, abstract, and body of the different documents.

The requests mentioned above generated 3480 documents. We then picked the articles with headings related to "data" and "quality" or "information" and "quality". This resulted in 604 relevant scientific reports for our investigation. Figure 1 provides an overview of this detailed analysis.

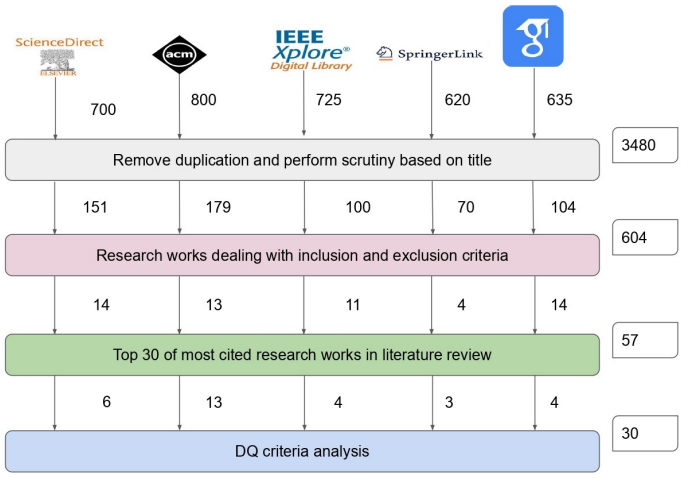


Fig. 1. Selection of relevant research articles

### B. Selection of relevant papers through inclusion and exclusion criteria

Between the 604 previous documents, we selected those with these features:

- Title includes "data" and "quality" or "information" and "quality",
- The objective expressed in the abstract concerns the assessment of data or information quality,
- The body of the article provides clear definitions of cited data quality dimensions,
- The articles have an important number of citations (20 citations at least).

Moreover, we deleted some research works where:

- Authors cited DQC without defining them,
- The paper focused on service, audio or video quality with a little attention to generic data quality dimensions.

This selection reduced the list to 57 research articles. After that, we sort the articles based on how many times they have been cited and which libraries have them. We created a file and saved the documents, noting details such as the year of publication, authors, title, citation count, and number of criteria. We identified the top 30 research articles based on their number of citations. The last article on the list had 24, whereas the first had 2412

### C. Synthesis of data quality definitions

To make sure that we clearly explain DQC and prevent any confusion, we've established the following algorithm :

*Step 1* Consider the most relevant definitions and names of DQC come from articles having the highest number of citations.

*Step 2* For each name of DQC, identify all the definitions in the literature review and classify them into four categories:

- C1 is the definition that has the highest number of citations in all the 30 selected paper,
- C2 contains all synonym definitions of the C1 definition, i.e., the semantic is similar but words are different,
- C3 is the set of definitions that complements C1 definition by adding new characteristics,
- C4 lists definitions have no connection with the C1 definition.

*Step 3* For each definition in a C4 category, evaluate if the definition should be associated to another DQC name. If so, move the definition to the category C2 or C3 of the related DQC name.

*Step 4* Propose the final definition based on C1 definition complemented by new characteristics coming from C3 definitions, i.e., new characteristics not related to a specific.

*Step 5* Neglect the definitions of C4 if the article has less than 10 citations and was published before 2010 (we chose 2010 because the most cited articles were published between 1996 and 2005). Otherwise, mention the definitions in the final definition as alternative perceptions of the criterion.

## IV. Analysis and Results

### A. Analysis of research works

This part of the paper shows the details gathered within the 30 relevant research works mentioned above. We studied all the data quality criteria presented in these papers. These papers specified a total of 270 DQC. However, several research articles proposed the same DQC. Thus, we regrouped the DQC having the same name and a similar definition to avoid redundancy. Then, we realized there were just 30 common DQC in the entire set of papers. Table I shows the resulting list of DQC.

| Nb | Criteria names | All names of DQC in literature review |
|---|---|---|
| 1 | Accessibility | accessibility [3], [8], [9], [11], [12], [14], [20], [26]; availability [8], [17], [35]; access [36], [37] |
| 2 | Accuracy | accuracy [3], [8]–[10], [23], [25], [28], [36], [38]–[41] [4], [11], [17], [22], [26], [29], [31], [35], [37], [42], [43] [14]; data accuracy [44]; precision [42]; correctness [11] |
| 3 | Appropriate amount of data | appropriate amount of data [8], [12]; appropriate amount of information [20], [26]; amount of data [8], [17]; amount of information [36]; data resolution [45] |
| 4 | Auditability | auditability [3]; verifiability [17] |
| 5 | Authorization | authorization [3] |
| 6 | Believability | believability [8], [12], [17], [20], [22], [26], [36]; credibility [3] |
| 7 | Communication | communication [45] |
| 8 | Completeness | completeness [3], [8], [10], [12], [20], [23], [24], [28], [38]–[41] [4], [9], [11], [17], [22], [26], [35]–[37], [42] [14], [29]; data coverage [8]; information completeness [44] |
| 9 | Concise representation | concise representation [12], [17], [20], [26], [36]; concise [8]; format [11], [14], [28]; representational conciseness [22], [35]; conciseness [11] |

| | | |
|---|---|---|
| 10 | Consistency | consistency [3], [4], [8]–[11], [24], [31], [35], [38], [40]–[42] [14], [29], [37]; conciseness [24]; consistency and synchronization [8]; data consistency [44] |
| 11 | Consistent representation | consistent representation [8], [12], [17], [20], [26], [36]; representational consistency [22], [25]; presentation [37] |
| 12 | Currency, timeliness and volatility | Currency [8], [28], [37]–[39]; timeliness [3], [8]–[12], [17], [20], [35], [36], [38], [41], [42] [14], [29]; volatility [8], [38]; currentness [23]; freshness [8]; timeliness and availability [8]; data currency [44]; temporal reliability [42]; up to date [26]; temporal relevance [45]; chronology of data and goal [45] |
| 13 | Data integration | data integration [45]; interlinking [25]; navigation [8] |
| 14 | Duplication | duplication [8], [22]; data deduplication [44]; uniqueness [42] |
| 15 | Ease of manipulation | ease of manipulation [8], [12], [20], [26]; useability [8]; ease of operation [11]; flexibility [11]; reuse [35] |
| 16 | Free of error | free of error [8], [12], [20] |
| 17 | Generalizability | generalizability [45] |
| 18 | Integrity | integrity [3], [11]; data integrity fundamentals [8] |
| 20 | Objectivity | objectivity [8], [11], [12], [14], [20], [36]; objective [17]; unbiased [26] |
| 21 | Relevancy | relevancy [8], [11], [12], [14], [17], [20], [25], [36]; fitness [3]; effectiveness [8]; useful [8]; efficiency [8]; transactibility [8]; convenience [36]; appropriateness [11]; relevance [26]; operationalization [45] |
| 22 | Reliability | reliability [8], [10], [11], [17], [31], [42] |
| 23 | Reputation | reputation [8], [12], [14], [20], [26], [36] |
| 24 | Safety | safety [8] |
| 25 | Security | security [8], [11], [12], [14], [20], [26], [35], [36] |
| 26 | Structure | structure [3] |
| 27 | Traceability | traceability [23], [26], [35] |
| 28 | Understandability | understandability [8], [11], [12], [14], [17], [20], [23], [26]; metadata [3]; learn ability [8]; data specification [41]; ease of understanding [36]; meaningfulness [11]; usefulness [35] |
| 29 | Validity | validity [11], [39], [42]; compliance [23] |
| 30 | Value added | value added [8], [11], [12], [20], [26], [36] |

TABLE I
30 FINAL DATA QUALITY CRITERIA AND THEIR DIFFERENT NAMES IN LITERATURE REVIEW

### *B. Definitions of DQC*

Once we finalized the consolidated list of DQC names, we summarised the definitions according to the methodology described in Section III-C. We ended up with the definitions presented in Table II.

| **Nb** | **DQ criteria names** | **Other names** | **Definitions or significations of criteria** |
|---|---|---|---|
| 1 | Accessibility | Availability, access | The ability, ease or difficulty with which data is available, easily and quickly retrievable [3], [12], [20] |
| 2 | Accuracy | Data accuracy, precision, correctness | The data are accurate when their values in the database match up to real world values. Again, accuracy refers to the closeness between a data value to a known reference value of one object [3]. Furthermore, it is the degree to which data are correct, reliable, certified, free of error, believable, and valid [9], [25], [38] |
| 3 | Appropriate amount of data | Appropriate amount of information, amount of data, amount of information, data resolution | It verifies if the data volume and amount are sufficient and appropriate for the task (neither too much nor too little) [17], [20], [26] |
| 4 | Auditability | Verifiability | Auditability means that data quality (accuracy, integrity...) shall be properly assessed by auditors in a sufficient time period, with low numbers of staff, during the different phases [3] |
| 5 | Authorization | | Authorization is the degree to which an individual or organization has the right to use the data [3] |
| 6 | Believability | Credibility | The degree to which data is regarded as true, credible, trustworthy specially as within range of known possibility or probability [11], [12], [36] |
| 7 | Communication | | It requires that data must reach the right person at the right time in a clear and understandable way [45] |
| 8 | Completeness | data coverage, information completeness | The quality of an information system or database to represent all valid and meaningful values describing a real world system [10]. It is also the level to which data are of sufficient breadth, depth and scope for a given task [12]. Moreover, completeness means the level to which data units are present in between collected data. Percentage of the real-world information entered in the sources and/or the data warehouse [24]. It also shows whether or not data have all required parts to answer some questions and or cover specific needs [44]. Completeness deals with the ratio between the number of non-null values in a source and the size of the universal relation. All values that are supposed to be collected as per a collection theory. It can be evaluated at database, dataset, data record, and data rows levels. |
| 9 | Concise representation | Concise, format, representational conciseness, conciseness | The measure of the level to which data is compactly represented without being overwhelming (i.e. brief in presentation, yet complete and comprehensible) [10], [12]. It indicates whether data is without superflous detail, well formatted and well presented to data users [8], [28], [36]. |

| | | | |
|---|---|---|---|
| 10 | Consistency | Conciseness, consistency and synchronization, data consistency | Consistency indicates the degree to which all elements in the data set follow the same semantic rules including format, structure, range of values, data type, intervals and representation [11], [29], [38]. This property avoid any contradictions and violations of constraints within the data set [37]. Moreover, consistency infers that data objects of the same real world entity cannot have different values when they are collected in the identical conditions [10], [31] |
| 11 | Consistent representation | Representational consistency, presentation | The degree to which data is presented and structured in the same format [12], [25]. Then, it refers to the ability to be marked by harmony, regularity and free of change, deviation or contradiction [36] |
| 12 | Currency, timeliness and volatility | Currentness, freshness, timeliness and availability, data currency, temporal reliability, up to date, temporal relevance, chronology of data and goal | **Currency:** The degree to which the data are sufficiently up to date to perform a task [12]. **Timeliness:** It is the degree to which age of the data is relevant for the task at hand [8]. More, it specifies the delay between a change of the real-world state and the correspondent modification of the state in databases. Again, it is the length of time before data was recorded [10]. **Volatility:** it refers to the period for which information is valid in the real world [8], [38] |
| 13 | Data integration | Interlinking, navigation | Data is often spread out across multiple data sources. Data integration evaluates whether relevant data sources are identified and linked between them. Furthermore, it assesses if the relevant data are collected when integrating them [8], [25], [45] |
| 14 | Duplication | Data deduplication, uniqueness | The ability to be free of duplication and redundancy when collecting or integrating data [8], [22], [44] |
| 15 | Ease of manipulation | Useability, ease of operation, flexibility, reuse | It indicates the abilities of data to be easily manipulated, customized and able to be assigned to multiple purposes [11], [26]. |
| 16 | Free of error | | The degree to which the data is correct and reliable [8], [12], [20] |
| 17 | Generalizability | | Generalizability includes statistical and scientific generalizability. Statistical generalizability refers to inferring from a sample to a target population. Scientific generalizability means apply a model based on a singular target population to other populations [45] |
| 18 | Integrity | Data integrity fundamentals | Integrity also deals with the fact to maintain and ensure the correctness and consistency of data over their entire life cycle. It indicates that any data characteristics, including those relating to business rules, relations, dates, content, format, definitions, must be correct and unchanged [8], [11]. |
| 19 | Interpretability | Readability, definition, documentation | This criterion refers to the degree to which data is in relevant languages, symbols, units, ranges of valid values, business rules [3] and the definitions are clear [8], [10], [26]. It ensures that data should be interpretable, with clear and appropriate representation so that data users can understand this data [11], [22] |
| 20 | Objectivity | Objective, unbiased | Objectivity checks if data is based on objective, based on facts, unbiased, unprejudiced, impartial and was objectively collected [11], [20]. Moreover, it means that data collection is not influenced by personal feelings or opinions, collected without using subjective judgements which introducing to bias [17] |
| 21 | Relevancy | Fitness, effectiveness, useful, efficiency, transactibility, convenience, appropriateness, relevance, operationalization | This criterion evaluates if data outputs will lead to the desired business expectation and whether concrete actions derived from these outputs [8], [45]. Relevancy also indicates whether data is applicable and useful for one or many contexts [8], [20]. Furthermore, it assesses if data is provided in accordance with specific purpose and deals with customer's needs [11], [36]. |
| 22 | Reliability | | This criterion indicates the degree to which data is correct, free of error and provided by a trustworthy source [8]. A reliable data has the ability to conform with customer needs by answering questions and problems for which it is generated or collected, [10], [31] |
| 23 | Reputation | | This criterion measures if data sources and content are trustworthy [8], [12], [20] |
| 24 | Safety | | Data have acceptable risks of damage to persons, processes, assets or the environment [8] |
| 25 | Security | | The level to which data access is restricted in a manner that ensures their security [26]. Thus, security criterion challenge the existence of adopted measures to protect against espionage, sabotage, crime, attack, escape, people and natural disasters [11], [36]. It assesses also the compliance with confidentiality and privacy requirements [14], [35] |
| 26 | Structure | | It consists in identifying the best type of data for a given objective [45]. Then, data structure is the level of difficulty when transforming semi structured or unstructured data into structured data through technology [3] |
| 27 | Traceability | | Traceability assesses the existence of history and documentation serving to record and trace all actions on data like collection and modification [23], [26], [35] |

| | | | |
|---|---|---|---|
| 28 | Understand-ability | Metadata, learn ability, data specification, ease of understanding, meaningfulness, usefulness | Firstly, it shows if data is clear, readable, unambiguous and easily comprehendible for human and machines. Secondly, understandability displays the level to which metadata describe all aspects of datasets to avoid misunderstandings and inconsistencies [3], [23] |
| 29 | Validity | Compliance | Validity measures the extent to which data items comply with their respective standards and value domains [11], [23]. [39] said that "a data item is invalid in all these following conditions: if it is defined to be integer but contains a non-integer value, or defined as an element of a finite set of possible values but contains a value not included in this set, or contains a NULL value where a NULL is not accepted" |
| 30 | Value added | | The extent to which data are beneficial then have an operational or organisational advantage for its use [12], [36] |

TABLE II
30 AGREED-UPON DATA QUALITY CRITERIA WITH THEIR RESPECTIVE NAMES AND DEFINITIONS

Table II lists the DQCs in alphabetical order. It also lists some of the synonyms found in the literature and their definitions. As each definition is an aggregation of ideas proposed by other authors, we include within the definitions the references of the source of each idea for traceability.

When developing the consolidated definitions, we had to deal with different situations.

Firstly, some groups of similar DQC had almost the equivalent definition and name. For instance, the common definition of believability is the extent to which data is considered true and credible [12]. One thing which changes with the other definitions is the subject or the data instance. Indeed, [12] and [11] used the term **data**, while [8] and [20] considered **information**, and [26] dealt with a **sequence record**. Because [36] defined this criterion as the ability *of being believed especially as within the range of known possibility or probability*, and given that information and sequence record are instances of data, the final definition of **believability** we proposed is "the degree to which data is regarded as true, credible, trustworthy specially as within range of known possibility or probability". It is the same situation for criteria ***accessibility, appropriate amount of data, ease of manipulation, free of error, objectivity, reputation, value added, and security***.

The second situation appears when similar criteria have definitions which are synonyms. For instance, [8] indicated that **reliability** is defined by "*the extent to which information is correct and reliable. Then the capability of the function to maintain a specified level of performance when used on specified condition*". While [10] thought that "*reliability has been linked to probability of preventing errors or failures, to consistency and dependability of the output information, and to how well data ranks on accepted characteristics*". The criteria of **structure, data integration, auditability and traceability** are in the same case.

On another side, some groups of DQC contain multiple phrasing of similar properties that change according to the context targeted by the authors. For instance, the difference among all definitions of **completeness** concerned the fact that:

- some authors defined completeness as a property of a value [3], [4], [22], [29], [31], [37], [42], a data [9], [11], [12], [40], an information [20], [35], [36] or a sequence record [26];
- others defined it as a property of information system [10], database [44] or Big Data analytic system [28]
- more, some of the authors represented it as a characteristic of a data set [24] or a characteristic a set of cells and rows [23]
- one of them described completeness as a criteria of a data collection [38];
- Blake Roger et al. [41] gave a mathematical representation to this criterion as it said that *completeness, K, is the ratio of the number of tuples with null values to the total number of tuples in a relation and is defined as 1 - (MT/NK), where MTis the number of tuples in a relation having a null value and NK the total number of tuples*
- finally [14] interpreted it as a set of sub-criteria encompassing coverage, comprehensiveness, appropriate amount, adequate, integrity.

Thus, our final definition of **completeness** encompasses all aspects of the instances that can represent data. It is based on the definition of [8] because their definition includes all the others. Other definitions of DQC like ***concise representation, consistent representation, accuracy, consistency, integrity, duplication, and validity*** were formalized through the same analysis process.

Moreover, some definitions concern the fitness of the data with the users' expectations or the ability to achieve an objective. For instance, **relevancy** refers to *"the extent to which information is applicable and useful for the task in hand"* [20] or *the ability of data to be adequate to customer needs* [11], [36]. More, **interpretability** was presented as the *degree to which user can understand data that they get* [22] or the fact *sequence records are in appropriate languages, symbols, and units, and the definitions are clear for interpretation* [26]. The explanation of **understandability** is similar to those mentioned above.

The criteria of **safety, generalizability and communication** had unique definition in our literature review. We kept their original definition and we paraphrased them.

The literature review provides four DQC types related to the time (time of data generation, use or update). The first type is *the delay between a change of the real-world state and the resulting modification of the information system state*. This criterion is called timeliness by [10] and [39]. The second type is *"the extent to which age of the data is appropriated for the task at hand"*. The authors of [9], [17], [36] and [11] named it timeliness. The third type is *the extent to which the data is sufficiently up-to-date for the task at hand*. This criterion is introduced as currency by [28], [39], [44] and [37]. Though,

it is named as timeliness by [12], [20] and [35]. The fourth DQC type describes *"the time period for which information is valid in the real world"*. It is called volatility by [38] and [8]. Thus, we combined the first and second types into one QDC called timeliness, and we kept currency and volatility for respectively the third and fourth types.

## V. Challenges and future research directions

Private and public organizations have produced and/or collected various types of data which today are stored in data silos. These silos need to be integrated to enable knowledge to emerge [46]. The European Commission has adopted a European strategy to engage stakeholders in collaborative data spaces [2]. As data processing is organized in several different stages, each involving organizational, technological and legal entities [47], private and public organizations need to collaborate through decentralized governance and distributed technologies [48]. This collaborative data processing raises several issues and challenges, especially, ensuring the reliability of distributed systems [49], trust in the decentralized governance of data processing, and compliance with legal requirements concerning data processing [50], [51]. Data quality plays a central role in these challenges to build a data economy.

### A. Automated assessment of data quality

Automated data quality assessment describes the use of algorithms, rules, or models to assess data quality without the need for direct human intervention. These methods can evaluate different facets of data quality, including the criteria described in this paper. However, until data quality criteria are standardised and unified, this will not be possible. While some previous studies have classified criteria primarily based on data-related dimensions [52], none of them have considered the contextual aspects of data processing to the best of our knowledge. Thus, there is a clear need for a more comprehensive and precise classification that takes into account other critical dimensions of data processing, including data lifecycle, governance, and regulatory considerations. Our work represents a first step towards this goal. Our research aims to develop a framework that not only considers traditional data-related dimensions but also the broader context of data processing. This framework will serve as a foundation for future efforts to automate data quality assessments, and thereby enhancing the reliability and trust of shared data across diverse domains and applications.

### B. Data quality in collaborative data processing

In the time when the EU is building the European Common Data Spaces, the need to ensure the quality of shared data has become increasingly important. A data space encompasses all its participants, including data providers, intermediaries, and users. A critical aspect of the data Space concept is that data is not stored centrally but remains at its source. Therefore, data is only transferred when necessary. This joint approach to data management in the EU highlights the importance of effective collaborative data processing and data quality assurance strategies. Collaborative data processing not only helps to manage data within decentralized governance models but also addresses the challenges of compliance, trust, and reliability, which are essential components for ensuring the quality of shared data in such dynamic and distributed environments.

This collaborative data processing raises challenges at three levels: (i) Compliance with legal requirements and regulatory Frameworks, (ii) trust in decentralized governance and (iii) reliability of distributed systems.

*1) Achieve compliance with regulations:* The concept of a Data Space implies a community-based approach for managing and exploiting data. Within such a community, commitments need to be formalized through contractual agreements, such as confidentiality or non-disclosure. This raises questions about how the underlying layers will ensure compliance and define a level of adherence to regulations. How can we formalize data quality through contracts, guarantee it, and establish a means of evaluation? Especially, regulatory requirements may change according to the type of data and its use. To the best of our knowledge, there is no existing work that categorizes data quality criteria according to regulations (RGPD, Data Act, Data Governance Act, AI Act, Open Data EU regulation...). There is a need for scientific research to develop a framework that aligns data quality criteria with the evolving regulatory landscape, particularly in the context of community-driven data spaces.

*2) Increase trust in decentralized governance:* Trust is a subjective concept characterizing a relationship among two or multiple entities with a common purpose [53]. In decentralized data governance, it is the belief that all participants in that data governance are willing and able to produce high quality data outputs. Indeed, having multiple entities managing data raises trust concerns, as the self-interest of one entity processing the data may conflict with the overall benefit of other entities [54]. Assessing the trustworthiness of individual contributors and their commitment to producing high quality data outputs is essential. This need leads to the following questions: are data governance stakeholders able to make the right decisions to maintain data quality? What are the data quality criteria that can be used to assess trust in all data governance stakeholders based on their actions and decisions? What are the data quality criteria relevant to data governance?

*3) Ensure reliability of distributed systems:* Reliability is the likelihood that a system, including hardware and software materials, will successfully perform its intended tasks under specified conditions and over specified periods of time [55]. Distributed reliability extends this concept to distributed systems. Akimova et al. [56] define the reliability of distributed systems as the ability to maintain the defined criteria required to perform a given task under the influence of failures, breakdowns, hardware-based or human errors, etc. In the context of data processing, distributed system reliability means that each step is performed with fault tolerance by all the technologies involved, resulting in high quality outputs. The reliability of distributed systems depends on the systems, the data life cycle

phases (data collection, data preparation, data analysis, etc [33]), the data transactions and, most importantly, the quality of the data outputs. It is therefore necessary to assess the reliability of all entities in distributed systems, i.e. the ability of each component to perform correctly and not degrade the quality of the data. Future research should focus on to create data quality contracts at each phases of the data life cycle based on appropriate data quality criteria.

## VI. Conclusion

Data quality can be expressed in terms of criteria that explain the three levels of requirements (compliance with regulations, trust in governance and reliability of data processing). Ensuring this quality means automating its assessment and monitoring its maintenance. This becomes an essential part of the data management activity and will become challenging when governance is decentralised and processing is distributed among different stakeholders.

This paper provides a framework to assess data quality in data processing. It identified the relevant data quality criteria in the literature review. We selected 57 research articles providing data quality criteria. Then, we choose the 30 most relevant among them. From these papers, we identified 30 DQC. From the definitions of the literature review, we proposed a unified and standardized definition for each criterion. This study fixed some difficulties and limitations. We examined ACM, ScienceDirect, IEEE, Springer and Google Scholar digital research libraries as we considered them more consistent with this research.

Our future work will focus on the practical validation of this framework to reduce the gap between academic methodology and industrial applicability. As a consequence, we need to validate this list of 30 DQC and associate the different DQC to each phases of the data lifecycle in order to build a data quality traceability system. Access to data processes and datasets of private companies is difficult. Nonetheless, we can start by assessing academic datasets as researchers make available their data management plan, data papers and FAIR principles, which describe the process they followed to produce their research data.

## Acknowledgments

This work was partially supported by the European research projects H2020 CyberSec4Europe (GA 830929) and LeADS (GA 956562), Horizon Europe DUCA (GA 101086308), and CNRS EU-CHECK.

## References

[1] B. Marr, "15 Amazing Real-World Applications Of AI Everyone Should Know About," May 2023. Section: Enterprise Tech.

[2] "Building a data economy — Brochure | Shaping Europe's digital future," Sept. 2019.

[3] L. Cai and Y. Zhu, "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era," *Data Science Journal*, vol. 14, p. 2, May 2015. Number: 0 Publisher: Ubiquity Press.

[4] I. Taleb, H. T. E. Kassabi, M. A. Serhani, R. Dssouli, and C. Bouhaddioui, "Big Data Quality: A Quality Dimensions Evaluation," in *2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld)*, pp. 759–765, July 2016.

[5] S. Moore, "How To Create A Business Case For Data Quality Improvement," June 2018.

[6] N. West, J. Gries, C. Brockmeier, J. C. Göbel, and J. Deuse, "Towards integrated Data Analysis Quality: Criteria for the application of Industrial Data Science," in *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*, pp. 131–138, Aug. 2021.

[7] "A Quantitative Study of the Relationship of Data Quality Dimensions and User Satisfaction with Cyber Threat Intelligence - ProQuest."

[8] 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," in *2012 International Conference on Information Retrieval & Knowledge Management*, pp. 300–304, Mar. 2012.

[9] C. Cichy and S. Rass, "An Overview of Data Quality Frameworks," *IEEE Access*, vol. 7, pp. 24634–24648, 2019. Conference Name: IEEE Access.

[10] Y. Wand and R. Y. Wang, "Anchoring data quality dimensions in ontological foundations," *Communications of the ACM*, vol. 39, pp. 86–95, Nov. 1996.

[11] G. Tejay, G. Dhillon, and A. G. Chin, "Data Quality Dimensions for Information Systems Security: A Theoretical Exposition (Invited Paper)," in *Security Management, Integrity, and Internal Control in Information Systems* (P. Dowland, S. Furnell, B. Thuraisingham, and X. S. Wang, eds.), IFIP International Federation for Information Processing, (Boston, MA), pp. 21–39, Springer US, 2005.

[12] L. L. Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment," *Communications of the ACM*, vol. 45, pp. 211–218, Apr. 2002.

[13] M. Sarafidis, M. Tarousi, A. Anastasiou, S. Pitoglou, E. Lampoukas, A. Spetsarias, G. Matsopoulos, and D. Koutsouris, "Data Quality Challenges in a Learning Health System," *Studies in health technology and informatics*, vol. 270, pp. 143–147, June 2020.

[14] J. Alipour and M. Ahmadi, "Dimensions and assessment methods of data quality in health information systems," *Acta Medica Mediterranea*, vol. 2017, pp. 313–20, Mar. 2017.

[15] S. Juddoo, C. George, P. Duquenoy, and D. Windridge, "Data Governance in the Health Industry: Investigating Data Quality Dimensions within a Big Data Context," *Applied System Innovation*, vol. 1, p. 43, Dec. 2018. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

[16] M. A. Jabar and A. S. M. Alnatsha, "Knowledge management system quality: A survey of knowledge management system quality dimensions," in *2014 International Conference on Computer and Information Sciences (ICCOINS)*, pp. 1–5, June 2014.

[17] P. Shamala, R. Ahmad, A. Zolait, and M. Sedek, "Integrating information quality dimensions into information security risk management (ISRM)," *Journal of Information Security and Applications*, vol. 36, pp. 1–10, Oct. 2017.

[18] P. H. S. Panahy, F. Sidi, L. S. Affendey, and M. A. Jabar, "The impact of data quality dimensions on business process improvement," in *2014 4th World Congress on Information and Communication Technologies (WICT 2014)*, pp. 70–73, Dec. 2014.

[19] A. Ahlemeyer-Stubbe and S. Coleman, "How to Create Profit Out of Data," in *Monetizing Data: How to Uplift Your Business*, pp. 187–202, Wiley, 2018. Conference Name: Monetizing Data: How to Uplift Your Business.

[20] B. K. Kahn, D. M. Strong, and R. Y. Wang, "Information quality benchmarks: product and service performance," *Communications of the ACM*, vol. 45, pp. 184–192, Apr. 2002.

[21] Munawar, N. Salim, and R. Ibrahim, "Comparative Study of Data Quality Dimensions for Data Warehouse Development: A Survey," in *Advanced Machine Learning Technologies and Applications* (A. E. Hassanien, A.-B. M. Salem, R. Ramadan, and T.-h. Kim, eds.), Communications in Computer and Information Science, (Berlin, Heidelberg), pp. 465–473, Springer, 2012.

[22] M. Souibgui, F. Atigui, S. Zammali, S. Cherfi, and S. B. Yahia, "Data quality in ETL process: A preliminary study," *Procedia Computer Science*, vol. 159, pp. 676–687, Jan. 2019.

[23] A. Vetrò, L. Canova, M. Torchiano, C. O. Minotas, R. Iemma, and F. Morando, "Open data quality measurement framework: Definition and application to Open Government Data," *Government Information Quarterly*, vol. 33, pp. 325–337, Apr. 2016.

[24] P. N. Mendes, H. Mühleisen, and C. Bizer, "Sieve: linked data quality assessment and fusion," in *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, EDBT-ICDT '12, (New York, NY, USA), pp. 116–123, Association for Computing Machinery, Mar. 2012.

[25] A. Zaveri, D. Kontokostas, M. A. Sherif, L. Bühmann, M. Morsey, S. Auer, and J. Lehmann, "User-driven quality evaluation of DBpedia," in *Proceedings of the 9th International Conference on Semantic Systems*, (Graz Austria), pp. 97–104, ACM, Sept. 2013.

[26] H. Huang, B. Stvilia, C. Jörgensen, and H. W. Bass, "Prioritization of data quality dimensions and skills requirements in genome annotation work," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 1, pp. 195–207, 2012. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.21652.

[27] A. Bernasconi, "Data quality-aware genomic data integration," *Computer Methods and Programs in Biomedicine Update*, vol. 1, p. 100009, Jan. 2021.

[28] N. Côrte-Real, P. Ruivo, and T. Oliveira, "Leveraging internet of things and big data analytics initiatives in European and American firms: Is data quality a way to extract business value?," *Information & Management*, vol. 57, p. 103141, Jan. 2020.

[29] M. A. Serhani, H. T. El Kassabi, I. Taleb, and A. Nujum, "An Hybrid Approach to Quality Evaluation across Big Data Value Chain," in *2016 IEEE International Congress on Big Data (BigData Congress)*, pp. 418–425, June 2016.

[30] I. Taleb, M. A. Serhani, and R. Dssouli, "Big Data Quality: A Survey," in *2018 IEEE International Congress on Big Data (BigData Congress)*, pp. 166–173, July 2018.

[31] S. A. Sheppard and L. Terveen, "Quality is a verb: the operationalization of data quality in a citizen science community," in *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, WikiSym '11, (New York, NY, USA), pp. 29–38, Association for Computing Machinery, Oct. 2011.

[32] S. I. H. Shah, V. Peristeras, and I. Magnisalis, "Government Big Data Ecosystem: Definitions, Types of Data, Actors, and Roles and the Impact in Public Administrations," *Journal of Data and Information Quality*, vol. 13, pp. 8:1–8:25, May 2021.

[33] S. I. H. Shah, V. Peristeras, and I. Magnisalis, "Dalif: a data lifecycle framework for data-driven governments," *Journal of Big Data*, vol. 8, no. 1, pp. 1–44, 2021.

[34] A. Bowling, "Mode of questionnaire administration can have serious effects on data quality," *Journal of Public Health*, vol. 27, pp. 281–291, Sept. 2005.

[35] D. Gürdür, J. El-khoury, and M. Nyberg, "Methodology for linked enterprise data quality assessment through information visualizations," *Journal of Industrial Information Integration*, vol. 15, pp. 191–200, Sept. 2019.

[36] J. Michnik and M.-C. Lo, "The assessment of the information quality with the aid of multiple criteria analysis," *European Journal of Operational Research*, vol. 195, pp. 850–856, June 2009.

[37] A. Umar, G. Karabatis, L. Ness, B. Horowitz, and A. Elmagardmid, "Enterprise Data Quality: A Pragmatic Approach," *Information Systems Frontiers*, vol. 1, pp. 279–301, Oct. 1999.

[38] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Computing Surveys*, vol. 41, pp. 1–52, July 2009.

[39] A. Even and G. Shankaranarayanan, "Utility-driven assessment of data quality," *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, vol. 38, pp. 75–93, May 2007.

[40] S. Schelter, D. Lange, P. Schmidt, M. Celikel, F. Biessmann, and A. Grafberger, "Automating large-scale data quality verification," *Proceedings of the VLDB Endowment*, vol. 11, pp. 1781–1794, Aug. 2018.

[41] R. Blake and P. Mangiameli, "The Effects and Interactions of Data Quality and Problem Complexity on Classification," *Journal of Data and Information Quality*, vol. 2, pp. 8:1–8:28, Feb. 2011.

[42] B. Piprani and D. Ernst, "A Model for Data Quality Assessment," in *On the Move to Meaningful Internet Systems: OTM 2008 Workshops* (R. Meersman, Z. Tari, and P. Herrero, eds.), Lecture Notes in Computer Science, (Berlin, Heidelberg), pp. 750–759, Springer, 2008.

[43] M. A. Hossain, P. K. Atrey, and A. E. Saddik, "Modeling and assessing quality of information in multisensor multimedia monitoring systems," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 7, pp. 3:1–3:30, Feb. 2011.

[44] W. Fan, "Data Quality: From Theory to Practice," *ACM SIGMOD Record*, vol. 44, pp. 7–18, Dec. 2015.

[45] R. S. Kenett, A. Zonnenshain, and G. Fortuna, "A road map for applied data sciences supporting sustainability in advanced manufacturing: the information quality dimensions," *Procedia Manufacturing*, vol. 21, pp. 141–148, Jan. 2018.

[46] C. Jean-Quartier, M. Rey Mazón, M. Lovrić, and S. Stryeck, "Collaborative data use between private and public stakeholdersmdash;a regional case study," *Data*, vol. 7, no. 2, 2022.

[47] S. I. H. Shah, V. Peristeras, and I. Magnisalis, "DaLiF: a data lifecycle framework for data-driven governments," *Journal of Big Data*, vol. 8, p. 89, June 2021.

[48] W. Gan, J. C.-W. Lin, H.-C. Chao, and J. Zhan, "Data mining in distributed environment: a survey," *WIREs Data Mining and Knowledge Discovery*, vol. 7, no. 6, p. e1216, 2017.

[49] A. Jonathan, M. Uluyol, A. Chandra, and J. Weissman, "Ensuring reliability in geo-distributed edge cloud," in *2017 Resilience Week (RWS)*, pp. 127–132, Sept. 2017.

[50] "Directive (EU) 2016/1148 of the European Parliament and of the Council of 6 July 2016 concerning measures for a high common level of security of network and information systems across the Union," July 2016.

[51] *European Union, Human rights, EU regulation, Journal officiel, 2016-05-04, n° L119, pp. 1-88*. Apr. 2016.

[52] K. D. Foote, "Data Quality Dimensions," Feb. 2022.

[53] F. Moyano, C. Fernandez-Gago, and J. Lopez, "A Conceptual Framework for Trust Models," in *Trust, Privacy and Security in Digital Business* (S. Fischer-Hübner, S. Katsikas, and G. Quirchmayr, eds.), Lecture Notes in Computer Science, (Berlin, Heidelberg), pp. 93–104, Springer, 2012.

[54] P. H. K. F. H. Y. K. Z. H. L. Y. Yang, "A Collaborative Auditing Blockchain for Trustworthy Data Integrity in Cloud Storage System," *in IEEE Access*, pp. pp. 94780–94794, Nov. 2022.

[55] W. Ahmed and Y. W. Wu, "A survey on reliability in distributed systems," *Journal of Computer and System Sciences*, vol. 79, pp. 1243–1255, Dec. 2013.

[56] G. Akimova, A. Solovyev, and I. Tarkhanov, "Reliability assessment method for geographically distributed information systems," in *2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT)*, pp. 1–4, Oct. 2018. ISSN: 2472-8586.