IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 23, 2019, accepted February 11, 2019, date of publication February 15, 2019, date of current version March 7, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2899751

An Overview of Data Quality Frameworks

CORINNA CICHY[1,2] AND STEFAN RASS[1]
[1] Alpen-Adria-Universität Klagenfurt, 9020 Klagenfurt, Austria
[2] Volkswagen Bank GmbH, 38112 Braunschweig, Germany

Corresponding author: Stefan Rass (stefan.rass@aau.at)

ABSTRACT Nowadays, the importance of achieving and maintaining a high standard of data quality is widely recognized by both practitioners and researchers. Based on its impact on businesses, the quality of data is commonly viewed as a valuable asset. The literature comprises various techniques for defining, assessing, and improving data quality. However, requirements for data and their quality vary between organizations. Due to this variety, choosing suitable methods that are advantageous for the data quality of an organization or in a particular context can be challenging. This paper surveys data quality frameworks in a comparative way regarding the definition, assessment, and improvement of data quality with a focus on methodologies that are applicable in a wide range of business environments. To aid the decision process concerning the suitability of these methods, we further provide a decision guide to data quality frameworks. This guidance aims to help narrow down possible choices for data quality methodologies based on a number of specified criteria.

INDEX TERMS Data quality assessment, data structures, decision making, information management, quality management.

I. INTRODUCTION

In many cases, business decisions strongly rely on some proportions of data that are available for the organization. Thus, a high standard of data quality plays an important role for most businesses. Also, regulations issued by supervisory authorities and directed at data management departments further emphasize the significance of improving data aggregation capacities within an organization. Low levels of data quality can have far-reaching consequences for a business, such as poor decision-making and missed business opportunities, since the data might not provide a clear picture of the circumstances [42], [56], [63]. Relying on manual data collection often causes a threat to the quality of data [7], [41], [73]. Moreover, increased costs can occur when data have to be corrected at some point in the process. A recent Gartner study claims that for businesses, poor data quality leads to an average loss of around $15 million [71]. A 2016 IBM research estimates that in the U. S., the total annual costs resulting from poor data quality are larger than 3 Trillion US dollars [82]. Apart from financial costs, low levels of data quality also affect the decision-making processes within an organization. The KPMG 2017 Global CEO Outlook has found that 56% of CEOs are worried about the integrity of data quality concerning the data that build the basis for their decisions [54]. Especially in a Big Data environment, new challenges, costs and impacts arise [59], [85], [94] including the value, usability and overall quality of the data [13], [18], [92], [84]. Tasks such as storage, processing of data but also the management of data quality are critical in these environments [68]. Poor data quality can also result in a compliance risk, i.e., when the standard of data quality does not match the expectations from supervisory authorities [63]. For many years, data quality has been considered as a multidimensional concept in the literature [3], [5], [81], [97]. Consequently, its measurement is regarded as a complex process including a number of challenges. Moreover, data warehouses have increased in size and complexity and the number of data sources within an organization has grown in recent years [11], [90]. Thus, it is no surprise that a significant increase in data quality literature can be observed [64], [101]. In addition, data quality can be described as a multidisciplinary problem concerning, for example, topics in computer science, quality control, human factors research and statistics [53]. In [102], data quality research is divided into a number of categories regarding methods and topics with further categories to be expected in the future due to the growth of the research area. The appropriate handling and usage of data within an organization further requires a form of decision strategy.

The associate editor coordinating the review of this manuscript and approving it for publication was Shaojun Wang.

According to [67], this involves steps regarding planning, obtaining, storing and sharing, maintaining, applying, and disposing of data (POSMAD). The POSMAD approach supports the decision-making process regarding data on the strategic, tactical and operational level while recognizing the life cycle of data. Also, data strategy is often influenced by IT strategy that focuses mainly on the use of technology. Reference [26] suggests an approach to dealing with data quality using a strategic approach to system complexity which underlines their dependencies. Based on the significance of achieving and maintaining a collection of high-quality data, it is not surprising that the literature on data quality management comprises a variety of frameworks and methodologies regarding the assessment, and improvement of data quality. Different approaches exist to summarize the state of research in the area. For example, [11] review a few well-known and established methodologies for the assessment and improvement of data quality for different types of data. In the authors' approach, criteria such as the type of data and systems, costs, and data quality dimensions are defined, and the aspects are compared for each of the chosen methods. A classification of data quality methods regarding data quality software tools and specific methods has been performed in [14].

The motivation behind this paper is to provide an overview of complete data quality frameworks that are widely applicable by summarizing and comparing their main components including the data quality definition, assessment, and improvement processes.

This paper proceeds as follows: In Section II, the methods used for literature search and selection are briefly described including an overview of the chosen methodologies as well as a brief summary of the main components of each framework. In Section III, an overview of data quality definitions according to the different frameworks is provided. This is followed by the survey and comparison of assessment methods including measurement types and processes in Section IV. Then, aspects of data quality improvement are considered in Section V, including the considerations of data quality costs and decision strategies. Section VI contains a selection guide for choosing suitable data quality frameworks under given circumstances. Finally, Section VII concludes this paper with a few summarizing notes.

II. SCOPE OF THIS SURVEY

Only those frameworks that apply to data in any field were selected for this survey. Moreover, works that only consider one or a few specific aspects of data quality management such as data stewardship, metadata management, or information systems were excluded. We aim to provide an overview of different comprehensive data quality frameworks to the reader. Thus, one aspect of the selection was that the method includes aspects for each of the identified main steps: Data quality definition, data quality assessment, and data quality improvement. Overall, the chosen frameworks fulfill the following criteria: The framework is generally applicable with regard to the

- context of data,
- information system, and
- type of business.

In addition, the framework should provide

- a definition of relevant data quality attributes,
- data quality assessment steps, and
- data quality improvement steps.

TABLE 1. Overview of frameworks.

Acronym	Name of Methodology	Year	Main Ref.
AIMQ	A Methodology for Information Quality Assessment	2002	[58]
CDQ	Comprehensive Methodology for Data Quality Management	2006	[9]
COLDQ	Cost-effect of Low Data Quality	2001	[61]
DQA	Data Quality Assessment	2002	[79]
DQAF	Data Quality Assessment Framework	2013	[87]
DQPA*	A Data Quality Practical Approach	2009	[33]
HDQM	A Data Quality Methodology for Heterogeneous Data	2011	[8]
HIQM	Hybrid Information Quality Management	2006	[20]
OODA DQ	The Observe-Orient-Decide-Act Methodology for Data Quality	2017	[93]
TBDQ	Task-Based Data Quality Method	2016	[96]
TDQM	Total Data Quality Management	1998	[98]
TIQM**	Total Information Quality Management	1999	[34]

* The acronym DQPA was chosen by the authors of this paper to represent the framework proposed in [33], but unnamed there.
** Formerly known as TQdM

Twelve different data quality frameworks in the literature were identified to meet the above criteria, listed in Table 1 and included in this work. Besides these, some frameworks provide techniques for data quality assessment without specifying improvement methods. For example, [72] estimates data quality in databases by adding and calculating quality specifications for each relation instance. In [77], control matrices are used while [48] proposes the use of control chart methods for data quality assessment. Data quality is modeled utilizing artificial neural networks in [57]. Reference [40] defines assessment processes for raw data in databases and information products (IPs) that are used by the consumers. Moreover, business process modeling is used in [2] and [22], and prediction markets are proposed by [78] for approaching data quality. Although these references contain interesting and valuable approaches to data quality assessment, they are not considered complete in the context of this survey. Finally, contributions that propose combinations of existing techniques (e.g., [100]) are not surveyed hereafter.

A. OVERVIEW OF DATA QUALITY FRAMEWORKS

Table 1 presents the generally applicable frameworks according to our aforementioned criteria. The overview contains the short and extended names of the methodology along with the year of publication and the main reference used. From this point onward, the frameworks will be referred to by a respective acronym.

TABLE 2. Overview of methods and their components.

Framework	Main Components
AIMQ	Data quality categorization model, information quality assessment instrument, bench-marking gap analysis, role gap analysis
CDQ	State reconstruction, assessment of data quality dimensions and setting targets, choice of optimal improvement process
COLDQ	Information chain mapping, cost categorization, impact analysis, cost determination, return on investment analysis
DQA	Subjective and objective data quality assessments (metrics and surveys), comparative analysis, root cause analysis, actions for improvement
DQPA	Identification of data quality problems, identification of relevant data, business rule development, data quality assessment, business impact determination, data cleansing, data quality monitoring
DQAF	Initial One-time Assessment, automated process controls, In-line measurement, periodic measurement
HDQM	State reconstruction, quantitative evaluation of data quality problems, selection of appropriate improvement activities
HIQM	Data quality definition, data quality evaluation, data quality monitoring and recovery support
OODA DQ	Iterative process: Observe-Orient-Decide-Act Cycle based on [15]
TBDQ	Planning and evaluating assessment, evolution and execution of improvement
TDQM	TDQM Cycle: Define, Measure, Analyze and Improve. Focus: Information product
TIQM	Establish data quality environment, assess data definition and architecture quality, assess data quality, measure non-quality data costs, re-engineer and cleanse data, improve data process quality

Each methodology consists of a number of steps regarding the definition, assessment and improvement of data quality. These steps will be examined and compared in greater detail in Sections III to V. For an initial overview of the contents of each of the frameworks; Table 2 shows the main components in the form of steps and phases.

B. SPECIAL PURPOSE FRAMEWORKS

Apart from the generally applicable frameworks, there are a number of methods that have been developed for a special purpose. For example, some frameworks are only relevant for data within a specific field, such as census, healthcare or financial data (e.g., the ISTAT [38], [46], the CIHI [60] and the QAFD [28], respectively). Another well-studied point of interest is the quality of Web data (e.g., the IQM [36] or the PDQM [21]). Moreover, the DaQuinCIS methodology (see [86]) focuses only on Cooperative Information Systems while some frameworks are directed exclusively at quality within data warehouses (e.g., the proDQM [45] or the DWQ [47]). These methods are excluded from this survey based on their lack of generality. The well-known method of Su and Jin [91] only takes into consideration product data quality in the context of manufacturing businesses. Moreover, there exist frameworks such as the MAMD by [23] and the MMPRO [17] that are developed purely based on ISO standards. Table 3 shows a number of these special purpose frameworks together with their main focus and application. The reader who is interested in data quality in the context of one of these fields of applications may also refer to the works directly in addition to the generally applicable methods that are surveyed in this paper. Although Table 3 only shows one well-known data quality framework representing healthcare environment, a vast number of further frameworks exist in this context. Chen *et al.* [24] provide an extensive review of this type of frameworks. The management of data quality in Big Data environments also includes aspects that exceed the purpose of this survey. For an overview of data quality challenges with emphasis on Big Data quality, we refer to [49].

TABLE 3. Overview of special purpose data quality frameworks.

Framework	Purpose	Main Ref.
AMEQ	Manufacturing Product Data	[91]
CIHI	Healthcare Data	[60]
DaQuinCIS	Cooperate Information Systems	[86]
DWQ	Data Warehouse Quality	[47]
IQM	Web Data	[35]
ISTAT	Census Data	[46]
MAMD	ISO Standards	[23]
MMPRO	ISO Standards	[17]
ORME-DQ	Operational Risk Management	[12]
PDQM	Web Data	[21]
proDQM	Data Warehouse Quality	[45]
QAFD	Financial Data	[28]
UDQA	Utility-Driven Data Quality Assessment	[37]

III. DATA QUALITY DEFINITION

Data can be defined as real-world objects, with the ability of storing, retrieving and elaborating through a software process and communicating via a network [10]. The way data and data quality are defined in the early development stages of a methodology is an important aspect that varies across the literature. This includes the context, nature, and type of data. Moreover, the data quality attributes or dimensions that are chosen to be relevant for the chosen data type and context have significant impacts on the whole methodology. These aspects are described and compared in this section.

A. TYPES AND NATURE OF DATA

The nature of data or the data type can be defined and categorized in different ways. In [11] and [89], two different classifications are mentioned. The first definition is based on the concept of manufacturing products and categorizes data into the three types: raw data item, component data item and information product. Table 4 shows this so-called orthogonal classification as proposed by [88].

The second way to classify data is the separation into structured data, semi-structured data, and unstructured data. This classification as well as methods on how to extract the information correctly from the different types of data is found in the literature (e.g., [1], [16], [19]). Table 5 describes and exemplifies this.

The heterogeneous nature of data depending on the domain is also recognized in [69], where data structure and

TABLE 4. **Data types.**

Data Type	Description
Raw Data Item	Smaller data units which are used to create information and component data items
Component Data Item	Data are constructed from raw data items and stored temporarily until final product is manufactured
Information Product	Data which are the consequence of performing manufacturing activity on data

TABLE 5. **Data structure.**

Data Structure	Description	Example
Structured Data	Generalization or aggregation of items described by elementary attributes defined within a domain items	Relational data tables
Semi-Structured Data	Data that have a structure with some degree of flexibility	Web page, XML file
Unstructured Data	A generic sequence of symbols, typically coded in natural language	Email text

type are classified into more specific categories such as time-continuous data and event based data. In [95], enterprise data are classified into the three categories: master data, transactional data and historical data. However, the categorization shown in Table 5 and the depiction of data quality as an information product seem to be more widely recognized.

The vast majority of methodologies consider mainly structured data, either by specification in preliminary definitions or indirectly, by referring to examples of structured data such as relational tables in their application. Semi-structured data that contain a degree of flexibility are also either explicitly or implicitly considered. Exceptions are the DQPA and the TBDQ, which only consider structured data but whose inventors envision to extend the models by an investigation of data quality for semi-structured and, in the case of the TBDQ, also unstructured data for future work. Unstructured data seem to pose a great challenge since many techniques for assessing structured and semi-structured data cannot be applied to unstructured data. The HDQM is one of the very few methodologies that consider structured, semi-structured as well as unstructured data explicitly. In particular, the HDQM can be seen as an extension of the CDQ that incorporates heterogeneous data. To do so, the model translates the different types of data resulting from heterogeneous resources into a common, conceptual representation. Although not explicitly stated, the measurement techniques used in the AIMQ may apply to both structured and unstructured data. In the TDQM, while data are viewed as an information product which relates to the classification by [88], the structure of data is not specified.

B. DIMENSIONS AND CLASSIFICATIONS

Data Dimensions are attributes of data quality that can, when measured correctly, indicate the overall quality level of data. The identification of relevant quality dimensions can be seen as a starting point to the subsequent assessment phase and builds the basis for various improvement activities [76]. Dimensions are highly context dependent and their relevancy and importance can vary between organizations and types of data. Depending on their focus, distinct frameworks recognize different attributes for data quality within their methodology. This section provides an overview of the chosen data quality dimensions for each selected work. Apart from the differences in choosing the attributes, their definition may also vary. A basis for the definitions of data quality in this paper is the work of [99]. An extensive survey on data quality dimensions that takes a look at different definitions for the referenced dimensions can be found in [89]. Table 6 shows the data quality dimensions that are explicitly mentioned in the frameworks. It further shows whether an extension to further quality aspects is generally supported in the methodology.

There is a relatively high variation in data quality dimensions considered per framework. Some attributes appear very frequently, while there exist some data quality dimensions that are only recognized by one framework. For a better overview of frequencies, Figure 1 shows the number of occurrences of data quality dimensions (if larger than one) based on the selected frameworks.

The most common dimensions are completeness, timeliness, and accuracy, followed by consistency and accessibility. The definitions of these dimensions according to [99] are as follows:

- **Completeness**: The extent to which data are of sufficient breadth, depth and scope for the task at hand.
- **Accuracy**: The extent to which data are correct, reliable and certified.
- **Timeliness**: The extent to which the age of the data is appropriate for the task at hand.
- **Consistency**: The extent to which data are presented in the same format and compatible with previous data.
- **Accessibility**: The extent to which information is available, or easily and quickly retrievable.

A number of methodologies strongly rely on a study by [99], in which an extensive survey was performed to identify those aspects of data quality that are most important to the data consumer. This includes the TDQM that views data as an information product based on the methodology of the TQM [75], a total quality management methodology for manufacturing products. It is noted that even though this adaption is useful, there is a limitation to the similarity between manufacturing and information products. In the TDQM, the information product is defined as the output of an information manufacturing system. In the first step of the TDQM cycle, data (here information product) characteristics are defined. These characteristics are basic units such as client accounts in a client account database and components of the database, for example in form of an entity-relationship model. Moreover, the IP characteristics include functionalities to the data consumer. Their expectations together with those of suppliers, manufacturers and managers are then specified in form

TABLE 6. Data quality dimensions overview.

Framework	Data Quality Dimensions	Flexible?
AIMQ	Accessibility, Appropriate Amount, believability, completeness, concise representation, consistent representation, ease of operation, free-of-error, interpretability, objectivity, relevancy, reputation, security, timeliness, understandability	no
CDQ	Structured: Accuracy, completeness, currency, Unstructured: Currency, relevance, reliability	yes
COLDQ	Data model: Clarity of definition, comprehensiveness, flexibility, robustness, essentialness, attribute granularity, precision of domains, homogeneity, naturalness, identifiability, obtainability, relevance, simplicity, semantic and structural consistency. Data Values: Accuracy, completeness, consistency, currency, null values, timeliness. Information Policy: Accessibility, metadata, privacy, redundancy, security, unit cost. Presentation: Appropriateness, correct interpretation, flexibility, format precision, portability, consistent representation, representation of null values, use of storage	no
DQA	Accessibility, appropriate amount of data, objectivity, believability, reputation, security, relevancy, value-added, timeliness, completeness, interpretability, ease of manipulation, understandability, concise representation, consistent representation, free-of-error	yes
DQAF	Completeness, timeliness, validity, consistency, integrity	no
DQPA	Accuracy, completeness, consistency, currency, timeliness, uniqueness, volatility	no
HDQM	Accuracy, currency	no
HIQM	Accuracy, completeness, consistency, timeliness	yes
OODA DQ	Speed, volume	no
TBDQ	Accuracy, completeness, consistency, timeliness	yes
TDQM	Accuracy, objectivity, believability, reputation, access, security, relevancy, value-added, timeliness, completeness, amount of data, interpretability, ease of understanding, concise representation, consistent representation	no
TIQM	Definition conformance, completeness, validity (business rule conformance), accuracy (to surrogate source/to reality), precision, non-duplication, equivalence of redundant or distributed data, accessibility, timeliness, contextual clarity, derivation integrity, usability, rightness (fact completeness)	no

of data quality dimensions. The importance of the different attributes across roles is assessed by a survey tool. The 16 data quality dimensions are classified into the four categories: Intrinsic, accessibility, contextual and representational. Similarly, the DQA makes use of a survey tool for assessing data quality dimensions from different perspectives. The AIMQ also recognizes the data quality dimensions and classifications that are proposed by [99]. However, the authors suggest an interesting categorization that differs from the existing ones. Instead of the previously mentioned categories, the AIMQ considers the four types: Sound, dependable, useful, and usable information. In particular, the dimensions are classified into four quadrants as it is shown in Table 7.

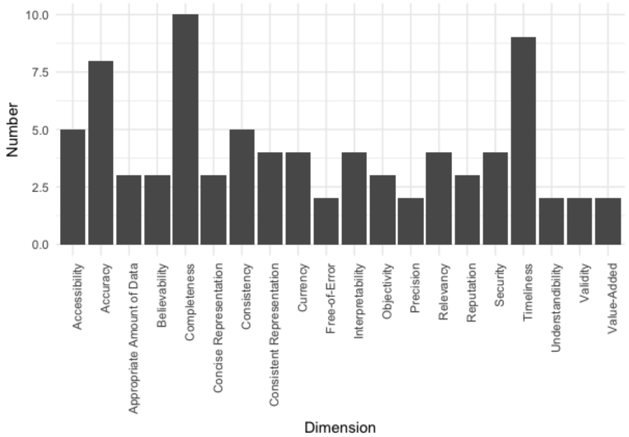

FIGURE 1. Number of frameworks using certain data dimensions.

TABLE 7. Data quality dimensions according to the AIMQ.

	Conforms to specifications	Meets or exceeds consumer expectations
Product Quality	Sound Information	Useful Information
Service Quality	Dependable Information	Usable Information

For example, the accessibility dimension is considered as a service quality and can be assessed by consumer expectations while the completeness dimension is a product quality that can be assessed by specifications. For more information on this model of data quality dimensions, we refer to [50], [51], and [58]. According to the authors, this classification is helpful later on in the process when deciding which aspects of data quality require improvement.

In the COLDQ methodology, the dimensions defined in [81] and [99] are recognized but modified and supplemented by further dimensions and categories. The dimensions are classified into the five categories: data model, data values, information domains, data presentation and information policy. This is somewhat different from most categorizations utilized by other frameworks.

As opposed to many other frameworks, the TIQM classifies the data quality dimensions into only two categories: inherent data quality and pragmatic quality. The authors mention a number of attributes such as precision, non-duplication and validity for inherent data quality or timeliness, usability or accessibility for pragmatic information quality. In order to establish relevant data quality dimensions in an organization, the TIQM suggests a survey of data quality expectations including expectations of current and prospective data warehouse consumers as well as knowledge workers on an operational level.

The data quality definition phase according to the HIQM is an extensive phase consisting of three separate steps. It starts with an information environment analysis, in which knowledge on data sources, processes, and stakeholders is gathered. This is followed by the resource management steps, where resources within the data quality program are defined according to the management plan. The most crucial part of the data quality definition based on the HIQM is identifying the data quality requirements by considering the needs of enterprise, supplier, as well as user-end consumers. In the HIQM, this definition is performed using an extensive analysis to model the different preferences and impacts on the business formally. The mathematical methods used include a fuzzy-linguistic approach as well as an analytical hierarchy process method.

In the TBDQ, only a small number of important dimensions is mentioned. However, the model is flexible and extensible to further data quality dimensions. Similar to the TIQM and DQA, it is particularly suggested to identify those dimensions that are most important for the organization, for example using questionnaires. Likewise, while the HDQM only addresses the two dimensions accuracy and currency explicitly, the adaptability to other dimensions in different contexts is suggested. In the DQPA, data quality attributes are based on [29] and are selected by means of expert user judgments and depend on the type of information system as well as the relevant data.

In the CDQ, different data quality dimensions are suggested depending on the structure of the data. The context-independent dimensions accuracy, completeness, and currency are mentioned for structured and semi-structured data [5], [74], [81]. For unstructured data, dimensions such as condition and originality are mentioned, for example to assess the relevance of the data [55]. However, the selection of dimensions based on observed data quality issues in the organization is recommended.

While the previously mentioned frameworks are concerned with both subjective and objective data quality characteristics, in the DQAF, only objective characteristics of data are considered. It is noticeable, that the accuracy dimension is not part of this model, which is unusual in data quality management. This is explained by the statement that accuracy is difficult or even impossible to measure in practice and that it is more useful to try to derive accuracy from the validity dimension.

Overall, the methodologies all recognize the multidimensionality of data quality while the specifications of attributes varies. Although these differences may be reasonable depending on the approach of the method, there appears to be a gap regarding an effective standardization of dimensions [56].

IV. DATA QUALITY ASSESSMENT

The goal of data quality assessment can be defined as the identification of erroneous data elements and the measurement of the impact of various data-driven business processes [65]. The data quality assessment is a crucial process within the management of data quality often comprising a number of different steps and involving several people groups within the organization. The assessment phase includes selecting and defining data quality measurement types that are then applied to the existing data in order to get an indication of how well each dimension is performing. Apart from defining various data quality measurements, the assessment of data quality has to take into account a number of aspects concerning the data process that have an influence on data such as data lineage, i.e., the consideration of the source data, and process by which the data item was produced [27], as well as data aggregation. Thus, depending on the organization, data quality assessment is a complex process. This section provides an overview of the different data quality assessment methods that were chosen by each of the selected works. Firstly, the use of data quality measurement types such as questionnaires or quantitative metrics throughout the frameworks is investigated. In many cases, these measurements make up the foundation for the subsequent assessment process. Secondly, the different phases and steps according to the chosen methodologies are summarized and related to each other in a comparative manner to highlight the differences, similarities and relationships between the frameworks.

A. DATA QUALITY MEASUREMENT TYPES

Data quality can be measured subjectively, for example by asking the data consumer to rate the level of quality of the dimensions. Alternatively, data quality metrics can be defined consisting of computations that can give an indication of the data quality level. The metrics are used to measure dimensions of data quality objectively. In many cases, one metric is not sufficient to accurately measure a data quality dimension, and the key is to combine different metrics to get a clear picture of the overall data quality. Many of these metrics measure the number or percentage of some specified constraints that are being violated, or qualitatively measure the number of erroneous decisions that were made based on the data [44]. It can be argued that some data quality dimensions cannot be assessed by objective measures and that for those dimensions, subjective measures are needed [80]. Depending on the framework, different measurement types are utilized. Table 8 shows an overview of the types of measurements suggested in each of the methodologies.

It can be seen that most methodologies strongly rely on objective data quality metrics. The DQAF shall be particularly emphasized here, as it provides a comprehensive set of objective data quality metrics that the organization can choose from. These measurements are classified into different types, i.e., initial one-time assessment, automated process control, in-line measurement, and periodic measurement.

The TDQM presents some common metrics to measure a subset of the identified data quality dimensions. This includes metrics for accuracy, freeness of error, timeliness, completeness, and consistency. Examples of those metrics are the percentages of incorrect values, an indicator of when data was updated, a percentage of non-existent accounts and the number of records that violate referential integrity. Moreover, the TDQM takes into account that certain business rules need to be considered when assessing data quality.

TABLE 8. Types of measurement.

Framework	Main Components
AIMQ	Subjective assessment: Survey questionnaire
CDQ	User Interviews and definition of data quality metrics for accuracy and currency
COLDQ	Consumer surveys and definition of various data quality metrics
DQA	Stakeholder expectations and definition of quantitative metrics (functional forms)
DQPA	Definition of primary data source and derived data quality metrics
DQAF	Definition of set of data quality metrics for different types of measurement
HDQM	Definition of data quality metrics for accuracy and currency
HIQM	Objective assessment through measurement algorithm suggested
OODA DQ	Not specified
TBDQ	Survey questionnaire and simple ratio
TDQM	Consideration of business rules and definition of data quality metrics
TIQM	User expectations and definition of data quality metrics

For example, the exposure for clients should be monitored. The TDQM further briefly discusses metrics that come from the information manufacturing background, such as security and credibility measures. In addition, the TDQM mentions so-called information-manufacturing-oriented data quality metrics such as measuring unauthorized access to assess the security dimension.

TABLE 9. Functional forms in the DQA.

Functional Form	Description	Dimensions measured
Simple Ratio	Ratio of desired outcomes to total outcomes	free-of-error, completeness, consistency, concise representation, relevancy, ease of manipulation
Min or Max Operation	Minimum or maximum value among normalized individual data quality indicator values	Believability, appropriate amount of data, timeliness, accessibility
Weighted Average	Assigning weighting factors to represent the importance of the variables to the evaluation of a dimension	Believability, appropriate amount of data

In the DQA, the so-called functional forms are presented for the assessment of objective data quality attributes (see Table 9). Moreover, for each of the data quality dimensions, an application of these functional forms as metrics are explained in detail.

In the DQPA, the authors differentiate between metrics for the assessment of primary data sources and assessment of derived data. In particular, the proposed framework extends the metrics proposed by [79] along with those presented in [6] and [72] to metrics that allow the measurement of data quality at different levels of granularity.

Other frameworks only mention the use of data quality metrics briefly. For example, the CDQ mentions the two objective metrics percentage of duplicate objects and that of matching objects to measure accuracy and currency. In the HIQM, no specific metrics are defined, but the need for a measurement algorithm for each data quality dimension is expressed. Similarly, the OODA DQ methodology refers to the use of existing data quality metrics and tools for measurement.

Some frameworks suggest combining objective measures with the use of subjective assessments such as survey questionnaires. The COLDQ suggests a number of objective ways to measure dimensions such as completeness and accuracy but also strongly recommends to survey the data consumer directly in order to measure dimensions related to data presentation. Similarly, the TBDQ performs an initial assessment by means of survey questionnaires followed by the objective assessment using metrics such as a simple ratio. The TIQM suggests identifying the user's expectation in order to decide which objective measures are required.

The AIMQ is the only chosen framework that relies on subjective measurements only. For the purpose of self-assessment, the AIMQ includes an "IQA Instrument", i.e., a questionnaire including several items that help with measuring data quality.

B. DATA QUALITY ASSESSMENT PROCESS

This section provides an overview of the different steps that are suggested in different data quality frameworks for the purpose of assessing the current state of data quality within an organization.

Some frameworks such as the TDQM do not provide formal steps within the assessment process. Here, the assessment phase consists of developing the previously mentioned data quality metrics and implementing them by means of a new information manufacturing system or add them to an existing system. Similarly, the DQA suggests that, in order to measure data quality in practice, the previously developed metrics should be applied to the data but no formal process is provided. Alongside these measurements, the DQA also suggests the use of additional subjective assessments that are then to be compared with the objective measurements. The results fall into one of four quadrants: low subjective and objective assessments, high objective and low subjective assessment, low objective and high subjective assessment, or high subjective and objective assessments. This builds the basis for the subsequent analysis and improvement process. While the OODA DQ methodology does not provide formal steps for this part of the process either, the assessment phase according to this methodology for data quality relates to the first phase of the method's iterative process, namely, the Observe part. In this phase, existing data is observed and data quality issues are identified by means of tools such as regular reports and dashboards. Moreover, a notification service for potential data quality issues as well as feedback from external agencies are suggested.

During the assessment according to the AIMQ, the values determined by means of the defined IQA instrument are aggregated into the four quadrants that were defined in Section III-B. In particular, the mean values of the dimensions of each quadrants are computed in this step. Similarly, in the CDQ, problem identification through interviews and quantitative evaluation of quality issues are performed. In particular, data quality issues are determined by means of user interviews including the identification of the precise part of data and the corresponding processes. The problem identification is followed by applying data quality metrics to the identified dimensions.

The HIQM only mentions the evaluation of data quality in general, based on non-specified metrics. However, the model provides a unique contribution regarding the data quality assessment in form of a data quality monitoring and recovery support method. In particular, the model proposes a warning phase comprising the components diagnoser, feedback modules, message generator, a warning log database, a warning analyzer, a warning/recovery database, as well as a real-time recovery module.

Rather than providing several steps, the assessment process according to the DQAF is based on the different measurement types that were previously defined. This includes the initial one-time assessment, automated process control and in-line as well as periodic measurement.

Other methodologies provide more detailed steps and processes for the execution of data quality assessment. For COLDQ, the process starts with a number of preliminary steps. In particular, data customers are identified, the information chain is mapped and dimensions and appropriate rules are chosen. It follows the measurement of each of the data quality categories that were defined in the previous section. The assessment according to the TIQM is described in great detail and composed of two processes. Firstly, data definitions and the whole information architecture quality are assessed. Secondly, the actual assessment of data quality takes place. Each of the two processes consists of a number of steps that are summarized in Table 10.

The assessment phase of the HDQM consists of two steps. Firstly, as opposed to many other methodologies, it starts by ranking its resources in order to establish feasibility and risk for the subsequent improvement phase. Secondly, the actual quantitative measurement of data quality is performed. In this assessment, the relevant dimensions are measured by applying appropriate metrics. For the accuracy dimension, the HDQM model suggests a specific distance ratio. When dealing with data that are not structured, the authors emphasize that preliminary steps have to be taken in order to be able to use the presented metric. In particular, values have to be related to their domains in semi-structured data sets and data objects have to be extracted and classified when dealing with unstructured data (as proposed in [43] and [66], respectively).

The assessment phase of the TBDQ consists of two steps. Firstly, the goals and scope of data quality for the business are defined in the planning phase. This includes specifying a minimum level of data quality, defining and assigning weight to the dimensions and identifying those tasks within the process that can lead to data quality issues. In the subsequent evaluation step, weights are assigned to the data quality problems by using a pair-wise comparison matrix of the analytical hierarchical process [83]. A questionnaire-based approach is suggested for subjective assessments, combined with a simple ratio metric for objective assessments of data quality. Based on this, the data quality issues receive value. It is also suggested that subjective and objective assessment results are compared as a form of validation.

TABLE 10. Assessment steps of the TIQM.

Process	Step
Assess data definition and information architecture quality	1) Identify data definition quality measures 2) Identify information group to assess 3) Identify the information stakeholders 4) Assess data definition technical quality 5) Assess information architecture and database design quality 6) Assess customer satisfaction with data definition quality
Assess information quality	1) Identify an information group for assessment 2) Establish information quality objectives and measures 3) Identify the information value and cost chain 4) Determine files or processes to assess

The DQPA provides seven different steps for applying data quality assessment. In the first step, useful data quality properties are identified for the assessment. Then, existing metrics are analyzed about their suitability to provide unbiased, user-independent evaluations of data quality aspects. In the third step, methods to represent, interpret and assess data quality indicators are described. The notion of data lineage is regarded as an important aspect of this model and crucial to the process. In the fourth step, quality scores of primary data sources are estimated and stored as metadata. Then, the derived data is assessed in the fifth step. In step six, the data quality is analyzed either by selecting the best data sources before the query execution based on its quality scores or by comparing data quality aggregated scores that correspond to different query plans for the same business question. Finally, in the seventh step, data sources are ranked according to the data quality stores and priorities provided by the user. The DQPA further makes use of a data lineage algorithm with a conflict resolution function for tracing back towards providing more information on the data quality [30]–[32].

The assessment phases are structured differently depending on the framework. However, many similarities can be observed regarding the types of measurements. Most methods rely on objective metrics or a combination of metrics and subjective measurements. Overall, the steps of assessment processes differ significantly, especially in the degree of detail.

V. DATA QUALITY IMPROVEMENT

After assessing the current data quality, in most situations, the goal is to take measures in order to improve its standard. To do so, organizations need to consider different techniques, and tools while taking into account the resulting costs. When measures to improve data quality are considered, this process also involves decision theory. Thus, data quality improvement is often composed of a number of steps. The improvement phase according to the data quality methodologies presented here are summarized and compared in this section.

A. DATA QUALITY IMPROVEMENT PROCESS

As a preliminary step to data quality improvement, some frameworks start by performing an analysis of root causes of data quality issues. For example, the TDQM suggests investigating these issues utilizing methods such as statistical process control and introducing dummy accounts. A further method introduced by [25] is proposed in the TDQM for analyzing the causes of poor data quality assessment results. The improvement process proceeds with identifying areas for improvement and allocating resources accordingly. They refer to the "Information Manufacturing Analysis Matrix" developed by [5] as well as the work of [4] for a methodology. Similar to the TDQM, root cause analysis is also part of the DQA framework. In this methodology, a comparison of the previously performed objective and subjective measurements where discrepancies are identified is suggested in combination with a root cause analysis. The findings should then be processed by taking the necessary measures for data quality improvement. Similarly, in the analysis phase of the HIQM, values obtained for the data quality dimensions in the assessment step are compared to the data quality requirements. This is followed by a both data- and process-oriented improvement phase that also includes modifications at a strategic level, which are planned and performed in a so-called strategy correction step. In the DQAF, comparing results of data quality assessment against assumptions or expectations is emphasized. Moreover, the author presents an overview of possible root causes along with their origin and possible improvement actions that can be useful for data quality improvement projects.

Some methods provide more detailed steps and instructions for the data quality improvement process. According to the TIQM methodology, the improvement phase consists of two main processes: Information Product improvement (reengineer and cleanse data), and improve information process quality (data defect prevention). These two processes are broken down further into smaller steps that are summarized in Table 11. For each step of the processes, the author presents a detailed process description including an exemplary flowchart of activities. Moreover, the model provides a list of inputs, outputs as well as techniques and tools that can be useful in the corresponding part of the improvement phase.

TABLE 11. Improvement steps of the TDQM.

Process	Step
Information product improvement	1) Identify data sources 2) Extract and analyze source data 3) Standardize data 4) Correct and complete data 5) Match and consolidate data 6) Analyze data defects 7) Transform and enhance data into target 8) Calculate derivations and summary data 9) Audit and control data extract, transformation and loading
Improve information process quality	1) Select process for information quality Improvement 2) Develop plan for information quality improvement 3) Implement information quality improvement 4) Check impact of information quality improvement 5) Act to standardize information quality improvements

The model also describes a number of best practices in form of checklists for the categories: Data definition and information architecture, business process and application design, business procedures and data capture and, management and environment.

The COLDQ also provides detailed instructions for the data quality improvement. After the current state of the data quality is assessed, it is suggested to first perform a requirements assessment in which data quality problems are prioritized and their scopes are defined. Moreover, the model gives advice on how to assign responsibilities, to choose a data quality project and to build the corresponding team of people. A number of tools that can be helpful for data quality improvement such as data cleansing and a rules definition system are suggested. The next step is the definition of data quality rules, followed by data mining techniques such as clustering, decision trees and link analysis. Moreover, the methodology proposes that the supplier management process is specified. After these preliminary steps, the actual improvement phase incorporates a solution architecture, static cleansing, integrating and testing rules and rules system as well as building the non-conformance resolution system that generates workflow tasks. Another important step that follows is the measurement of improvement, for example by means of statistical process control.

In the DQPA, a detailed example for data quality improvement steps can be found. The main components of data quality improvement according to the DQPA are business impact determination, data cleansing, and monitoring and assessment of data quality on a regular basis. For the impact analysis, among methods such as cost benefit analysis, usage and anecdotes (told from past events), the prioritization and ranking of data sources and business questions were found to be a crucial aspect. For the subsequent data cleansing process, metadata usage, data profiling, and data matching are suggested. Finally, in order to enhance data quality successfully, standardized improvements and continuous assessment and monitoring are proposed.

The improvement phase of the AIMQ approach focuses strongly on two different gap analysis methods. Prior to these methods, the values of each quadrant that were measured in the assessment step are analyzed and compared to each other. This directly helps with the decision of which quadrant requires improvement. Two important techniques that are mentioned in the model are the bench-marking gap analysis and the role gap analysis. The first technique contains the systematic comparison of performance across organizations in order to evaluate the relative state. In particular, the data quality of an organization is compared with a ‘‘benchmark’’, i.e., a best-practice data quality organization. When large gaps are identified, a possibility for improvement can be deducted for data quality in that area. The role gap analysis evaluates assessments of data quality and compares them across different roles within the organization. This helps understanding the awareness of data quality problems of different people groups. If the gap is large, the information consumers and information systems professionals disagree about the level of data quality. Thus, in that case, they should start by discussing these differences and come to an agreement. Overall, the AIMQ focuses on prioritizing areas for data quality improvement.

As opposed to many other frameworks in which data quality targets are set in the beginning of the process, the CDQ methodology suggest to set these targets after the evaluation of actual data quality values as part of the improvement process. To do this, process-oriented, as well as cost-oriented analyses, are performed. For the improvement of data quality, the HDQM proposes three steps. Similar to the CDQ, the phase starts with an analysis of data quality requirements which in this model is performed using a process-oriented approach [10]. It can be noted that in most frameworks, this step is located at an earlier stage. It follows a selection of activities for data quality improvement. This is done by using both a data-driven and process-driven strategy in order to produce a ReSource/Improvement Activity matrix. Finally, an improvement process is chosen and evaluated based on this matrix. The selected process should incorporate all relevant dimensions and resources.

Just as the assessment phase, the improvement phase according to the TBDQ consists of two steps. Firstly, prioritization of data units is suggested and data improvement tasks such as data correction or notification designed and proposed in the evolution step. The decision on a suitable task is based on an ‘‘award system’’ comparing the different tasks based on execution costs and level of improvements. In the execution steps, the tasks and modified process units are performed. In addition, the execution is analyzed in terms of scope and achieved amount of improvement.

The OODA DQ methodology proposes a rather different approach to structuring the data quality improvement process. This phase of the framework comprises the remaining steps of its iterative cycle, i.e., Orient, Decide and Act. The Orient phase includes a root cause analysis that should be performed by a data governance team as well as the assessment of the severity of the previously identified data quality issues. Decisions ranging from data cleansing to modifications in application systems are the main concern in the Decide phase of the process. The decisions can be on a tactical as well as on an operational level and also include decisions regarding the number of people needed for fixing the issues in an appropriate manner. Finally, the Act phase is where identified actions are performed, implemented and validated.

B. DATA QUALITY COST CONSIDERATIONS

Costs can be defined as ‘‘Resources sacrificed or forgone to achieve a specific objective or the monetary effect of certain actions or lack thereof.’’ [35]

Naturally, there are a number of different types of costs that result from a low standard or data quality but also those that are involved in quality improvement measures. When considering data quality improvement projects or initiatives, it is important that they lead to benefits for the organization. In the work of [35], low data quality costs such as lost opportunity costs, higher maintenance costs or process failure costs as well as data improvement costs such as training costs and infrastructure improvement costs are reviewed and classified. This categorization helps with proving feasibility of new initiatives and with bench-marking, i.e., comparing data quality costs among organizations to set data quality goals. Moreover, cost classifications are particularly important when it comes to assessing the risk resulting from low data quality.

TABLE 12. Costs considerations.

Framework	Budget constraints	Non-quality costs	Improvement costs	Cost classification	Cost-benefit analysis
CDQ	yes	yes	yes	yes	yes
COLDQ	-	yes	yes	yes	yes
DQAF	yes	yes	-	-	yes
DQPA	-	yes	-	-	-
HDQM	yes	yes	yes	-	yes
TBDQ	yes	yes	yes	-	yes
TDQM	yes	-	-	-	-
TIQM	yes	yes	yes	yes	yes

Table 12 shows to which extend data quality costs are incorporated and considered in the frameworks that are surveyed in this paper. This list excludes AIMQ, DQA, HIQM and OODA DQ, since these do not consider costs explicitly.

A cost-benefit analysis (CBA) often builds the basis for decision-making processes in an organization. The CBA can be defined as a process in which benefits and costs of a project are compared systematically and analytically in order to assess its value [70].

Some frameworks propose specified methods to deal with data quality costs. For example, in the TBDQ, a cost-benefit analysis is considered from a qualitative perspective. The model uses an award system in order to choose the improvement tasks. The processes are evaluated on the basis of their execution costs and the level of improvement. This is done by means of a Time-Driven Activity-Based Costing (TDABC) as proposed in [52]. In the TDQM, the problem of allocating resources appropriately is emphasized. An integer programming model [5] is proposed for the purpose of maximizing the improvement under certain budget constraints. Costs play an important role throughout the steps of the CDQ. Non-quality costs (potential savings) are examined in detail and compared to quality costs, i.e., improvement costs. The HDQM provides a novel approach to data quality cost considerations. Apart from stating quantitative methods to cost-benefit analysis (e.g., [62]), the model proposes to qualitatively compare costs with benefits. First, a ReSource/Activity matrix is used to identify candidate improvement processes followed by an evaluation of costs for each of the candidate processes. The qualitative approach consists of categorizing costs (very low, low, medium, high, very high) and subsequently, comparing the values along with their effects on dimensions (data quality dimension/cost ratio) in order to find the appropriate improvement process. The TIQM provides a detailed classification of data quality costs. Non-quality costs consist of process failure costs, information scrap and rework costs, as well as missed opportunity costs. Assessment costs arise from the data quality assessment processes, including software as well as labor costs. Finally, improvement costs result from improving and maintaining data quality. The framework provides many cost examples along with measurement methods. The COLDQ also provides a detailed classification of data quality costs in which the costs are classified into impacts on the operational, the tactical and the strategical level. The operational impacts include various types of costs such as detection, correction and prevention costs. The tactical and strategical impacts consist of costs regarding lost opportunities, delays and organizational mistrust. As a basis for the cost-benefit analysis, the COLDQ suggests the evaluation of return on investments (ROI) which directly relates investments to profit [39]. This is useful when justifying the implementation of improvement projects or activities.

Other methods recognize the significance of data quality costs without establishing concrete methods to approach them within the framework. The DQPA underlines the importance of data quality prevention, correction costs as well as cost effectiveness. However, the model itself only incorporates non-quality costs, i.e., impacts on the business. For future work, an evaluation of data quality costs comprising prevention and correction costs is planned.

C. IMPROVEMENT DECISION STRATEGIES

It is observed that many frameworks emphasize the significance of having a decision strategies in place when it comes to choosing the objects of improvement. Among other aspects of the data quality process, the improvement decisions strongly depend on the required costs that were discussed in Section V-B. Table 13 shows the strategies and methods explicitly mentioned in the frameworks that support the improvement decision processes.

TABLE 13. Improvement decision strategies.

Framework	Method
AIMQ	Based on the data quality categorization and gap analysis
CDQ	Based on the state reconstruction phase
COLDQ	Data Quality Scorecard to identify best opportunities for improvement
DQAF	Definition of directives for data quality strategy
DQPA	Ranking of business questions and data sources to determine impact and required enhancements
HDQM	ReSource-Improvement-Activity Matrix, cost-benefit analysis
TBDQ	Prioritization of data units based on weight and measurement score, award system
TDQM	Information manufacturing analysis Matrix [4] and integer programming model [5]

VI. DATA QUALITY FRAMEWORK SELECTION

The previous sections have shown that the different data quality frameworks use different methods during the process of quality assessment and improvement. This section provides a decision guide that can help selecting an appropriate data quality framework for a given situation in a systematic way.

To start, two preliminary questions are presented which should be considered carefully before following the subsequent decision guide, presented in Table 14.

1) What are the user's general requirements? If the user is interested in finding a methodology, that supports the definition, assessment and improvement processes within an organization in a comprehensive way, the twelve frameworks presented in this paper constitute possible options. If the user seeks tools or software to support aspects of data quality management such as data profiling or data validation, digging into more specialized literature on these techniques can be beneficial; a summary of which reference [14] provides.
2) What is the context of the data? If the data of interest belong to one of the specialized and well-studied areas of research mentioned in Table 3, we also refer to the relevant literature suggested for special purpose frameworks (see Section II-B). Otherwise, the twelve frameworks surveyed in this paper are the ones applicable in any chosen context (besides further ones that are more specialized or tailored to specific applications and hence excluded from our treatment here).

Table 14 is the decision guide, designed in order to narrow down the set of candidates out of the twelve surveyed data quality frameworks, based on the application at hand. Table 14 shows the differences between the methodologies based on a number of identified key criteria (similar to a decision tree), on which a systematic questionnaire shown

TABLE 14. Decision guidance. (a) General comparative overview. (b) Decision tree questions.

(a) General Comparative Overview

Framework	Number of Dimensions	Identification of dimensions supported	Can handle structured data?	Can handle semi-structured data?	Can handle unstructured data?	Metrics specified	Subjective Assessment supported	Costs considered
AIMQ	15	-	(yes)	(yes)	(yes)	-	yes	-
CDQ	5	yes	yes	yes	partially	yes	yes	BC, NQC, QIC, CC, CBA
COLDQ	35	-	yes	(yes)	(yes)	yes	partially	NQC, QIC, CC, CBA
DQA	16	yes	yes	-	-	yes	yes	-
DQAF	5	-	yes	-	-	yes	-	BC, NQC, CBA
DQPA	7	-	yes	-	-	yes	-	NQC
HDQM	2	-	yes	yes	yes	yes	-	BC, NQC, QIC, CBA
HIQM	4	yes	yes	(yes)	-	-	-	-
OODA DQ	2	-	yes	(yes)	(yes)	-	-	-
TBDQ	4	yes	yes	-	-	yes	yes	BC, NQC, QIC, CBA
TDQM	15	-	yes	(yes)	(yes)	yes	-	BC
TIQM	13	-	yes	(yes)	(yes)	yes	yes	BC, NQC, QIC, CC, CBA

BC = Budget Constraints, NQC = Non-quality Costs, QIC = Quality Improvement Costs, CC = Cost Classification, CBA = Cost-Benefit Analysis.
(yes) denotes that the aspect is not explicitly mentioned in the reference but can be inferred.

(b) Decision Tree Questions

Question	Answer	Relevant Column in Table 14a	Instruction
What is the structure of the data?	Structured	Can handle structured data?	If "yes" then framework is applicable
	Semi-structured	Can handle semi-structured data?	If "yes" then framework is applicable
	Unstructured	Can handle unstructured data?	If "yes" then framework is applicable
Which dimensions are relevant?	Subset of dimensions specified in one or more frameworks	-	Choose a framework that contains relevant dimensions
	Set of dimensions that is not specified in any framework	Identification of dimensions supported	If "yes" then framework is applicable
	Not yet known	Identification of dimensions supported	If "yes" then framework is applicable
What type of measurements are preferred?	Objective metrics	Metrics specified	If "yes" then framework is applicable
	Subjective assessments	Subjective assessment supported	If "yes" then framework is applicable
To what extent should costs be considered?	Analysis of non-quality costs	Costs considered	If entry includes "NQC" then framework is applicable
	Analysis of improvement costs		If entry includes "QIC" then framework is applicable
	Detailed Cost classification		If entry includes "CC" then framework is applicable
	Cost-benefit analysis		If entry includes "CBA" then framework is applicable

in Table 14a towards a final selection of a method can be conducted.

VII. CONCLUSION

In this survey, twelve general-purpose applicable data quality frameworks that contain data quality definitions, assessment and improvement processes were systematically surveyed and compared. As opposed to many other frameworks, the selected works are generally applicable in most circumstances in practice. A variation in data quality definitions was observed since the frameworks all chose different data quality dimensions to be relevant. However, most frameworks recognize that the relevance of the dimensions should be assessed individually by the organization. Nevertheless, completeness, timeliness and accuracy appear to be the most important quality attributes. Most frameworks focus on structured and semi-structured data, while few works can also handle unstructured data. Moreover, the assessment processes vary strongly in methods and complexity. Many authors suggest the use of objective metrics or a combination of metrics and subjective measurements while one frameworks relies solely on subjective assessment. In this survey, improvement processes were compared with special emphasis on costs and decision strategies. While most frameworks recognize the importance of considering non-quality as well as improvement costs, the depth of considerations ranges from simply recognizing budget constraints to detailed cost-benefit analyses. The majority of frameworks provide methods that help the improvement decision process. Here, different approaches were observed.

This paper further provided a decision guide that can help the reader to identify the most suitable framework(s) among the presented works. The selection is based on a number of key aspects that are written as questions and that can, in the same manner as a decision tree, narrow down the choices to the suitable frameworks for a given situation.

Possible future research directions include more comprehensible prediction of impacts of poor data quality which is directly related to the mentioned non-quality costs. Moreover, there appears to be a lack in research concerning the impacts and interactions of data quality dimensions on data quality and thus, regulatory compliance with a sophisticated, statistical basis.

REFERENCES

[1] S. Abiteboul, P. Buneman, and D. Suciu, *Data on the Web: From Relations to Semistructured Data and XML*. San Francisco, CA, USA: Morgan Kaufmann, 2000.

[2] S. Bagchi, B. Xue, and J. Kalagnanam, "Data quality management using business process modeling," in *Proc. IEEE Int. Conf. Services Comput. (SCC)*, Chicago, IL, USA, Sep. 2006, pp. 398–405.

[3] D. P. Ballou and H. L. Pazer, "Modeling data and process quality in multi-input, multi-output information systems," *Manage. Sci.*, vol. 31, no. 2, pp. 150–162, 1985.

[4] D. P. Ballou and G. K. Tayi, "Methodology for allocating resources for data quality enhancement," *Commun. ACM*, vol. 32, no. 3, pp. 320–329, 1989.

[5] D. Ballou, R. Wang, H. Pazer, and G. K. Tayi, "Modeling information manufacturing systems to determine information product quality," *Manage. Sci.*, vol. 44, no. 4, pp. 462–484, 1998.

[6] D. P. Ballou and G. K. Tayi, "Examining data quality," *Commun. ACM*, vol. 41, no. 2, pp. 54–57, 1998.

[7] K. A. Barchard and L. A. Pace, "Preventing human error: The impact of data entry methods on data accuracy and statistical results," *Comput. Hum. Behav.*, vol. 27, no. 5, pp. 1834–1839, 2011.

[8] C. Batini, D. Barone, F. Cabitza, and S. Grega, "A data quality methodology for heterogeneous data," *Int. J. Database Manage. Syst. (IJDMS)*, vol. 3, no. 11, pp. 60–79, 2011.

[9] C. Batini, F. Cabitza, C. Cappiello, C. Francalanci, and P. di Milano, "A comprehensive data quality methodology for Web and structured data," in *Proc. 1st Int. Conf. Digit. Inf. Manage.*, Bangalore, India, Dec. 2006, pp. 448–456.

[10] C. Batini and M. Scannapieco, *Data Quality: Concepts, Methodologies and Techniques*. New York, NY, USA: Springer-Verlag, 2006.

[11] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Comput. Surv.*, vol. 41, no. 3, p. 16, 2009.

[12] C. Batini, D. Barone, M. Mastrella, A. Maurino, and C. Ruffini, "A framework and a methodology for data quality assessment and monitoring," in *Proc. MIT ICIQ*, Cambridge, MA, USA, 2007, pp. 333–346.

[13] C. Batini, A. Rula, M. Scannapieco, and G. Viscusi, "Data quality to big data quality," *J. Database Manage.*, vol. 26, no. 1, pp. 60–82, 2015.

[14] A. Borek, P. Woodall, M. Oberhofer, and A. K. Parlikad, "A classification of data quality assessment methods," in *Proc. MIT ICIQ*, Adelaide, Australia, 2011, pp. 189–203.

[15] J. Boyd, "A discourse on winning and losing," Maxwell Air Force Base, Air Univ., Tech. Rep. M-U 43947, 1987.

[16] P. Buneman, "Semi-structured data," in *Proc. ACM PODS*, 1997, pp. 117–121.

[17] I. Caballero, E. Verbo, C. Calero, and M. Piattini, "MMPRO: A methodology based on ISO/IEC 15939 to draw up data quality measurement processes," in *Proc. MIT ICIQ*, Cambridge, MA, USA, 2008, pp. 326–340.

[18] L. Cai and Y. Zhu, "The challenges of data quality and data quality assessment in the big data era," *Data Sci. J.*, vol. 14, no. 2, p. 2, 2015.

[19] D. Calvanese, D. Giacomo, and M. Lenzerini, "Modeling and querying semi-structured data," *Netw. Inf. Syst. J.*, vol. 2, no. 2, p. 253–273, 1999.

[20] C. Cappiello, P. Ficiaro, and B. Pernici, "HIQM: A methodology for information quality monitoring, measurement, and improvement," in *Advances in Conceptual Modeling—Theory and Practice*, J. F. Roddick, Ed. Berlin, Germany: Springer, 2006, pp. 339–351.

[21] A. Caro, C. Calero, and M. Piattini, "A portal data quality model for users and developers," in *Proc. MIT ICIQ*, Cambridge, MA, USA, 2007, pp. 462–476.

[22] A. Caro, A. Rodriguez, C. L. Yonke, C. Cappiello, and I. Caballero, "Designing business processes able to satisfy data quality requirements," in *Proc. MIT ICIQ*, Paris, France, 2012, pp. 31–45.

[23] A. G. Carretero, A. Freitas, R. Cruz-Correia, and M. Piattini, "A case study on assessing the organizational maturity of data management, data quality management and data governance by means of MAMD,' in *Proc. MIT ICIQ*, Ciudad Real, Spain, 2016, pp. 75–84.

[24] H. Chen, D. Hailey, N. Wang, and P. Yu, "A review of data quality assessment methods for public health information systems," *Int. J. Environ. Res. Public Health*, vol. 11, no. 5, p. 5170–5207, 2014.

[25] D. J. Corey, L. Cobler, T. Lonsdale, and K. Haynes, "Data quality assurance activities in the military health services system," in *Proc. MIT ICIQ*, Cambridge, MA, USA, 1996, pp. 127–153.

[26] N. Couture, "Reducing data management complexity in the enterprise," *Bus. Intell. J.*, vol. 17, no. 1, pp. 31–37, 2012.

[27] Y. Cui, J. Widom, and J. L. Wiener, "Tracing the lineage of view data in a warehousing environment," Database Group, Stanford Univ., Stanford, CA, USA, Tech. Rep., Nov. 1997. [Online]. Available: http://www-db.stanford.edu/pub/papers/lineage-full.ps

[28] F. D. Amicis and C. Batini, "A methodology for data quality assessment on financial data," *Stud. Commun. Sci.*, vol. 4, no. 2, pp. 115–137, 2004.

[29] M. del Pilar Angeles and L. M. MacKinnon, "Detection and resolution of data inconsistencies, and data integration using data quality criteria," in *Proc. Qual. Inf. Commun. Technol.*, Porto, Portugal, 2004, pp. 87–94.

[30] M. del Pilar Angeles and L. M. MacKinnon, "Tracking data provenance with a shared metadata" in *Proc. Postgraduate Res. Conf. Electron., Photon., Commun. Netw.*, Lancaster, U.K., 2005, pp. 120–121.

[31] M. del Pilar Angeles and L. M. MacKinnon, "Quality measurement and assessment models including data provenance to grade data sources," in *Proc. Int. Conf. Comput. Sci. Inf. Syst.*, Athens, Greece, 2005, pp. 101–118.

[32] M. del Pilar Angeles and F. García-Ugalde, "Assessing quality of derived non atomic data by considering conflict resolution function," in *Proc. 1st Int. Conf. Adv. Databases, Knowl., Data Appl.*, Cancún, Mexico, Mar. 2009, pp. 81–86.

[33] M. Del Pilar Angeles and F. García-Ugalde, "A data quality practical approach," *Int. J. Adv. Softw.*, vol. 2, no. 2, pp. 259–274, 2009.

[34] L. P. English, *Improving Data Warehouse and Business Information Quality*. New York, NY, USA: Wiley, 1999.

[35] M. Eppler and M. Helfert, "A classification and analysis of data quality costs," in *Proc. MIT ICIQ*, Cambridge, MA, USA, 2004, pp. 311–325.

[36] M. J. Eppler and P. Muenzenmayer, "Measuring information quality in the Web context: A survey of state-of-the-art instruments and an application methodology," in *Proc. MIT ICIQ*, Cambridge, MA, USA, 2002, pp. 187–196.

[37] A. Even and G. Shankaranarayanan, "Understanding impartial versus utility-driven quality assessment in large datasets," in *Proc. MIT ICIQ*, Cambridge, MA, USA, 2007, pp. 265–279.

[38] P. D. Falorsi, S. Pallara, A. Pavone, A. Alessandroni, E. Massella, and M. Scannapieco, "Improving the quality of toponymic data in the italian public administration," in *Proc. ICDT Workshop Data Qual. Cooperat. Inf. Syst.*, Rome, Italy, 2003, pp. 71–74.

[39] G. T. Friedlob, *Understanding Return on Investment*. New York, NY, USA: Wiley, 1996.

[40] M. Ge, M. Helfert, and D. Jannach, "Information quality assessment: Validating measurement dimensions and processes," in *Proc. ECIS*, 2011, p. 75. [Online]. Available: http://aisel.aisnet.org/ecis2011/75

[41] T. Haegemans, M. Snoeck, and W. Lemahieu, "A theoretical framework to improve the quality of manually acquired data," *Inf. Manage.*, vol. 56, no. 1, pp. 1–14, 2019.

[42] A. Haug, F. Zachariassen, and D. Van Liempd, "The costs of poor data quality," *J. Ind. Eng. Manage.*, vol. 4, no. 2, pp. 168–193, 2011.

[43] J. Hegewald, F. Naumann, and M. Weis, "XStruct: Efficient schema extraction from multiple and large XML documents," in *Proc. 22nd ICDE Workshops*, Atlanta, GA, USA, Apr. 2006, p. 81.

[44] B. Heinrich, M. Kaiser, and M. Klier, "Metrics for measuring data quality–foundations for an economic oriented management of data quality," in *Proc. 2nd ICSOFT*, Barcelona, Spain, 2007.

[45] M. Helfert and C. Herrmann. (2002). *Proactive Data Quality Management for Data Warehouse Systems—A Metadata Based Data Quality System*. [Online]. Available: https://www.alexandria.unisg.ch/213765/

[46] (2004). *Guidelines for the Data Quality Improvement of Localization Data in Public Administration*. [Online]. Available: http://www.istat.it

[47] M. A. Jeusfeld, C. Quix, and M. Jarke, "Design and analysis of quality information for data warehouses," in *Proc. Int. Conf. Conceptual Modeling*, Singapore, 1998, pp. 349–362.

[48] A. L. Jones-Farmer, J. D. Ezell, and B. T. Hazen, "Applying control chart methods to enhance data quality," *Technometrics*, vol. 56, no. 1, pp. 29–41, 2014.

[49] S. Juddoo, "Overview of data quality challenges in the context of Big Data," in *Proc. ICCCS*, Pamplemousses, Mauritius, Dec. 2015, pp. 1–9.

[50] B. K. Kahn, "Product and service performance model for information quality: An update," in *Proc. MIT ICIQ*, Cambridge, MA, USA, 1998, pp. 102–115.

[51] B. L. Kahn, D. M. Strong, and R. Y. Wang, "Information quality benchmarks: Product and service performance," *Commun. ACM*, vol. 45, no. 4, pp. 184–192, 2002.

[52] R. S. Kaplan and S. R. Anderson, "Time-driven activity-based costing," Tech. Rep. SSRN 485443, 2003. [Online]. Available: https://ssrn.com/abstract=485443

[53] A. F. Karr, A. P. Sanil, and D. L. Banks, "Data quality: A statistical perspective," *Stat. Methodol.*, vol. 3, no. 2, pp. 137–173, 2006.

[54] (2017). *KPMG: Disrupt and Grow, 2017 Global CEO Outlook*. [Online]. Available: https://assets.kpmg.com/content/dam/kpmg/xx/pdf/2017/06/2017-global-ceo-outlook.pdf

[55] H. Krawczyk and B. Wiszniewski, "Visual GQM approach to quality driven development of electronic documents," in *Proc. 2nd Int. Workshop Web Document Anal.*, Edinburgh, U.K., 2003, pp. 43–46.

[56] N. Laranjeiro, S. N. Soydemir, and J. Bernardino, "A survey on data quality: Classifying poor data," in *Proc. IEEE Pacific Rim Int. Symp. Dependable Comput.*, Nov. 2015, pp. 179–188.

[57] R. Laufer and V. Schwieger, "Modeling data quality using artificial neural networks," in *Proc. Int. Workshop Qual. Geodetic Observ. Monit. Syst.*, Munich, Germany, 2015, pp. 3–8.

[58] Y. W. Lee, D. M. Strong, B. K. Kahn, and R. Y. Wang, "AIMQ: A methodology for information quality assessment," *Inf. Manage.*, vol. 40, no. 2, pp. 133–146, 2002.

[59] I. Lee, "Big data: Dimensions, evolution, impacts, and challenges," *Bus. Horizons*, vol. 60, no. 3, pp. 293–303, 2017.

[60] J. Long and C. Seko, "A cyclic-hierarchical method for database data-quality evaluation and improvement," in *Information Quality* (Advances in Management Information Systems). New York, NY, USA: Routledge, 2005.

[61] D. Loshin, *Enterprise Knowledge Management: The Data Quality Approach*. San Francisco, CA, USA: Morgan Kaufmann, 2001.

[62] D. Loshin, *Master Data Management*. Burlington, MA, USA: Morgan Kaufmann, 2009.

[63] D. Loshin, "Evaluating the business impacts of poor data quality," Knowl. Integrity Incorporated Bus. Intell. Solutions, Silver Spring, MD, USA, Tech. Rep., 2011.

[64] R. Lukyanenko, "Information quality research challenge: information quality in the age of ubiquitous digital intermediation," *J. Data Inf. Qual.*, vol. 7, nos. 1–2, p. 3, 2016.

[65] A. Maydanchik, *Data Quality Assessment*. Bradley Beach, NJ, USA: LLC, 2007.

[66] A. McCallum, "Information extraction: distilling structured data from unstructured text," *Queue*, vol. 3, no. 9, pp. 48–57, 2005.

[67] D. McGilvray, *Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information*. Boston, MA, USA: Morgan Kaufmann, 2008.

[68] J. Merino, I. Caballero, B. Rivas, M. Serrano, and M. Piattini, "A data quality in use model for big data," *Future Gener. Comput. Syst.*, vol. 63, pp. 123–130, Oct. 2016.

[69] N. Micic, D. Neagu, F. Campean, and E. H. Zadeh, "Towards a data quality framework for heterogeneous data," in *Proc. IEEE iThings, IEEE GreenCom, IEEE CPSCom, IEEE Smart Data*, Exeter, U.K., Jun. 2017, pp. 155–162.

[70] E. J. Mishan and E. Quah, *Cost-Benefit Analysis*, 5th ed. London, U.K.: Routledge, 2007.

[71] S. Moore. (2018). *How to Create a Business Case for Data Quality Improvement*. [Online]. Available: https://www.gartner.com/smarterwithgartner/how-to-create-a-business-case-for-data-quality-improvement/

[72] A. Motro and I. Rakov, "Estimating the quality of databases," in *Proc. Int. Conf. Flexible Query Answering Syst.* (Lecture Notes in Computer Science), vol. 1495, T. Andreasen, H. Christiansen, and H. L. Larsen, Eds. Berlin, Germany: Springer, 1998, pp. 298–307.

[73] G. D. Murphy, "Improving the quality of manually acquired data: Applying the theory of planned behaviour to data quality," *Rel. Eng. Syst. Saf.*, vol. 94, no. 12, pp. 1881–1886, 2009.

[74] F. Naumann, J.-C. Freytag, and U. Leser, "Completeness of integrated information sources," *Inf. Syst.*, vol. 29, no. 7, pp. 583–615, 2004.

[75] J. S. Oakland, *Total Quality Management*. Portsmouth, NH, USA: Heinemann, 1989.

[76] P. H. S. Panahy, F. Sidi, L. S. Affendey, and M. A. Jabar, "The impact of data quality dimensions on business process improvement," presented at the WICT, Bandar Hilir, Malaysia, Dec. 2014, pp. 70–73.

[77] E. M. Pierce, "Assessing data quality with control matrices," *Commun. ACM*, vol. 47, no. 2, pp. 82–86, 2004.

[78] E. Pierce and L. Thomas, "Assessing information quality using prediction markets," in *Proc. MIT ICIQ*, Cambridge, MA, USA, 2007.

[79] L. L. Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment," *Commun. ACM*, vol. 45, no. 4, pp. 211–218, 2002.

[80] R. Price, D. Neiger, and G. Shanks, "Developing a measurement instrument for subjective aspects of information quality," *Commun. AIS*, vol. 22, no. 1, pp. 49–74, 2008.

[81] T. C. Redman, *Data Quality For The Information Age*. Boston, MA, USA: Artech House, 1996.

[82] T. C. Redman, *Getting in Front on Data: Who Does What*. Baskin Ridge, NJ, USA: Technics Publication, 2016.

[83] T. L. Saaty, *The Analytic Hierarchy Process*. New York, NY, USA: Wiley, 1980.

[84] S. Sadiq and P. Papotti, "Big data quality—Whose problem is it?" in *Proc. IEEE 32nd Int. Conf. Data Eng.*, Helsinki, Finland, May 2016, pp. 1446–1447.

[85] B. Saha and D. Srivastava, "Data quality: The other face of big data," in *Proc. IEEE 30th Int. Conf. Data Eng.*, Chicago, IL, USA, Mar./Apr. 2014, pp. 1294–1297.

[86] M. Scannapieco, A. Virgillito, C. Marchetti, M. Mecella, and R. Baldoni, "The DaQuinCIS architecture: A platform for exchanging and improving data quality in cooperative information systems," *Inf. Syst.*, vol. 29, no. 7, pp. 551–582, 2004.

[87] L. Sebastian-Coleman, *Measuring Data Quality for Ongoing Improvement*. Waltham, MA, USA: Morgan Kaufmann, 2013.

[88] G. Shankaranarayan, R. Y. Wang, and M. Ziad, "Modeling the manufacture of an information product with IP-MAP," in *Proc. MIT ICIQ*, Boston, MA, USA, 2000.

[89] F. Sidi, P. H. S. Panahy, L. S. Affendey, M. A. Jabar, H. Ibrahim, and A. Mustapha, "Data quality: A survey of data quality dimensions," in *Proc. Int. Conf. Inf. Retr. Knowl. Manage.*, Kuala Lumpur, Malaysia, Mar. 2012, pp. 300–304.

[90] F. Sidi, A. Ramli, M. A. Jabar, L. S. Affendey, A. Mustapha, and H. Ibrahim, "Data quality comparative model for data warehouse," in *Proc. Int. Conf. Inf. Retr. Knowl. Manage.*, Kuala Lumpur, Malaysia, Mar. 2012, pp. 268–272.

[91] Y. Su and Z. Jin, "A methodology for information quality assessment in the designing and manufacturing processes of mechanical products," in *Proc. MIT ICIQ*, Cambridge, MA, USA, 2004, pp. 447–465.

[92] S. R. Sukumar, N. Ramachandran, and R. K. Ferrell, "Quality of Big Data in health care,"Ì *Int. J. Health Care Qual. Assurance*, vol. 28, no. 6, pp. 621–634, 2015.

[93] A. Sundararaman and S. K. Venkatesan, "Data Quality Improvement Through OODA Methodology," in *Proc. MIT ICIQ*, Cambridge, MA, USA, 2017.

[94] I. Taleb, H. T. El Kassabi, M. A. Serhani, R. Dssouli, and C. Bouhaddioui, "Big data quality: A quality dimensions evaluation," in *Proc. IEEE UIC, IEEE ATC, IEEE ScalCom, IEEE CBDCom, IEEE Internet People, Smart World Congr.*, Toulouse, France, Jul. 2016, pp. 759–765.

[95] J. A. Vayghan, S. M. Garfinkle, C. Walenta, D. C. Healy, and Z. Valentin, "The internal information transformation of IBM," *IBM Syst. J.*, vol. 46, no. 4, pp. 669–683, 2007.

[96] R. Vaziri, M. Mohsenzadeh, and J. Habibi, "TBDQ: A pragmatic task-based method to data quality assessment and improvement," *PLoS ONE*, vol. 11, no. 5, pp. 1–30, 2016.

[97] Y. Wand and R. Y. Wang, "Anchoring data quality dimensions in ontological foundations," *Commun. ACM*, vol. 39, no. 11, pp. 86–95, 1996.

[98] R. Y. Wang, "A product perspective on total data quality management," *Commun. ACM*, vol. 41, no. 2, pp. 58–66, 1998.

[99] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *J. Manage. Inf. Syst.*, vol. 12, no. 4, pp. 5–33, 1996.

[100] P. Woodall, A. Borek, and A. K. Parlikad, "Data quality assessment: The hybrid approach," *Inf. Manage.*, vol. 50, no. 7, pp. 369–382, 2013.

[101] Y. Xiao, L. Y. Y. Lu, J. S. Liu, and Z. Zhou, "Knowledge diffusion path analysis of data quality literature: A main path analysis," *J. Informetrics*, vol. 8, no. 3, pp. 594–605, 2014.

[102] H. Zhu, S. E. Madnick, Y. W. Lee, and R. Y. Wang, "Data and information quality research: Its evolution and future," *Comput. Handbook*, no. 2, pp. 1–20, 2014.

CORINNA CICHY received the B.Sc. degree in mathematics with actuarial science from the University of Southampton, U.K., in 2016, and the M.Sc. degree in statistics from the University of Warwick, Coventry, U.K., in 2017. She is currently pursuing the Ph.D. degree in applied computer science with Alpen-Adria-Universität Klagenfurt, Austria. She is a part of the Volkswagen Bank Ph.D. program and works in their risk management department. Her work and research interests include data quality assessment based on machine learning techniques, data modeling, and regulatory compliance.

STEFAN RASS graduated with a double master's degree in mathematics and computer science from Alpen-Adria-Universität Klagenfurt (AAU), in 2005, the Ph.D. degree in mathematics, in 2009, and the Habilitation degree in applied computer science and system security, in 2014. He is currently an Associate Professor with AAU, teaching courses on theoretical computer science, complexity theory, and security and cryptography. He authored numerous papers related to security and applied statistics and decision theory in security. Closely related to the project is his (co-authored) book *Cryptography for Security and Privacy in Cloud Computing*, (Artech House). He participated in various nationally and internationally funded research projects. His research interests include applied system security, and complexity theory, statistics, decision theory, and game-theory.