Use of Context in Data Quality Management: a Systematic Literature Review

FLAVIA SERRA, Instituto de Computación, Universidad de la República Facultad de Ingeniería, Montevideo, Uruguay
VERÓNICA PERALTA, University of Tours, Tours, France
ADRIANA MAROTTA, Instituto de Computación, Universidad de la República Facultad de Ingeniería, Montevideo, Uruguay
PATRICK MARCEL, Université d’Orléans, ORLEANS, France

The importance of context in data quality (DQ) was shown many years ago and nowadays is widely accepted. Early approaches and surveys defined DQ as fitness for use and showed the influence of context on DQ. This paper presents a Systematic Literature Review (SLR) for investigating how context is taken into account in recent proposals for DQ management (DQM). We specifically present the planning and execution of the SLR, the analysis criteria and our results reflecting the relationship between context and DQ in the state of the art and, particularly, how this context is defined and used for DQM. The SLR is instrumental to the identification of context components and the design of a context formal model.

CCS Concepts: Information Systems—Data Management Systems—Data Quality—Context in Data Quality Management;

Additional Key Words and Phrases: Systematic Literature Review, Data Quality, Context, Data Quality Methodology

1 Introduction

Data quality (DQ) is a very wide research area, which involves many different aspects, problems and challenges. The growing need to discover and integrate reliable information from heterogeneous data sources, distributed in the Web, social networks, cloud platforms or data lakes, makes DQ an unavoidable topic, particularly hot in recent years (see e.g., [64, 68, 87, 90]). In addition, DQ has enormous relevance for the industry, due to its great impact on information systems in all application domains.

Advances in information and communication technologies have led organizations to manage increasingly large amounts of data. This generates interesting opportunities, both for the daily operations as well as for decision-making and strategic planning. However, these opportunities may be limited by DQ issues in this kind of scenarios [62], because DQ problems can lead to making erroneous decisions and domain strategies planning that do not comply with DQ requirements. In addition, such applications involve many services, processes and tasks, using varied data and metadata, for many purposes, and involving different users and roles. These variations mean different contexts for data, and DQ perceptions and expectations may strongly vary from one context to another.

Authors’ Contact Information: FLAVIA SERRA, Instituto de Computación, Universidad de la República Facultad de Ingeniería, Montevideo, Uruguay; e-mail: fserra@ing.edu.uy; Verónica Peralta, University of Tours, Tours, Centre-Val de Loire, France; e-mail: veronika.peralta@univ-tours.fr; Adriana Marotta, Instituto de Computación, Universidad de la República Facultad de Ingeniería, Montevideo, Uruguay; e-mail: amarotta@ing.edu.uy; Patrick Marcel, Université d’Orléans, ORLEANS, France; e-mail: patrick.marcel@univ-orleans.fr.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM 1936-1963/2024/6-ART
https://doi.org/10.1145/3672082

ACM J. Data Inform. Quality
In this work we focus on the context of the data, and we investigate how this context influences the quality of such data. Early approaches and surveys defined DQ as *fitness for use* because, according to Wang & Strong [103], one of the DQ area foundational papers, DQ must be considered in the context of the task carried out by the data users. According to Watts et al. [104], from this usage perspective, DQ assessment tends to be contextual. Therefore, a single set of DQ metrics is not enough, hence task-dependent DQ metrics, developed for specific application contexts, should be included [88]. Enriching DQ metrics definitions with context management would provide with: (i) flexibility, since they could adapt to context variations, (ii) generality, since they could include many particular context-dependent cases, and (iii) richness, leading to include more aspects to the metric. Even & Shankaranarayanan [74], rely on the idea that a dataset may be used in multiple contexts and contribute to utility (a measure of the benefits gained by using data) differently in each. But how is the context defined?

We observed that in the literature there is no single definition, rather it seems that context can be made up of several components. For instance, Pipino [88] considers that several elements form the context: organization’s business rules, company and government regulations, and constraints provided by the database administrator. Other authors include the users [72], or characteristics of the decision task and the decision-maker [104]. Despite its recognized importance for DQ management (DQM), context is typically taken as an abstract concept, which is not clearly defined nor formalized. In order to evidence this lack, we analyzed the state of the art on context definition for DQM, with special attention to modeling and formalization aspects.

This paper presents a Systematic Literature Review (SLR) for investigating how, when and where context is taken into account in proposals for DQM, published between 2010 and 2020. Our motivation is that, to the best of our knowledge, there are no works that analyze data context definitions in DQ, nor the contextual aspects of DQ activities addressed in the different steps of a DQM process. At this stage of the research, we only focus on surveying context proposals for DQM, it is not within the scope of this work to give a context definition.

A SLR is a methodology for identifying, evaluating and interpreting all available research literature that is related to a particular research question or topic area. It is a widely used methodology [80] in multiple scientific domains of information technology, computing and education. In particular, we highlight SLR in the DQ area [77, 89, 106]. SLR is not a conventional literature review, because while the latter just provides a high level summary of the literature in connected fields, the SLR provides a more complete, rigorous and reproducible framework for the literature study [81].

**Contributions.** Our main contribution is the analysis of the bibliography, which allows us: i) to identify works that address the context of data for DQM, ii) to identify which are the proposals that formalize a definition of context, iii) to identify and analyze which are the context components, and iv) to study the influence of the context on the DQ activities developed at each step of a DQM methodology.

**Outline.** The paper is organized as follows: Section 2 provides background on the important notions used throughout this paper. Section 3 presents the planning and execution of the SLR and describes its results. Section 4 presents an analysis of the contexts identified in the literature by defining a taxonomy. Section 5 answers the research questions of the SLR and presents the research challenges that arise from such answers. Finally, in Section 6, we present our conclusions, proposals and future works. In addition, in Appendix A we list all the works (named primary studies) selected by applying the SLR. The bibliography referenced in this work is classified into 2 groups: the works selected by the SLR and the general bibliography.

2 Background
This section presents background concepts about DQ, context and context for DQ.
2.1 Data Quality

DQ is defined as fitness for use and is widely recognized to be multidimensional [103]. In the literature, we find the terms data quality and information quality (IQ) interchangeably, as for example in [88]. However, Batini and Scannapieco [62] use DQ when referring to dimensions, models, and techniques strictly related to structured data, while, in all other cases, they use IQ. Therefore, there is no consensus on whether there is or not a difference between DQ and IQ, since the difference is established or not in each proposal.

DQ dimensions express the characteristics that data must have, such as its correctness, completeness and consistency. In the literature, there is no agreement on the set of the dimensions characterizing DQ nor on their meanings [91].

DQ methodologies organize DQ activities by executing a series of ordered steps, in particular, to measure, evaluate and improve DQ. A DQ methodology is defined by Batini et al. [61] as a set of guidelines and techniques that, starting from input information describing a given application context, defines a process to assess and improve DQ.

In the next subsections, we list the most representative DQ dimensions, which are considered in our review. In addition, we present the DQ activities addressed by DQ methodologies. In particular, we focus on a DQ methodology, called Complete Data Quality Methodology (CDQM) [61], whose initial phase proposes analyzing the context of the data. Considering this approach, we aim at deepening the analysis of the contextual aspects of the 3 proposed phases. For this, below we describe each of the CDQM phases.

2.1.1 DQ Dimensions and DQ Metrics. In a large industrial survey, Wang & Strong inventoried 175 DQ attributes used by data consumers (attributes that came to mind when the data consumer thought about DQ) [103] and organized them in 15 DQ dimensions grouped in 4 DQ categories [100]: intrinsic DQ (namely, accuracy, objectivity, believability, reputation); accessibility DQ (namely, accessibility, access security); contextual DQ (namely, relevancy, value-added, timeliness, completeness, amount of data); and representational DQ (namely, interpretability, ease of understanding, concise representation, consistent representation).

The ISO/IEC 25012 standard [98] also proposes 15 DQ dimensions (called characteristics), and presents them in 3 groups: i) inherent DQ (namely, accuracy, completeness, consistency, credibility, currentness), ii) between inherent and system-dependent DQ (namely, accessibility, compliance, confidentiality, efficiency, precision, traceability, understandability), and iii) system-dependent DQ (namely, availability, portability, recoverability). Note that the focus of our research is not the impact of the set of standards on the data context, but rather which DQ dimensions proposed by the standard are context-dependent.

All DQ dimensions mentioned above are those traditionally used in the DQ domain. Additionally, they are reused by Merino et al. in the domain of big data [41]. This is aligned with the work of Shankaranarayanan & Blake in [50], where they mention that DQ research has not tackled big data, having to consider traditional DQ dimensions. They claim that what makes big data difficult to analyze using traditional tools are the "three V's": volume, velocity and variety. However, the authors also mention that the industry, already in 2017, was discussing a fourth "V", veracity, to describe the accuracy and believability of the data. Nowadays, the term data veracity is widely used in big data applications [2–4, 18, 24], and sometimes ambiguously relating to several traditional DQ dimensions.

While DQ dimensions are used to express the characteristics that data must satisfy, DQ metrics provide quantitative means for assessing to what extent these characteristics are satisfied. For example, according to Pipino et al. in [88], the analysis of the satisfaction of the Codd’s Referential Integrity constraint, across tables in a relational database, can be addressed through the DQ dimension consistency. Meanwhile, a DQ metric measuring this DQ dimension is the ratio of violations of a specific consistency type to the total number of consistency checks subtracted from one.
On the other hand, Pipino et al. in [88] argue that it is difficult to define which aspects of a DQ dimension are related to a specific application in a company. However, once this task is done, proposing DQ metrics associated with these dimensions is easier. Their research addresses DQ assessment based on task-dependent and -independent DQ metrics. According to the authors, companies must consider the subjective perceptions of the users, and the objective measurements based on the used data.

The term DQ attribute is also often used in the DQ literature, but ambiguously, as it may refer to DQ dimensions [70], to specific aspects of DQ dimensions [103], or even to DQ indicators [102]. For instance, Jiang et al. [78] use the terms DQ dimension and DQ attribute interchangeably, while Wang & Strong [103] define DQ dimension as a set of DQ attributes that represent a single aspect of DQ. Another approach is presented by Batini & Scannapieco [62], where an attribute in a relational database is extended with DQ attributes (namely accuracy, currency, and completeness). In accordance with the authors, the values of such DQ attributes represent values of DQ dimensions.

2.1.2 DQ methodologies. In this section we do not want to delve into DQ methodologies, but rather into DQ activities addressed by them. For this reason, in this survey we focus on a particular DQ methodology that covers the overall activities proposed in the literature. A comparison of methodologies can be found in [61, 70, 96].

Several DQ methodologies have been proposed in the literature, in particular, in [61] thirteen of them are compared. Within the framework of this comparison, the authors mention that, in general, the sequence of activities of a DQ methodology is composed of 3 phases: i) state reconstruction, ii) measurement/assessment, and iii) improvement. For comparison, they focus mainly on various tasks or steps that are included at the last 2 phases.

For our analysis, we select one of the methodologies analyzed in [61]: the Complete Data Quality methodology (CDQM), because it is the only one that addresses the 3 phases and most of the analyzed steps. Indeed, the other methodologies neglect state reconstruction phase, where requirements are elicited. Fig. 1 sketches its phases and steps, which are described below.

ACM J. Data Inform. Quality
Phase 1 (state reconstruction). According to Batini et al. [61], the objective of this phase is to collect contextual information on organizational processes and services. This phase involves identifying data types, databases and related management procedures, DQ problems and corresponding costs, business rules and norms. That is, eliciting everything that is important to an organization, in terms of users, processes, tasks, data, and the relationships among them. This phase consists of a unique step, ST1, also called state reconstruction.

Phase 2 (measurement/assessment). This phase focuses on the DQ model definition, measurement and assessment. It comprises 3 steps: data analysis (ST2), DQ requirements analysis (ST3), and DQ measurement/assessment (ST4). At ST2 the data are analyzed to get a first impression of its quality status. Data profiling techniques (which are considered as the process of analyzing a dataset in a quality domain defined by a set of DQ and data requirements [52]) can be used and interviews with data users may be conducted. At ST3 DQ requirements are analyzed and prioritized, and at ST4 the DQ model is defined (i.e. DQ dimensions relevant to the DQ requirements are selected and DQ metrics, that calculate the DQ values, are defined), which allows measuring and evaluating DQ.

Batini et al. clarify in [61] that the term “measurement” is used to address the task of measuring the DQ values associated with the selected DQ dimensions, in order to enable a DQ evaluation. In turn, the term “assessment” is used when the DQ values obtained in the measurement are compared with reference values (e.g. thresholds or DQ requirements), in order to enable a diagnosis of DQ. In this work, we adopt the same meaning for each of the aforementioned terms.

Phase 3 (improvement). This phase is focused on improving DQ, and it also comprises 3 steps: ST5, ST6 and ST7, called error causes identification, strategies and techniques selection, and costs evaluation, respectively. At ST5 the causes of DQ problems are identified. Based on them some strategies and techniques are selected to achieve new DQ targets at ST6. In the end, ST7 costs evaluation is addressed. According to Batini et al. [61], there are two types of costs: i) costs associated with poor DQ, also called indirect costs (i.e. process costs caused by data errors, and opportunity costs due to lost and missed revenues); ii) costs of assessment and improvement activities, also called direct costs.

We believe that the first phase is very important, since, in order to execute the other 2 phases efficiently, it is necessary to know all the organizational resources. Assuming that the results obtained in the first phase are known can lead to a DQ assessment that does not correspond to the reality of the organization. This, in the end, will have an impact on organizational costs, whether direct or indirect.

As we saw, Batini et al. [61] only analyze contextual aspects at the first phase of CDQM. However, relevant literature from the DQ area indicates that the context is relevant in DQ activities (such as DQ dimensions selection [100] and DQ measurement/assessment [88]), present at steps corresponding to other CDQM phases. Taking this into account, we hypothesize that the context is relevant in all phases of CDQM. Thus, in Subsections 4.3 and 4.4, we analyze the elements that make up the context and how they are linked to each CDQM phase.

2.2 Context

Seeking to understand what the context is, Bazire and Brézillon [63] analyzed a corpus of 150 definitions in different domains of cognitive sciences and close disciplines, and build a model of context representing the components of a situation and the different relations between those components. As pointed out by Dey [72], context is a poorly used source of information in computing environments, resulting in an impoverished understanding of what context is and how it can be used. Dey provided operational definitions of context and context-aware computing: context is a general term used to capture any information that can be used to characterize the situations of an entity, a system being context-aware if it uses context to provide relevant information and/or services to the user, where relevancy depends on the user’s task.
According to Bolchini et al. [65] the notion of context arises in the areas of Psychology and Philosophy, and nowadays it acquires more and more importance in the Computer Science area. This demonstrates the importance that the context has taken and the relevance it has gained in Computer Science. In addition, they point out that in a commonsense interpretation, the context is perceived as a set of variables that may be of interest for an agent and that influence its actions. This definition does not deviate from the one given by Dey [72]. In [65] different context models to support context-aware applications are identified, and are classified into 4 categories, according to the focus of interest: presentation, location, user, or community. In accordance with the authors, these approaches added a new perspective called data tailoring-oriented, since they aim at the reduction of the size of data by means of contextual preferences. Context-dependent preferences are also addressed in [69, 99]. Both proposals address the problem of selecting the appropriate context-based preferences for customizing a query. In particular, in [69] it is pointed out that user preferences are very important for personalized database applications, specifically in those in which the user context plays a key role. Taking into account the aforementioned investigations, we consider that preferences and data tailoring appear as context-related concepts. Therefore, including them as relevant terms in our research could lead us to discover interesting approaches to the context.

On the other hand, while ubiquitous computing is related to the context, it is a more restrictive concept. The term ubiquitous refers to the seamless integration of devices into the users everyday life [105]. For example, mobile devices such as notebooks, personal digital assistants, and smart phones are considered ubiquitous systems, and they are the so-called context-aware systems [60]. The authors of [59] point out that, in this type of system, the context is more than position and identity. They consider that although a complete definition of context is illusive, it is necessary to define a minimum set of context components, answering to who, what, where, when and why questions. But this is not the only important thing, since related to the context definition is the question of how to represent it. All these visions were analyzed in a preliminary investigation [92].

So far we have presented different approaches to context, however, this research is focused exclusively on the relationship between data context and DQ. Literature that addresses this relationship and establishes the importance of considering context in DQ is presented in the next section.

### 2.3 Context for Data Quality

As we mentioned before, early approaches defined DQ as fitness for use [103], and this concept has been widely adopted in the literature [50, 66, 79, 85]. According to Watts et al. [104], from this fitness for use perspective, DQ assessment tends to be contextual, and reference works agree with this approach [100]. Even nowadays it continues to be addressed [2, 10, 11, 20, 21, 42]. However, to the best of our knowledge, there is no work that fully exploits the DQ contextual aspects.

DQ assessment is a challenging task that requires data context and demands a great domain knowledge [84]. The importance and influence of data context in DQ has been stated many years ago [103], and is widely accepted. Early works only state whether a DQ dimension depends on context or not [100]. Already in 2006, Shankaranarayanan & Cai [97] pointed out that contextual factors have not been explicitly examined in the DQ literature, and recognized the contextual or subjective nature of DQ evaluation.

In 2009, Watts et al. [104] argue that contextual assessments can be as important as objective quality indicators, because they can affect the information used for decision making tasks. Even & Shankaranarayanan [74] say that contextual assessment measures the extent to which the presence of defects reduces a dataset’s utility. Loshin [83], in 2011, states that measures are contextual because they depend on the context. Furthermore, the author adds that contextual DQ dimensions depend on business policies that are implemented as business rules within systems and processes. However, neither the business rules nor the context are specified. Most recently, in 2019, Catania et al. [21] suggest that contextual information can help in interpreting users’ needs in linked data. While
in 2020, Davoudian and Liu [24] claim that traditional DQ tasks do not take into account big data characteristics, since the context of use is appropriate for big data quality analysis.

The need to consider the context for DQ assessment persists over time and for different application domains. Nevertheless, even today the literature lacks a concise and globally accepted definition for context in DQ [49]. Therefore, this work aims at investigating whether context is used and how it is used in recent proposals for DQM.

3 Application of a SLR to investigate the context usages in DQ

A SLR [81] is a methodology aiming at identifying, evaluating and interpreting all available research relevant to a particular research question or topic area. It starts by defining a review protocol that specifies the research questions being addressed, and the methods that will be used to perform the review. These research questions determine the criteria to assess potential relevant data to answer the questions. Relevant data are extracted from the scientific works found through the SLR application, called primary studies (from now on PS). According to Kitchenham [81], the SLR methodology and produced documents (protocol, questions, etc.) ensure the completeness, rigor and reproducibility of the review. A SLR is performed at 3 stages [81]: (i) planning (review objectives and research questions are defined), (ii) conducting (digital libraries are selected, search strings, and inclusion and exclusion criteria are defined for obtaining a set of PS), and (iii) reporting (SLR results are reported).

Next, we describe the planning and conducting of the SLR methodology applied to our research problem, and we present quantitative and qualitative results. More detailed findings, an overview of the selected PS, and descriptions of some proposals, are presented in Subsection 3.3.

3.1 Planning the Review

At this stage, the review objectives, the research questions that arise from such objectives, and the search strings that result from the research questions are defined. Note that both, research questions and search strings, are shown as they were defined and executed, that is, abbreviations were not used in any of them (i.e. neither DQ nor IQ).

Review Objectives: We apply a SLR to know current evidence in the relationship between context and DQ areas. In short, we want to identify any gap at the intersection of these areas in order to detect new research lines.

Research Questions: To define the research questions, the most important DQ concepts, such as DQ, IQ, DQ dimension, DQ metric, DQ attributes, and DQ model are taken into account. In particular, it is important to know if DQ models have been defined considering context and how context is represented in such models. In turn, we search whether context participates in the definition of DQ dimensions and their DQ metrics, and more generally, how context relates to DQ concepts at different stages of DQM. For all this, the following research questions are raised:

- RQ1: How is context used in data quality models?
- RQ2: How is context used within quality metrics for the main DQ dimensions?
- RQ3: How is context used within data quality concepts?

Note that RQ3 is intentionally more inclusive, in order to get works dealing with DQ and context but not mentioning specific DQ models, DQ dimensions or DQ metrics.

3.2 Conducting the Review

The SLR execution is carried out following a series of steps: search strings definition, selection of the digital libraries, definition of the inclusion and exclusion criteria, and selection of the primary studies. All of them are described below.
Search Strings: First, we extract the keywords from the research questions, namely: context, data quality model, quality metric, data quality dimension, quality concept and data quality. The latter two are the decomposition of a more complex keyword, data quality concept, appearing in RQ3. Then, to perform a thorough search, some keywords are complemented or detailed with alternative terms as follows (each of them is discussed and argued in Section 2):

- **context** is complemented with alternative terms data tailoring and preference.
- **data quality dimension** is refined by the DQ dimensions. Note that, contrarily to dimensions, metrics are not refined. Indeed, unlike dimensions, metrics are usually not managed with names in the DQ literature, since they are defined specifically for each measurement process.
- **quality concept** is refined by important DQ concepts: quality dimension, quality metric and quality attribute.
- **data quality** is complemented with alternative term information quality.

We performed several tests with partial search strings to create, for each research question, a final search string, which is built as a conjunction of keywords (AND connector). The alternative terms are joined with the keywords by means of disjunctions (OR connectors). We can see the alternative terms of each keyword in Table 1. RQ2 is very comprehensive, then the associated search string is too long to be supported by search engines of some digital libraries, as it should list 27 DQ dimensions. To solve this problem, we cut the search string in 7 smaller ones, each containing a subset of the DQ dimensions. The resulting search strings are:

- **SS1**: (context OR preference OR "data tailoring") AND ("data quality model")
- **SS2**: (context OR preference OR "data tailoring") AND ("quality metric") AND ("data believability" OR "data accuracy" OR "data objectivity" OR "data reputation")
- **SS3**: (Context OR preference OR "data tailoring") AND ("quality metric") AND ("data value-added" OR "data relevancy" OR "data timeliness" OR "data completeness" OR "appropriate amount of data")
- **SS4**: (Context OR preference OR "data tailoring") AND ("quality metric") AND ("data interpretability" OR "data ease" OR "data understanding" OR "data representational consistency" OR "data concise representation")
- **SS5**: (Context OR preference OR "data tailoring") AND ("quality metric") AND ("data credibility" OR "data currentness" OR "data veracity")
- **SS6**: (Context OR preference OR "data tailoring") AND ("quality metric") AND ("data accessibility" OR "data compliance" OR "data confidentiality")

| Keywords                       | Alternative terms                                                                 |
|-------------------------------|-----------------------------------------------------------------------------------|
| context                       | preference, "data tailoring"                                                      |
| data quality model            |                                                                                   |
| quality metric                |                                                                                   |
| data quality dimensions       | "data believability", "data accuracy", "data objectivity", "data reputation", "data value-added", "data relevancy", "data timeliness", "data completeness", "appropriate amount of data", "data interpretability", "data ease", "data understanding", "data representational consistency", "data concise representation", "data credibility", "data currentness", "data veracity", "data accessibility", "data compliance", "data confidence", "data efficiency", "data precision", "data traceability", "data understandability", "data availability", "data portability", "data recoverability" |
| quality concept               | "quality dimension", "quality metric", "quality attribute"                        |
| data quality                  | "information quality"                                                             |

Table 1. Alternative terms for each keyword.
- **SS7**: (Context OR preference OR “data tailoring”) AND (“quality metric”) AND (“data precision” OR “data traceability” OR “data understandability”)
- **SS8**: (Context OR preference OR “data tailoring”) AND (“quality metric”) AND (“data availability” OR “data portability” OR “data recoverability”)
- **SS9**: (Context OR preference OR “data tailoring”) AND (“data quality” OR “information quality” ) AND (“quality metric” OR “quality dimension” OR “quality attributes”)

**Digital Libraries**: We select digital libraries trying to cover the most important venues in the database domain and considering the most representative search engines of the DQ area. We look for the least possible overlap among the libraries, thus reducing the number of duplicate PS. In addition, the selected digital libraries should allow for search in both titles and abstracts. We initially avoided to use the search engines of DBLP and Google Scholar because most of the returned PS are already returned by the other digital libraries. As well, Google Scholar indexes many other types of documents (e.g. technical reports or slides) producing noise within search results and DBLP does not support search inside abstracts, loosing many relevant PS. However, we are aware that some important venues have online proceedings, in particular VLDB, EDBT and ICDT. So, we complete the SLR with a specific search in Google Scholar, but restricted to 3 venues: VLDB, EDBT and ICDT.

| Digital Library                          | Stage 1 | Stage 2 | Stage 3 | Stage 4 | Stage 5 |
|-----------------------------------------|---------|---------|---------|---------|---------|
| ACM                                     | 506     | 120     | 15      |         |         |
| IEEE                                    | 11      | 3       | 0       |         |         |
| ScienceDirect                           | 1027    | 48      | 9       |         |         |
| Springer                                | 1287    | 107     | 29      |         |         |
| Google Scholar (VLDB, EDBT, ICDT)      | 67      | 1       | 1       |         |         |
| From references                         |         |         |         |         | 4       |
| **Total**                               | 2898    | 1955    | 279     | 54      | 58      |

Table 2. Amount of selected PS by stage and Digital Library

The selected digital libraries are the following:
- **ACM digital library** (http://dl.acm.org)
- **IEEE Xplore** (http://ieeexplore.org)
- **Science Direct** (http://www.sciencedirect.com)
- **Springer LNCS** (http://www.springer.com)
- **Google Scholar** (https://scholar.google.com) restricted to VLDB, EDBT and ICDT.

**Inclusion and Exclusion Criteria**: To restrict the PS returned by each digital library, we apply a set of inclusion and exclusion criteria. The former are automatically applied by digital libraries, while the latter are considered later, while reading abstracts or full texts.

**Inclusion criteria**: (i) PS published in the period 2010-2020 inclusive (this time interval is large enough to ensure getting recent PS), (ii) written in English, (iii) published in PDF format, and (iv) being articles or book chapters.

**Exclusion criteria**: (i) full text is written in another language (even if the abstract is in English), (ii) works published in non peer-reviewed venues, (iii) DQ is not addressed or it is addressed superficially (i.e., the need for high-quality data is mentioned, but the work does not address any of the relevant activities of the DQM process, namely, data profiling, DQ model definition, DQ measurement, DQ assessment and/or DQ improvement), or (iv) context is not addressed. These criteria allow to discard PS addressing other subjects even if containing the good keywords.
Selection of Primary Studies: The selection process has 5 stages. They are listed below, highlighting the number of PS output by each stage, which are summarized in Table 2.

- **Stage 1: Execution of search strings.** 2898 PS were returned by the 5 digital libraries in response to search strings and inclusion criteria.
- **Stage 2: Duplicate elimination.** Some PS were returned by several search strings within the same digital library, and many PS from Google Scholar were also returned by other libraries. Duplicate elimination resulted in 1955 PS.
- **Stage 3: Selection by relevance.** 279 PS were selected by relevance, reading the title and abstract of each PS. A title and/or abstract is relevant when it suggests addressing at least one of the most important activities of the DQM process: data profiling, DQ model definition, DQ measurement, DQ assessment and/or DQ improvement.
- **Stage 4: Selection by full text.** A final selection, by a careful reading of full text and application of exclusion criteria, resulted in 54 PS.
- **Stage 5: Selection by references.** Finally, the references of the 54 selected articles are reviewed. At this stage we apply all the inclusion and exclusion criteria, except the year of publication, since in the final set of PS we also want to include relevant works prior to the year 2010. In this way, 4 additional articles were selected, obtaining a total of 58 PS.

According to Kitchenham [81] this methodology is reproducible. However, there are some exceptions as the methodology depends on external factors. In particular, we experienced some problems with the ACM library, which changed its search engine in the middle of the SLR process. Indeed, our first search returned very few PS, while a second search some month later, returned hundreds of PS. This evidences, that the reproducibility of the methodology, as described in [81], is impacted by the changes in digital libraries. The complete selection process is shown in Fig. 2.

### 3.3 Reporting the Review

In this subsection we present and quantify the main results of the review by analyzing the selected PS according to different criteria.

Firstly, for complementing Table 2, Fig. 3 shows the distribution of selected PS by (a) digital library, (b) search string, (c) year of publication and (d) venue quality. We remark that more than a half of the selected PS were retrieved with SS9, while none with SS4, and most of selected PS come from Springer and ACM, while none from IEEE. Interestingly, the number of published papers dealing with the use of context for DQ increased from 2016, except for 2017. Selected PS are classified according to the quality of venues, in accordance with rankings and metrics of Scopus\(^1\) for journals and Core\(^2\) for conferences. We use the quality labels (A*, A, B, C) proposed by Core, which correspond to Scopus relative rankings of [94%-100%), [84%-94%), [54%-84%) and [13%-43%), respectively. NR is used for referring to not ranked PS and CH for referring to book chapters. We remark that of 58 selected PS, only 13 PS are ranked A* or A (and interestingly, 11 of them were published in or after 2016). In addition, 23 PS are ranked B (including 11 from the International Journal on Data and Information Quality), and 14 PS are not ranked. Concerning venue type, 31 PS come from conference proceedings, 21 from journals and 6 from book chapters.

Seeking to give readers an overview of the selected PS, firstly, we observe general characteristics of each PS:

- **Research domain.** We consider the research domain addressed in the PS. We are not interested in going deeper into the research domain, but rather we want to identify those addressed by the authors. First, we

---

\(^1\)SCOPUS portal. https://www.scopus.com/sources (accessed December 2020)

\(^2\)CORE portal. http://portal.core.edu.au/conf-ranks/ (accessed December 2020)
highlight PS that only address DQ (or IQ) without restricting to a specific domain (we call it only DQ). On the other hand, other PS concern contextual DQ for a main research domain, namely: big data, decision making, linked data, internet of things, data mining, e-government, open data, machine learning or data integration. We consider that the research domain is the area of research addressed in the PS, and we name them keeping the exact terms used by the authors.

- **Type of work.** We highlight and classify the PS according to the type of proposal. We find several approaches: framework, model, methodology, analysis, case study, metrics and architecture.

- **Data model.** We observe the data model used in the selected PS and classify them according to this data model. In this way, we found PS with structured, semi-structured, unstructured and mixed data models. In some PS data are not specified or PS are general, i.e., not restricted to a data type or model. In both cases, PS are classified as N/A.

- **Case study.** We observe if the PS is or presents a case study. In particular, in works that present a case study, we are interested in identifying what type of data is used. Then, PS are classified according to the type of data used in the case study presented to evaluate the proposal, namely: real data, artificial data and not specified (N/A). The latter corresponds to PS without a case study.

Secondly, we analyze the most relevant aspects of each PS for our research. Specifically, we focus on the type of DQ activity that the PS addresses and whether it uses or proposes DQ dimensions and/or metrics.

Fig. 2. Primary studies selection process.
CDQM steps. We consider the seven steps of CDQM [61] described in Section 2. In this analysis, a CDQM step is relevant to a PS if the PS addresses DQ activities related to the step or at least present a discussion that involves the analyzed step. For example, if a PS proposes DQ metrics, we consider that such PS is aligned with step ST4 (measurement/assessment) of CDQM. When a PS does not address any of the suggested steps, it is classified as N/A.

DQ dimensions. Some authors affirm that there are contextual and non-contextual DQ dimensions (see for example the classification of Wang & Strong [100] introduced in Section 2). However, there is no agreement in the literature. Therefore, in order to investigate what DQ dimensions are more frequently considered as contextual we classify PS according to whether they emphasize contextual DQ dimensions or not. For the former, we also indicate the elements providing context, namely bibliography, the ISO/IEC 25012 standard [98], user, data/information, application domain, DQ requirements, metadata, task at hand, and utility (usefulness of the information for its anticipated purpose [62] or a measure of the benefits gained by using data [74]).

DQ metrics. We are also interested in identifying PS that propose contextual DQ metrics. Therefore, for us, an PS proposes contextual DQ metrics if they are defined, used or analyzed taking into account some element of the data context.

Below, we present the analysis results. We group the most general results, while the more specific results are presented individually. Fig. 4, 5, 6, 7 and 8 present quantitative results.

Case study, Data model and Research Domain. We analyze the PS that present a case study, the type of data model that they use, and the research domain that they address.
According to the Fig. 4(a), most PS use real data to develop their case studies but artificial data are also largely used. We remark that many PS do not provide a case study (N/A), being especially the case of reviews and surveys.

Regarding the data models used in the selected PS, in Fig. 4(b) we show their distribution. As we can see, most of research is based on structured data, especially using the relational schema. There are also a high number of PS that consider unstructured data, as CSV files or sensor data. In addition, there is an important number of PS that are general, i.e. they are not restricted to a data type or model (they are classified as N/A).

On the other hand, Fig. 5(a) shows the distribution of the main research domains addressed by the selected PS. We do not delve into the research domains, we are only interested in showing which are the research domains in which DQ issues are addressed. Obviously, most of them only concern DQ (or IQ) but their proposals are not domain-specific. As we mentioned before, we refer to them as only DQ. However, we can see that many PS combine DQ with other research domains. For example, with the big data domain, where the management of large volumes of data presents a clear need to incorporate DQ tasks [18, 46, 50] and to study the context into those DQ tasks [2, 3, 17, 24, 41, 52].

**Type of Work.** We present in detail the type of works proposed by the PS. When PS have several contributions, we focus on the main proposals in order to indicate the leading types of work.

The most frequent type of work, as evidenced in Fig. 5(b), is model, since several PS coincide in proposing a DQ model [12, 13, 17, 29, 33, 38, 41, 43, 52, 53]. In particular, [13] presents domain-specific characteristics of DQ.
while [12] extends this work by proposing DQ models. In addition, in [6], the authors discuss models proposed for utility, discovering that DQ dimensions and DQ metrics are deeply influenced by utility. This proposal considers the usage of data and the relevance of processes that adopt it in the measurement. In turn, [9] presents a model to identify opportunities for increasing monetary and non-monetary benefits by improved DQ. In other matters, we identify other kind of models, all for DQ assessment and motivated by the idea that DQ is context-dependent. For instance, the authors of [55] present a decision model to facilitate the description of business rules, and in [10, 11, 42] a model of context is developed. As well, there are PS that address the impact of poor DQ and propose improvement models, specifically [25] presents a machine learning model and [39] a neural network model. Similarly, [16] presents a model of the asset management data infrastructure and [5] a Bayesian network model that shows how DQ can be improved while satisfying budget constraints. In the case of [40], it presents a formal multidimensional model on which is applied a rule-based approach to DQ assessment, and [27] provides a cost and value model for issues related to DQ. Finally, [21] models context for source selection in linked data, and [1] proposes a conceptual model where the relationships among quality characteristics of e-government websites and users’ perceptions are represented.

Another PS characteristic that we observe concern frameworks for assessing DQ (or IQ) [3, 7, 14, 21, 27, 28, 37, 38, 40, 44, 45, 49, 53, 57, 58]. For instance, in [21] the authors use a framework called Luzzu [71] for linked DQ assessment. Additionally, [40] provides a framework for context-aware data warehouse quality assessment. Other proposals are task-specific, as [5] that presents a framework for discovering dependencies between DQ dimensions, or [19] that proposes a framework that encapsulates data cleaning procedures. Also in [57] we find an ontology-based quality framework for data integration.

On the other hand, in some frameworks DQ is not the goal, but it is a relevant element of the proposal. For example, in [1] a framework of citizens’ adoption of e-government services is presented. Furthermore, an alternative angle is presented in [27], since the authors develop a framework in the form of a guideline to manage DQ costs. They define DQ cost as financial impacts caused by DQ problems and resources required to achieve a specified level of DQ. Finally, the framework in [3] describes the implementation of a system that acquires tweets in real time, and applying DQ metrics to measure the quality of such tweets. These DQ metrics are defined formally.

Among the works classified as analysis, we find reviews and surveys. We consider them together because the goal of both is the same, to analyze the state of the art. For instance, [50] identifies the core topics and themes that define the identity of DQ research, whereas [26] analyzes the main concepts and activities of DQ. In the case of [36], the authors investigate DQ in the context of Internet of Thing. In [34], besides presenting a DQ methodology, the authors review DQ dimensions and other methodologies to assess DQ. In addition, some PS relate DQ with a particular domain, for example, big data [24, 46], data warehouses [49] and open data [22]. Moreover, [56] studies the mapping between metadata and DQ issues, in particular, the connection between metadata (such as count of rows, count of nulls, count of values and count of value pattern), and data errors. In other matters, [31, 48] apply a systematic literature review, the former to find works that consider DQ requirements during the process of developing an information system, the latter presents our preliminary results of review on DQ and context. In addition, in [8] the authors compare methodologies that have been proposed in the literature for IQ assessment and improvement.

Among the PS presenting methodologies, most of them are focused on methodologies for DQ (or IQ), specifically considering evaluation and improvement tasks [2, 4, 8, 18, 32, 34, 37, 47, 53, 54]. As seen before, [53] describes a DQ framework, but it also includes a DQ methodology. [37] employs a particular methodology, called Six-Sigma [82], to define critical factors to quality, measure current quality level, analyze deficiencies in information and identify the root causes of poor information. With the same purpose, [2] considers a methodology to build a DQ adapter module, which selects the best configuration for the DQ assessment. In turn, a data-focused methodology, based on DQ actions, is used in [4] to get smart data and to obtain more value for the datasets. Additionally, with
another focus, a methodology for selecting sources in the linked data domain is developed in [20]. In this work context-dependent DQ is taken into account, according to different DQ dimensions.

We also find works that present case studies focusing on the analysis of a particular problem. Firstly, [16] evaluates how data governance supports data-driven decision-making in asset management organizations. Also, it investigates the impact of data governance on DQ. Otherwise, the goal of [22, 23] is to examine the quality of open data in public sector, while [3] dives into the problem of re-defining traditional DQ dimensions and DQ metrics in the big data domain.

Regarding metrics, [30] investigates how DQ metrics for the DQ dimensions completeness, validity, and currency can be aggregated to derive an indicator for the accuracy dimension. In [35], a set of requirements for DQ metrics is presented. By last, [15] proposes a visual analytic approach that enables data analysts to utilize and customize DQ metrics, in order to facilitate DQ assessment of their specific datasets. In relation to works that focus on architectures we have two proposals. [51] presents a system architecture that includes mechanisms for DQ assessment and security, while [18] provides a DQ assessment architecture that manage streaming data in a context-aware manner at different levels of granularity.

Phases and steps of the CDQM. We also analyze the PS according to the CDQM steps that are addressed in the proposals, regardless of whether or not the PS propose a DQ methodology/process. In Fig. 6, we see the amount of PS that address each of the steps. We observe that many PS deal with one or several steps, only 1 PS do not focus on any of the steps (refered as N/A), and none of the PS addresses all the considered steps at the same time.

![Fig. 6. Total of PS by steps of CDQM.](image)

The least addressed phase is state reconstruction (step ST1), where all the relevant elements in an organization are collected, for example applications, user roles, databases, business processes, etc. (called contextual information by Batini et al. [61]). This result is in line with what is mentioned by Batini et al. in their analysis of DQ methodologies [61]. They point out that DQ methodologies, in general, assume that organizations have documents with contextual information that can be taken into account at later steps and, therefore, neglect this step.

The most addressed phase is assessment (which consists of ST2, ST3 and ST4), and in particular, most PS address ST4. At ST2, data analysis, the PS analyze relevant data identified at ST1. Some PS perform data profiling tasks at this step, which allows obtaining a first approximation to the quality of the data. The PS related to ST3, DQ requirements analysis, identify, analyze, and/or prioritize DQ requirements. They highlight the importance of classifying DQ requirements, which conditions the selection of DQ dimensions and the definition of DQ metrics at ST4. At ST4, PS deal with the definition of the DQ model and then, the measurement and assessment of DQ. In
Fig. 7. Total of PS that consider contextual DQ dimensions (a) and contextual DQ metrics (b).

particular, it is in this step that the contextual aspects of DQ dimensions are most discussed. Interestingly, some PS address measurement and/or assessment issues without proposing a DQ model, for example [11].

Concerning the improvement phase, less works concern its steps (ST5, ST6 and ST7). Actually, with so many PS focused on ST4, we expected to found more proposals dealing with this phase. Interestingly, we did not identify any PS that address, at the same time, the last 4 steps of CDQM. In particular, some PS analyze the causes of DQ problems (at ST5) without dealing with DQ assessment (at ST4) [22, 23, 26, 27], and other PS analyze improvement strategies (at ST6) without considering causes of DQ problems (at ST5) [4, 8, 20, 24, 32, 37, 43, 47, 53].

In summary, based on the results shown in Fig. 6, researchers have been mainly focused on measuring and assessing DQ, without delving into the causes of the results obtained, or the possibilities of DQ improvement. Further description of the PS addressing each step is presented in Subsections 4.3 and 4.4.

DQ dimensions. In 72% of PS (i.e. 42 out of 58 PS), authors affirm that there is a set of DQ dimensions that are contextual, i.e. they depend on context (see Fig. 7(a)). In fact, the authors present different arguments to justify the contextual characteristic of some DQ dimensions (see Fig. 8). For example, most of the PS that present contextual DQ dimensions justify them by referencing relevant bibliography of the DQ area, where it is already said that such DQ dimensions are contextual. Two examples of the referenced works are the classification of Wang & Strong [100] and the bibliographic review of Ge & Helfert [75, 100]. Other authors refer to the ISO/IEC 25012 standard [98] to affirm that certain dimensions are contextual. In these two cases, the authors only differentiate the contextual DQ dimensions from the non-contextual ones by taking definitions from the bibliography, without delving into the difference between the two types of DQ dimensions. Other PS considered in this analysis, indicate which are the elements that influence certain DQ dimensions, making them contextual. For instance, several PS include different aspects of the user (data profile, location, requirements, etc.) and data and/or information in use. To a lesser extent some aspects of the application domain are also mentioned, such as rules and constraints. In addition, DQ requirements, metadata, the task at hand, and the concept of utility are also considered.

Once the contextualizing elements are determined, it is interesting to identify the set of contextual DQ dimensions. We remind that in the classification of Wang & Strong [100] the DQ dimensions proposed as contextual are: relevancy, value-added, timeliness, completeness, and amount of data. In Table 3, we present the 5 DQ dimensions more commonly considered contextual by the authors of the PS analyzed: namely accuracy, completeness, timeliness, consistency, and relevancy, and we present the elements that, according to the PS, determine the contextual characteristics of these DQ dimensions. For example, this means that according to PS [17] the DQ dimension called Accuracy is contextualized by the task at hand, while in paper [35], the authors suggest that the same DQ dimension is contextualized by DQ requirements. We can see that, for both proposals, the accuracy dimension is context-dependent, however there is no agreement about which element contextualizes it. Note that when we mention the bibliography as a contextualizing element, we mean that the authors only reference bibliography to justify contextual aspects of DQ dimensions.

ACM J. Data Inform. Quality
We remark that accuracy and consistency dimensions were not considered as contextual in the classification of Wang & Strong, which is contradicted by our findings. Indeed, as seen in Table 3, nowadays both DQ dimensions, together with completeness, are the three DQ dimensions most recognized as contextual by researchers. In addition, Table 3 also highlights that DQ dimensions considered contextual are conditioned especially by the users, i.e. users characteristics, their data/quality requirements, their data profiles, their roles, etc. Then, according to this result, contextual DQ dimensions measure from different perspectives (e.g. accuracy, relevancy or completeness), the suitability of the data at hand (the data whose quality is evaluated), regarding the characteristics and needs of the users.

DQ metrics. In this analysis, we consider PS that define, use or analyze contextual DQ metrics. As shown in Fig. 7(b), they represent 38% of PS (i.e. 22 out of 58 PS). According to Heinrich et al. [76], DQ metrics need to be adaptable to the context of a particular application. Each contextual DQ metric takes into account different contextual aspects, since it strongly depends on the proposal. Specifically, DQ metrics are influenced by the type of data being measured and how they are measured, among others. An example of this type of proposal is [40], where the authors define DQ metrics that are applied taking into account other data (called external context), that give context to the data at hand.

Once the SLR execution was completed, we focused on the analysis of the context and its influence on DQ activities. This analysis resulted in a taxonomy, which is presented in the next section.

4 Analysis of the Contexts Identified in the Literature

Many works affirm that DQ depends on the context, but what is exactly the context?

In Subsection 2.2, we introduced Bazire and Brézillon’s model of context, representing the components of a situation and the different relations between those components [63]. According to Dey, such components capture any information that can be used to characterize the situations of an entity [72]. This notion of context is in accordance with the reviewed PS. Indeed, PS propose several components that make up context, represented with varied level of formalization, sometimes just mentioned, without an explicit definition.

Therefore, in the remaining of this paper we consider that the context is represented a the set of components and we investigate which components are proposed in the PS. We remark that, in this literature review, we do not aim to propose a context model, nor identify relationships among context components; however, note that the results of this SLR were used to devise a model of context that we proposed in [94].
Table 3. PS by the 5 DQ dimensions most commonly contextualized by different elements.

| Contextualizing element | Accuracy | Completeness | Timeliness | Consistency | Relevancy |
|-------------------------|----------|--------------|------------|-------------|-----------|
| **Utility**             | [6]      | [6]          | [6]        |             |           |
| **Task at hand**        | [17]     | [17]         | [17]       |             |           |
| **Metadata**            | [30, 40, 56] | [30, 56]   | [30, 56]   | [30, 40, 56] |           |
| **DQ requirements**     | [35]     | [35, 51, 55] | [35, 51]   | [35]        |           |
| **Application domain**  | [8, 47, 49] | [8, 15]    |            |             |           |
| **Data/Information**    | [5, 7, 40, 54] | [5–7]     | [6]        | [5, 40, 54] | [6]       |
| **User**                | [2, 9, 18, 34, 37, 38, 44] | [2, 18, 33, 37, 38, 44, 57] | [2, 18, 34, 37, 44, 45] | [28, 34, 44] |           |
| **ISO/IEC 25012**       | [31, 41, 46] | [31, 41, 46] | [46] |           | [31, 41, 46] |
| **Bibliography**        | [3, 24, 43, 50, 53] | [3, 14, 26, 29, 36, 39, 43, 50, 53, 58] | [14, 26, 29, 32, 36, 50] | [3, 43, 50, 53] | [14, 26, 29, 32, 50, 58] |

In this section we propose a taxonomy allowing to analyse PS according to three orthogonal analysis axes: (i) the level of formalization of the context, (ii) the components of context, and (iii) the steps of the DQM process where context is addressed.

The taxonomy is built following the method proposed in [86]. We mainly follow an empirical-to-conceptual approach for creating two of the taxonomy axes (called dimensions in [86]) and their characteristics, which are selected incrementally, as we progressed in the analysis of the PS. Characteristics of the third axis, i.e. DQM steps, are selected before examining the PS, with a conceptual-to-empirical approach. In particular, we use the steps of CDQM as characteristics, which are set before analysing the PS.

Next subsections describe the axes of analysis, their characteristics and some examples of use. Then, Subsection 4.4 presents the taxonomy, and Subsection 4.5 discusses the main findings.

4.1 Level of Formalization of Context Definition

Although all PS deal with the usage of context for DQ, many of them just mention its importance but do not provide a proper definition of context. In addition, when a definition is given, it is usually informal or even fuzzy. Therefore, as a first step for understanding the underlying notion of context, we classify the PS according to the existence of a proper definition of context and its level of formalization, in three categories: (i) formal definition, when the context is defined formally; (ii) informal definition, when the context is defined, but not formally, for example, in natural language; (iii) no definition, when the context is not defined (not even in natural language), but it is used implicitly. The latter occurs, for example, when the authors present the importance of data context, but they do not define what the context is.

For example, in [40] context is formally defined. The authors assess DQ in a data warehouse, taking into account the data context. The formalization of context and DQ metrics are given as logic rules. However, in [29], context definition is informal, given in natural language. The authors propose context clusters that can specify information about what (the type of the described information), where (spatial location), when (temporal location), and why (data motivations). In turn, in [53] a DQ model is specified, and its DQ dimensions (called

Note that this taxonomy, not being hierarchical, is presented in tabular format.

To avoid confusion, we use the term “dimensions” only to refer to DQ dimensions.
DQ characteristics) refer, according to the authors, to a specific context of use. However, such a context is not defined, either formally or informally.

Quantitatively, as illustrated in Fig. 9, 50% of the selected PS provide no context definition, while only 10% of the works (6 PS) present a formal context definition. We provide a detailed description of the PS formalizing context in a preliminary report [95].

The classification of PS according to this analysis axis is indicated by text styles and colors in Table 4, specifically, underline references correspond to formal definitions, italic-gray references to informal definitions, and bold-black ones to no definition.

4.2 Context Components

The state of the art revealed that although there are general operational definitions of context and context-aware computing [72], context representation is neglected in DQM. In particular, we highlight conclusions of Bertossi et al. [11] who report that the literature only deals with obvious contextual aspects of data, like geographic and temporal dimensions. As Bolchini et al. [65] suggest, other contextual aspects should be specified, e.g. users (person, application, device, etc.), presentations (system capabilities to adapt content presentation to different channels/devices), communities (set of relevant variables shared by a group of peers), and data tailoring (only selecting relevant data, functionalities and services). Preferences, documents content, DQ requirements and domain rules also emerge as important contextual aspects (a preliminary review can be found in [93]). Thus, context not only fits a single perspective, but could be defined by elements taken from different perspectives (user, application domain, data tailoring, etc.).

We remark that, in this literature review, we are only interested in identifying which are the elements that make up the context, called context components, but we do not aim to study their interrelationships nor proposing a model.

Among the selected PS only 9% do not propose any context component. In the rest, 91% of PS, we identified or deduced (when the authors do not define the context, but suggest that DQ depends on certain element) the context components suggested. We iteratively identify components in the PS and group close ones, getting to the 10 categories listed in the first column of Table 4.

Below we present each characteristic, giving a definition (in some cases abstracted from the literature) and describing some examples of use.

**DQ requirements.** This characteristic specifies *quality requirements that data must satisfy* [11]. So that, at the time of a query, users can retrieve data of a specific quality, i.e. within some acceptable range of DQ values in accordance with DQ requirements [101]. For example, in [41], a DQ requirement says that the soundness and the validity of data for the analysis must be the highest. According to the authors, this DQ requirement represents a deep concern about the extent to which data is good enough to perform reliable analysis.
Table 4. Taxonomy: PS by level of context formalization (underline formal, italic-gray informal, and bold-black no definition), context components and the steps of DQM where context is addressed

Data filtering needs. This characteristic concerns requirements that express specific needs about the data to be selected for a particular task. For example, filtering non-null data from an attribute for assessing their accuracy [34].

System requirements. In this case, we refer to functional and non-functional system requirements [24]. These requirements describe a function that must be performed by a system or a system element (functional requirement), or a required constraint on the functional effects of the software (non-functional requirement). Examples of big data system requirements [24] concern: data capability (network and storage requirements, e.g. the system must support PostgreSQL and MongoDB), data source (different characteristics of data sources, e.g. the system must collect data from sensors), data transformation (data processing and analysis, e.g. the system must support batch), data consumer (different characteristics of data consumers, including visualization, e.g. the system must support textual results), and data lifecycle (data lifecycle management functionality, e.g. the system must support DQM).

Business rules. This characteristic defines relationships among a set of attribute values, which derives a set of integrity constraints that accurately reflect an organization’s policies and domain semantics [67]. Contrarily to data filtering needs, in general, business rules are independent of the task on hand. They can take different forms, for example, in [41], they are simply constraints defined over data, while in [26] they are policies that govern business actions, which in turn result in constraints on data relationships and values. In [41], the authors affirm that for a big data quality project it is necessary to gather, together with the DQ requirements, the business rules that the data must satisfy.

Application domain. This characteristic specifies the domain of the evaluated data. Indeed, many PS consider some aspects of the application domain, in particular, for the definition of DQ metrics. For instance, the authors in [49] focus on the application domain identification and then, define DQ metrics around the information obtained in such domain, which results in domain-specific DQ metrics.
Task at hand. For this characteristic, as many PS, we consider the definition of Wang & Strong [103], as the task carried out with the data whose quality is evaluated. Many PS argue that DQ must be evaluated within the context of the task at hand. As an example, the proposal in [17] indicates that DQM for big data should prioritize those DQ dimensions really addressing the DQ requirements for the task at hand. In particular, the task at hand in this PS is to measure the level of DQ in big data.

User characteristics. This characteristic is defined as the type of users with the same profile and/or preferences. PS suggest several features of the user that provide context, and they arise from the user profile (general aspects of the user, such as his geographical location, language, profession, age, etc.), and/or the user preferences. For example, a proposal for e-government environment [1] shows the relationships among the perceived IQ and the perception, satisfaction, trust and demographic characteristics (e.g. identification, gender, age, education, and internet experience) of the users.

Metadata. Metadata is defined as data about data [73]. Some PS include metadata in the data context. In particular, the proposal in [56] investigates the connection between metadata and DQ issues to assist data scientists in DQ assessment. It provides an exhaustive analysis of the bibliography, where DQ dimensions and metadata are related. Metadata, such as count of rows, count of nulls, count of values, and count of value pattern are used to generate DQ rules. According to the authors, these rules can be used in conjunction with other data cleaning solutions.

DQ values. This characteristic concern the values obtained during DQ measurement. Specifically, DQ values give context to other DQ measures. For example, sometimes it is necessary to aggregate a set of DQ values to obtain an overall DQ measure. In fact, the authors of [30] highlight the need of a better analysis of the inter-dependencies between DQ dimensions (rarely exploited in the literature), and investigate how DQ values for the DQ dimensions completeness, validity, and currency can be aggregated to derive an indicator for accuracy. Although DQ values are a specific type of metadata, we consider important to have a separated characteristic for them, as they represent a key concept for DQM.

Other data. This characteristic concerns the data that give context to the data whose quality is evaluated. For instance, in a relational database, data from one table could give context to other tables. Examples of this case can be seen in [11, 40], where the authors use other data as contextual information, to verify if the data at hand have good quality. That is, the contextual information allows validating the data at hand.
We found that data are mostly contextualized by DQ requirements, followed by data filtering needs, application domain, other data and business rules. In turn, we cannot fail to mention that there are 5 PS ([22, 23, 25, 48, 50]) where the importance of considering the context for DQM is highlighted, however they do not mention which are its components.

Next, we discuss different context components representations, identified in the selected PS. We do not create a separate analysis axis in the taxonomy because this issue is sparsely addressed in the PS, which according to the method applied, would not result in interesting PS analysis groups [86]. Indeed, of the PS that propose components to form the context, only 55% suggest a representation of these components.

4.2.1 Context Components Representation. In this subsection, we describe the different proposed ways to represent each component of the context, beyond the level of formalization used to define it. We identified three kind of representations: entities, numerical indicators and rules. In turn, we also identify different ways of proposing the numerical indicators and expressing the rules.

The following paragraphs describe each kind of representations and their usage for representing specific context components, which is also summarized in Fig. 11. In addition, the number of PS dealing with each representation is shown in Fig. 10.

Entities. They are used to represent context components such as data filtering needs [20, 21], application domain [54] and user characteristics [54]. The context model of [54] is based in Dey’s definition [72] and application domain and user characteristics are represented as entities. In [20, 21], concepts that are entity categories identified by URIs, are used for representing data filtering needs. Notably, in [21], an extension of [20], authors use SKOS\(^5\) for defining, publishing, and sharing concepts over the Web. SKOS concepts represent different data filtering needs that may be required by users to select linked data sources.

Numerical indicators. They are proposed in the following ways:

- **Thresholds**: These are used to express numerical indicators that represent DQ requirements [20, 21, 28, 33]), application domains [35], and other data [2]. For example, in [21], users indicate thresholds for DQ values, then a ranked list of relevant sources are returned according to the quality thresholds.

- **Functions**: In this case, numerical indicators are calculated by applying functions and used to represent metadata [56] and DQ values [3, 30]. In [56], an indicator function maps the set of metadata to \(\{0,1\}\), where 1 denotes an incorrect data and 0 denotes “clean data”. The indicator function is designed as a product of the results of DQ metrics, namely, completeness, validity and currency in [30], while completeness, readability and usefulness are used in [3]. In the latter, each DQ metric has an associated weight, a scalar value between 0 and 1.

Rules. They are used to represent different context components, in particular we identified the following ways of expressing such rules:

- **Natural language**: It is used for expressing rules that represent DQ requirements [24, 32, 33, 35, 37, 43], data filtering needs [4, 32], business rules [4, 55], system requirements [24] and other data [4, 37]. They are the most used by PS, especially to represent DQ requirements. For instance, in [37], DQ requirements of customers, in natural language, are linked to DQ dimensions, while in [35] they are used to develop DQ metrics.

- **Logical expressions**: Also logical expressions are used for expressing rules that represent DQ requirements [10, 11, 42, 57], business rules [10, 49], application domain [49], user characteristics [49] and other data [10, 11, 40, 42]. Logical expressions, along with natural language, are the most widely used to express

\(^{5}\)SKOS (Simple Knowledge Organization System) is a W3C recommendation. https://www.w3.org/TR/skos-reference/
rules that represent context components. (see Fig. 11). For instance, in [49] the Datalog language is used to represent context components through a set of logical rules.

- **Scripts**: To a lesser extent, scripts are used to express syntactic or semantic rules, in particular for representing data filtering needs [15], business rules [5] and application domains [15]. As an example, in [15], characteristics of the application domain and data filtering needs arising from constraints and dependencies, and they are expressed through semantic and syntactic rules. In turn, in [5], business rules are applied as semantic rules over tuples of relational tables.

- **SQL statements**: These also express rules that represent DQ requirements [12, 13] and business rules [47]. In [12], an extension of [13], it is argued that SQL statements enable defining data objects using “select” clauses, and defining DQ requirements using “where” ones. In [47] SQL statements are used to represent functional dependencies that arise from business rules. According to the authors, conditional functional dependencies have a syntax that can clearly represent context rules to document dependencies between attributes. This syntax also makes easier to translate rules to SQL queries that look for records that violate the rules.
Considering the different aspects of the context presented above, we present in Fig. 11 the contexts components (framed), their definitions (penultimate column) and representations (last column), which are the foundations for a context definition and the specification of its components. The detail of PS addressing each characteristic of context components is shown in Table 4, where each line represents a characteristic.

4.3 Steps of DQM Process

In this subsection, we analyze the usage of context within the DQM process. Since, to the best of our knowledge, there is no DQ methodology that exploits data context at all its steps, we identify and analyze in the literature which steps of the DQM process are related to context.

We consider that context is related to a step in two cases: i) if the authors use context in the DQ activities performed at the step, or ii) if they mention that context is important for the considered step. As indicated before, the steps of the CDQM methodology are used as categories for this analysis axis. They were introduced in Subsection 2.1.2. Next paragraphs we provide some examples of PS arguing the importance of context at each step.

4.3.1 State reconstruction phase. It has a single step, also called state reconstruction.

ST1: State reconstruction. Many PS argue the importance of identifying context from the initial step. For example, in [4] all the necessary elements to carry out DQ corrections (specially business rules and data filtering needs) are identified at the first step. In addition, authors claim the importance of identifying business rules at the initial step, since when DQ actions are aligned with the business rules, there are more possibilities to exploit the value of data.

4.3.2 Measurement/Assessment phase. This phase has 3 steps, namely, data analysis, DQ requirements analysis and DQ measurement/assessment.

ST2: Data analysis. According to several PS, context should be considered within data analysis and profiling tasks. For example, authors of [15] suggest that information about the application domain should be integrated into the data profiling process to make the visualizations and other quantitative representations more adequate to usage. Therefore, they propose a visual data analysis tool that provides custom DQ checks (i.e. evaluation of tuples returning DQ values as boolean values), in order to assess different application domain-specific constraints and dependencies.

ST3: DQ requirements analysis. DQ requirements are the characteristic of the context component axis the most used in PS. Consequently, the step dealing with their analysis, is tightly related to context. As an example, in [32], a DQ methodology with 6 phases is proposed, which aims to develop a DQ assessment tool. At the second phase of such methodology, DQ requirements are extracted, analyzed and prioritized, taking into account different needs, in particular the data filtering needs of stakeholders.

ST4: DQ measurement/assessment. This is the step where context is the most addressed in the PS. Indeed, 46 out of 58 PS argue the need of handling context for DQ measurement/assessment. As an example, in [20], data sources of good quality are selected in the linked data domain. For this, DQ values obtained in the DQ measurement are associated with the sources. Later, DQ assessment is performed by comparing these DQ values with thresholds, that represent DQ requirements.

4.3.3 Improvement phase. It consists of 3 steps, namely, error causes identification, strategies and techniques selection, and costs evaluation.

ST5: Error causes identification. Some, less numerous PS, address context at this step. In general, the causes or sources of the DQ problems are analyzed taking into account the context. In [38] the authors claim that a
comprehensive DQ problems analysis under the application domain considering DQ requirements, business processes, users, among others, ensures that data correctly reflects the real situation. Based on this, they propose a model that allows them to manage DQ problems. In particular, they identify the problems sources, for example software quality defects and man-made problem.

**ST6: Strategies and techniques selection.** Several PS argue the importance of context for DQ improvement plan and its implementation. In particular, strategies can vary depending on the organization, on the data, as well as on the existing and target DQ maturity levels. For example, according to the authors of [53], the organization’s target DQ maturity level must be identified to develop the DQ improvement plan. They argue that DQ requirements play a very important role at all levels of DQ maturity, either because they are unknown to users or because DQ models are defined based on them.

**ST7: Costs evaluation.** In several PS, context influences the analysis of prevention costs, detection costs and repairing costs. For example, [39] focuses on evaluating the costs of improving data completeness. For the authors, data completeness implies that data are not missing and are sufficient for the task at hand. Therefore, the evaluation of improvement costs depends on the task at hand.

The detail of PS addressing each step of the DQM process is shown in Table 4, where each column represents a step.

### 4.4 Taxonomy of Analyzed PS

The taxonomy is depicted in Table 4. The first axis, level of formalization of context definition, is represented with text styles and colors (PS proposing a formal definition of context are listed with underline, those providing an informal definition in italic-gray, and those with no definition in bold-black). The second axis, context components, are represented in lines, and the third axis, steps of DQM process, are represented in columns.

We firstly observe that the only 6 PS presenting a context formalization, all address step ST4, and only one of them also address ST5.

The relationships among context components and DQM steps are richer. Fig. 12 quantitatively highlights the components used at each step. Below, we discuss some findings, organizing the discussion by DQM steps.

**ST1: State reconstruction.** We can see in Fig. 12, that most components are used at this step, the dominant ones being data filtering needs and business rules. This makes sense, as state reconstruction is the phase that collects contextual information on organizational processes and services. We remark that DQ requirements are not addressed at this step, while they are the dominant component in the other steps. The task at hand is not considered either, which is unexpected, since it is typically part of the information on organizational processes elicited at this step.

**ST2: Data analysis.** Surprisingly, although CDQM does not consider the analysis of DQ requirements until ST3, the most used components at ST2 are DQ requirements. They are mainly used for DQ requirements-driven data analysis. Metadata also plays an important role at this step. Conversely, user characteristics is the only component that is not considered probably because data analysis, in particular data profiling, is performed as objectively as possible.

**ST3: DQ requirements analysis.** As expected, DQ requirements is the dominant component, being addressed in more than half of the concerned PS, as evidenced in Fig. 12. We remark the absence of business rules among the studied components, which contradicts the intuition that DQ requirements are strongly based on business rules.
Fig. 12. Context components at each CDQM step.

**ST4: DQ measurement/assessment.** This step is the most addressed by the PS, with special attention to the tasks of defining the DQ model and DQ measurement/assessment. All components are used, the dominant one also being DQ requirements. We remark that most usages of user characteristics arise in this step.

**ST5: Error causes identification.** DQ requirements, data filtering needs, application domain and other data, are the sole components used by PS at this step. We remark the unexpected absence of DQ values, supposed to guide improvement tasks.

**ST6: Strategies and techniques selection.** DQ requirements and data filtering needs are also leading this step. It is also striking that DQ values are not taken into account in any PS, since these indicators could condition the actions to be taken on the data, processes, users, etc., in the work scenario.

**ST7: Costs evaluation.** As observed in Fig. 12, the context components most considered at this step are also DQ requirements and data filtering needs. We consider that this result is coherent, since the costs associated with poor quality must be evaluated, mainly, with respect to DQ requirements defined on data at hand, and data filtering needs related with the task at hand.

In light of this taxonomy, the next subsection presents the lessons learned.

### 4.5 Discussion

In this subsection we discuss the main findings resulting from the taxonomy.

Firstly, we point out that while in CDQM (the sole DQM methodology dealing with context) context is identified at ST1 and never updated or exploited in the final steps, our study revealed that this is not enough. Indeed, many PS claim that context can change along DQM cycle. For example, data users may change, their tasks may change, also introducing new DQ requirements and data filtering needs, etc. We believe that there is a major need to develop new DQM methodologies that both, are guided by context at each step, and enhance results of each step.
to discover new context components, which implies having an up-to-date context. This leads us to propose a DQM methodology in [96].

In particular, the measurement/assessment phase, especially step ST4, are the most addressed in the PS. Indeed, 46 out of 58 PS point that ST4 is context-dependent. All categories of context components are considered by many of those works. Furthermore, all PS that propose a context formal definition address activities of ST4 (one of them also addresses ST5). This attention from the literature evidences the subjective nature of the activities developed at ST4, and the need to handle context to perform such activities.

Another important lack in the literature is the shortage of context formalization. Indeed, only 6 PS present a formal definition of context, and more alarmingly, half of the PS present no context definition at all, which magnifies this lack.

In addition, we highlight that while most PS identify context components (53 out of 58), only 55% of them provide a representation for such components. Among the proposed representations, rules are the most used, especially rules expressed in natural language and as logical expressions. They are used for representing 7 categories of components. Undeniably, components representation is an open research issue, which is tightly related to the lack of context formalization.

Among categories of context components, DQ requirements is the most addressed in PS (36 out of 58 PS deal with them). DQ requirements are used in all steps except at ST1, though participating in almost the entire DQM cycle. Besides, they are proposed as the most relevant context components at each step in which they participate. This is a natural result, since in DQM, DQ requirements are the starting point for many tasks, and particularly for DQ measurement and DQ assessment. But we also highlight that all categories of context components are used in many DQM steps.

Finally, but not surprisingly, the 5 PS that do not propose any context component, do not propose a context definition either. However, they point out the importance of context for all steps of DQM.

5 Answers to Research Questions

Based on the obtained findings, in this section we answer the three research questions that guided the systematic literature review. Finally, we present the research challenges that arise from these answers.

5.1 RQ1: How is Context Used in Data Quality Models?

In the PS that address DQ models [12, 13, 17, 29, 33, 38, 41, 43, 52, 53], we mainly identify that they include context by:

- Including contextual DQ dimensions in the dimensions set that will compose the DQ model. In most cases, the contextual DQ dimensions are not selected consciously, by considering the context of the data evaluated, but rather by choosing DQ dimensions suggested by the bibliography.
- Defining contextual DQ metrics, through the inclusion of context components in their descriptions or algorithms. Business rules, DQ requirements and data filtering needs are among those most considered (also at dimensions selection activity). Other components, such as the tasks developed by the users and their characteristics, are not explicitly considered, in spite of being important aspects that frequently guide DQ experts. This may be because these aspects are not seen as context components by several authors.
- Considering the context through DQ requirements. DQ requirements are sometimes included as part of the DQ model, other times as an input for building the model, and other times for evaluating DQ after measurement execution. These requirements may come from business rules, users’ data or user preferences.

We consider that although the literature demonstrates great interest in the use of context, there has not yet been an in-depth discussion on the identification of context components and how they allow us to define appropriate
DQ models. We believe that the lack of study of the influence of context on the different concepts around a DQ model (i.e. dimensions, metrics, measurement and assessment), often causes that the resulting DQ models do not help us to address DQ improvement strategies or lead to strategies that are not exhaustive.

5.2 RQ2: How is Context Used within Quality Metrics for the main Data Quality Dimensions?

In Subsection 3.3 we already addressed contextual DQ metrics identified in the PS. Now, looking for answering this research question, we focus on a more detailed analysis.

In the answer to the previous question, we mentioned that many times DQ metrics include some context components in their algorithms. In addition, we highlight that several DQ metrics are defined as a combination of DQ values that were obtained in previous DQ measurements (i.e. DQ metadata). They are called aggregate metrics and are generally defined to give a global DQ measure. These DQ metrics can be seen as contextual if we consider the used DQ values as part of the context of data.

We observed that sometimes DQ metrics arise in the data profiling activity, and many times they are proposed based on the expertise of the data analyst and/or the DQ expert. Although we noted that literature analyzes certain context components (such as business rules and DQ requirements) when defining DQ metrics, we insist that other components, such as data user experience and other relevant data (that provide knowledge of the evaluated data), are still not sufficiently taken into account. This survey allows us to affirm that the need to delve deeper into the contextual aspects of DQ metrics is not new, since it has been raised in the bibliography for a long time. The authors even propose the context components that they consider most influential, however, no concrete proposals have been found on how to include these context components in DQ metrics definitions.

5.3 RQ3: How is Context Used within Data Quality Concepts?

According to our review, we identify that the main DQ concepts that usually consider data context are measurement and assessment activities, DQ methodologies, DQ requirements and DQ problems.

The findings regarding this question are somehow contradictory with respect to those identified for the previous one, because literature claims that DQ activities that most depend on the context are measurement and assessment, however, the contextual characteristics of these activities are not being widely exploited when defining DQ metrics.

The taxonomy presented in Table 4 clearly shows a need to address context in all steps of a DQ methodology, through the consideration of different context components, whether formally or informally. However, literature presents few proposals that use context (such as the CDQM methodology), not explicitly and only in the initial steps.

DQ requirements are the most considered in DQ issues, but the activities in which they can be refined, such as user requirements analysis, are not always addressed. We think that this is an important limitation of the literature, since ignoring some DQ requirements during DQM may result in a costly quality management that rarely has the expected results.

Regarding DQ problems, we saw that they are generally well identified. In fact, investigations arise based on them. However, once identified, few works delve into their causes. This lack in the existing research is also relevant, since not considering causes of DQ problems during DQM could lead to useless costs and efforts. Specifically, not investigating the DQ problems causes may imply incorrect selection of strategies and techniques to improve DQ.

In our opinion, the misuse (or lack of use) of context in DQ concepts is due to the lack of a fully context-dependent DQ methodology. This means that at each step of this methodology, the influence of the context on DQ activities should be analyzed, as well as the possibility that new context components emerge while the DQM process is carried out.
5.4 Identified Research Challenges

After analyzing in depth the answers we obtained for the research questions, as well as the collateral findings obtained throughout the process for obtaining these answers, we summarize the main research challenges that arose from our study.

We believe that there is an urgent need for developing context-aware methodologies that not only are guided by the context at each step, but also generate results at each step that lead to discover new context components. There is a need of methodologies that help DQ experts to conduct reliable context-aware DQM processes that are easy to maintain and reproduce, leaving as result DQ improvements and data provenance.

Surprisingly, we found that there is a overlooked research issue, and it is related to DQ requirements: user requirements analysis as part of DQM processes. In spite of the great importance of DQ requirements in the works we analyzed, the activities in which they can be refined, basically user requirements analysis, are barely addressed.

We also found that there is a promising research line that studies and propose solutions for including context components explicitly in the DQ metrics. This inclusion would allow DQ metrics to manage context as an independent asset, such that any variation in the context would automatically affect the metric behaviour. In the same manner, the explicit inclusion of context components into other DQ concepts, such as DQ dimensions definitions, or specific DQ tasks, would have the same consequences.

Finally, but not least, we identify the need for DQ context formalization. This formalization is essential to propose general solutions for DQM, allowing to manage context-specific DQ dimensions. In order to address context formalization the first step to take is to give a rigorous definition and representation of the context. As shown in this study, it is difficult to find in the literature a comprehensive definition of context for DQM, and much more difficult to find works that propose representations and formalizations for context.

6 Conclusions

This paper reports the protocol and results of a SLR on context in DQM. The SLR methodology presented in Section 3 allowed the selection of 58 papers based on their relevance with respect to the usage of context in DQM. Half of them, i.e. 29 PS, do not present any context definition, formal or informal, and only 6 PS present a formal context definition. Furthermore, even if many context components have been proposed, most works do not propose a representation for them. These findings indicate that context formalization is an open research issue.

The well-established hypothesis stating that DQ assessment must be guided by DQ requirements, is aligned with our findings, showing DQ requirements as the most used context component. However, we found enough evidences that other components are important. Therefore, identifying all context components at early steps of the DQM process would enable a global vision of the business, being able to obtain more precise DQ assessment. Although we found that the literature addresses the aforementioned context components, it does not delve into the level of influence that each of them has on the different DQ concepts (i.e. dimensions, metrics, measurement and assessment), related to DQ models. Indeed, there are very few works that briefly specify the use of context in DQ metrics, although they are largely recognized to be highly context-dependent.

There are even fewer works that propose DQ improvement solutions, and generally without investigating the causes of DQ problems. We believe that it is mandatory to know the context to be able to find the causes of DQ problems and propose adequate solutions. More generally, even though context for DQ is starting to draw attention, we claim that more context-aware and context-driven solutions are necessary, and this, for all DQM steps.

All these activities must be managed within a DQM methodology guided by the data context, whose steps integrate activities for eliciting, using, and enhancing context, without forgetting that the context is not static,
but changes according to the data usage. Indeed, the development of DQM methodologies better exploiting context at all their steps is another open research issue.

Finally, we highlight that the SLR was instrumental to the identification of the context components, because specific methodological guidelines allowed us to carry out exhaustive and incremental analyzes of the bibliography. This task helped us develop a taxonomy that proposes future lines of research.

Inspired by the findings of this bibliographic study, we proposed a model of context, made up of the reviewed components, and also studying the relationships among them and with the DQ model [94]. We also proposed a context-aware DQM (CaDQM) methodology [96].

Currently, we are working in a logical representation of context components. In addition, some user studies are being conducted in the health domain.

Our long-term outlook is to apply our context-aware DQM methodology in a large-scale project, in the domain of Digital Government with several public services of Uruguay.

**PRIMARY STUDIES SELECTED BY THE SLR**

[1] M. Shakaib Akram and A. Malik. 2012. Evaluating citizens’ readiness to embrace e-government services. In 13th Annual Int. Conf. on Digital Government Research, J.C. Bertot, L.F. Luna-Reyes, and S. Melloul (Eds.). ACM, 58–67. https://doi.org/10.1145/2307729.2307740

[2] D. Ardagna, C. Cappiello, W. Samà, and M. Vitali. 2018. Context-aware data quality assessment for big data. Future Generation Computer Systems 89 (2018), 548–562. https://doi.org/10.1016/J.FUTURE.2018.07.014

[3] F. Arolo and A. Vaisman. 2018. Data Quality in a Big Data Context. In Advances in Databases and Information Systems - 22nd European Conference, ADBIS 2018, Proceedings (LNCS, Vol. 11019), A. Benczúr, B. Thalheim, and T. Horváth (Eds.). Springer, Berlin, Heidelberg, 159–172. https://doi.org/10.1007/978-3-319-98398-1_11

[4] M.T. Baldassarre, I. Caballero, D. Caivano, B. Rivas Garcia, and M. Piattini. 2018. From big data to smart data: a data quality perspective. In Proc. of the 1st ACM SIGSOFT International Workshop on Ensemble-Based Software Engineering, A. Bucchiarone, M. Mongiello, F. Nocera, and M. Sheng (Eds.). ACM, 19–24. https://doi.org/10.1145/3281022.3281026

[5] D. Barone, F. Stella, and C. Batini. 2010. Dependency Discovery in Data Quality. In Advanced Information Systems Engineering, 22nd International Conference, CAiSE 2010, Hammamet, Tunisia, June 7-9, 2010. Proceedings (LNCS, Vol. 6051), B. Pernici (Ed.). Springer, Berlin, Heidelberg, 53–67. https://doi.org/10.1007/978-3-642-13094-6_6

[6] C. Batini and M. Scannapieco. 2016. Data and Information Quality - Dimensions, Principles and Techniques. Springer, Chapter Information Quality in Use. https://doi.org/10.1007/978-3-319-24106-7

[7] C. Batini and M. Scannapieco. 2016. Data and Information Quality - Dimensions, Principles and Techniques. Springer, Chapter Data Quality Dimensions. https://doi.org/10.1007/978-3-319-24106-7

[8] C. Batini and M. Scannapieco. 2016. Data and Information Quality - Dimensions, Principles and Techniques. Springer, Chapter Methodologies for Information Quality Assessment and Improvement. https://doi.org/10.1007/978-3-319-24106-7

[9] M. Belhiah, M. Salim Benqatla, and B. Bounabat. 2015. Decision Support System for Implementing Data Quality Projects. In Data Management Technologies and Applications, Revised Selected Papers (Comm. in Comp. and Inf. Sc., Vol. 584), M. Helfert, A. Holzinger, O. Belo, and C. Francalanci (Eds.). Springer, 1–16. https://doi.org/10.1007/978-3-319-30162-4_1

[10] L. Bertossi and M. Milani. 2018. Ontological Multidimensional Data Models and Contextual Data Quality. ACM Journal of Data and Information Quality 9, 3 (2018), 14:1–14:36. https://doi.org/10.1145/3148239

[11] L. Bertossi, F. Rizziolo, and L. Jiang. 2010. Data Quality Is Context Dependent. In Enabling Real-Time Business Intelligence - 4th International Workshop, BITE 2010, Held at the 36th International Conference on Very Large Databases, VLDB 2010, Singapore, September 13, 2010, Revised Selected Papers (Lecture Notes in Business Information Processing, Vol. 84), Malu Castellanos, Umeshwar Dayal, and Volker Markl (Eds.). Springer, 52–67. https://doi.org/10.1007/978-3-642-22970-1_5

[12] Z. Bicevskis, J. Bicevskis, and I. Oditis. 2017. Models of Data Quality. In Information Technology for Management, Extended Selected Papers (Lecture Notes in Business Information Processing, Vol. 311), E. Ziemba (Ed.). Springer, 194–211. https://doi.org/10.1007/978-3-319-77721-4_11

[13] J. Bicevskis, Z. Bicevskis, and I. Oditis. 2017. Domain-Specific Characteristics of Data Quality. In Proc. of the 2017 Federated Conference on Computer Science and Information Systems, FedCIS 2017 (Annals of Computer Science and Information Systems, Vol. 11), M. Ganzha, L.A. Maciaszek, and M. Paprzycki (Eds.). 999–1003. https://doi.org/10.15439/2017F279

*Research project developed by the e-Government and Information and Knowledge Society Agency (AGESIC) of Uruguay: https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/
REFERENCES

[56] L. Visengrityeva and Z. Abedjan. 2020. Anatomy of Metadata for Data Curation. ACM Journal of Data and Information Quality 12, 3 (2020), 16:1–16:30. https://doi.org/10.1145/3371925

[57] J. Wang, N.J. Martin, and A. Poulouvasilis. 2011. An Ontology-Based Quality Framework for Data Integration. In Workshops on Business Informatics Research - BIR 2011 International Workshops and Doctoral Consortium, Revised Selected Papers (Lecture Notes in Business Information Processing, Vol. 100), L. Niedrite, R. Strazdina, and B. Wangler (Eds.). Springer, 196–208. https://doi.org/10.1007/978-3-642-29231-6_16

[58] M. Zárraga-Rodríguez and M.J. Álvarez. 2015. Experience: Information Dimensions Affecting Employees’ Perceptions Towards Being Well Informed. ACM Journal of Data and Information Quality 6, 2-3 (2015), 12:1–12:14. https://doi.org/10.1145/2774223

[59] G.D. Abowd and E.D. Mynatt. 2000. Charting past, present, and future research in ubiquitous computing. ACM Transactions on Computer-Human Interaction 7, 1 (2000), 29–58. https://doi.org/10.1145/344949.344988

[60] M. Baldauf, S. Dustdar, and F. Rosenberg. 2007. A survey on context-aware systems. International Journal of ad Hoc and Ubiquitous Computing 2, 4 (2007), 263–277. https://doi.org/10.1504/IJAHUC.2007.014070

[61] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino. 2009. Methodologies for data quality assessment and improvement. Comput. Surveys 41, 3 (2009), 16:1–16:52. https://doi.org/10.1145/1541880.1541883

[62] C. Bolchini, C. Curino, G. Orsi, E. Quintarelli, R. Rossato, F.A. Schreiber, and L. Tanca. 2009. And what can context do for data? ACM Journal of Data and Information Quality - Dimensions, Principles and Techniques. Springer. https://doi.org/10.1007/978-3-905691-48-5

[63] M. Bazire and P. Brézillon. 2005. Understanding Context Before Using It. In Modelling and Using Context, 5th International and Interdisciplinary Conference, CONTEXT 2005, Proceedings (LNCS, Vol. 3554), A. Dey, B. Kokinov, D. Leake, and R. Turner (Eds.). Springer, 29–40. https://doi.org/10.1007/11508373_3

[64] L. Bertossi. 2019. Database Repairs and Consistent Query Answering: Origins and Further Developments. In Proc. of the 38th ACM SIGMOD-SIGACT-SIGAI Symp. on Principles of Database Systems, PODS 2019, D. Suchu, S. Skritek, and C. Koch (Eds.). ACM, 48–58. https://doi.org/10.1145/3294052.3322190

[65] C. Bolchini, C. Curino, G. Orsi, E. Quintarelli, R. Rossato, F.A. Schreiber, and L. Tanca. 2009. And what can context do for data? Commun. ACM 52, 11 (2009), 136–140. https://doi.org/10.1145/1592761.1592793

[66] C. Cappiello, C. Francalanci, and B. Pernici. 2004. Data Quality Assessment from the Users Perspective. In Conceptual Modeling - ER 2004, 30th Int. Conf., ER 2004, Proceedings (LNCS, Vol. 3217), Ma. Jeusfeld, L. Delcambre, and T. Wang Ling (Eds.). Springer, 304–317. https://doi.org/10.1007/978-3-540-26466-7_23

[67] F. Chiang and R.J. Miller. 2008. Discovering data quality rules. Proc. of the VLDB Endowment 1, 1 (2008), 1166–1177. https://doi.org/10.14778/1453856.1453980

[68] X. Chu, I.F. Ilyas, S. Krishnan, and J. Wang. 2016. Data Cleaning: Overview and Emerging Challenges. In Proc. of the 2016 Int. Conf. on Management of Data, SIGMOD 2016, F. Ozcan, G. Koutrika, and S. Madden (Eds.). ACM, 2201–2206. https://doi.org/10.1145/2882903.2912574

[69] P. Ciaccia and R. Torlone. 2011. Modeling the Propagation of User Preferences. In Conceptual Modeling - ER 2011, 30th Int. Conf., ER 2011, Proceedings (LNCS, Vol. 6998), M. Jeusfeld, L. Delcambre, and T. Wang Ling (Eds.). Springer, 304–317. https://doi.org/10.1007/978-3-642-24606-7_23

[70] C. Cichy and S. Rass. 2019. An Overview of Data Quality Frameworks. In Data and Information Quality - Dimensions, Principles and Techniques. Springer. https://doi.org/10.1007/978-3-030-37453-2_30

[71] E. Duval. 2001. Metadata Standards: What, Who & Why. Journal of Universal Computer Science 7, 7 (2001), 591–601. https://doi.org/10.3217/JUCS-007-07-0591

[72] A. Even and G. Shankaranarayanan. 2009. Dual Assessment of Data Quality in Customer Databases. ACM Journal of Data and Information Quality 1, 3 (2009), 15:1–15:29. https://doi.org/10.1145/1659225.1659228

[73] M. Ge and M. Helfert. 2007. A Review of Information Quality Research - Develop a Research Agenda. In Proc. of the 12th Int. Conf. on Information Quality, MIT, M.A. Robbert, R. O’Hare, M. Markus, and B. Klein (Eds.). MIT, 76–91.

[74] B. Heinrich, M. Klier, and M. Kaisser. 2009. A Procedure to Develop Metrics for Currency and its Application in CRM. ACM Journal of Data and Information Quality 1, 1 (2009), 5:1–5:28. https://doi.org/10.1145/1515693.1515697

ACM J. Data Inform. Quality
[77] S. Issa, O. Adekunle, F. Handi, S. Si-Said Cherfi, M. Dumontier, and A. Zaveri. 2021. Knowledge Graph Completeness: A Systematic Literature Review. *IEEE Access* 9 (2021), 31322–31339. https://doi.org/10.1109/ACCESS.2021.3056622

[78] L. Jiang, A. Borgida, and J. Mylopoulos. 2008. Towards a Compositional Semantic Account of Data Quality Attributes. In *Conceptual Modeling - ER 2008, 27th Int. Conf. on Concep. Modeling, Proceedings (LNCS, Vol. 5231)*, Q. Li, S. Spaccapietra, E. Yu, and A. Olivé (Eds.). Springer, 55–68. https://doi.org/10.1007/978-3-540-87877-3_6

[79] J. Juran. 1999. *Juran’s quality handbook. 5th. Edition*. McGraw-Hill Companies, Inc.

[80] F. Kamei, I. Wiese, C. Lima, I. Polato, V. Nepomuceno, W. Ferreira, M. Ribeiro, C. Pena, B. Cartaxo, G. Pinto, and S. Soares. 2021. Grey Literature in Software Engineering: A critical review. *Inf. Softw. Technol.* 138 (2021), 106699. https://doi.org/10.1016/J.IINFOSOF.2021.106699

[81] B. Kitchenham. 2004. *Procedures for Performing Systematic Reviews*. Technical Report 2004. 1–26 pages.

[82] K. Linderman, R.G. Schroeder, S. Zaheer, and A.S. Choo. 2003. Six Sigma: a goal-theoretic perspective. *Journal of Operations Management* 21, 2 (2003), 193–203. https://doi.org/10.1016/S0272-6963(02)00087-6

[83] D. Loshin. 2011. *8 - Dimensions of Data Quality*. In *The Practitioner’s Guide to Data Quality Improvement*, D. Loshin (Ed.). Morgan Kaufmann, Boston, 129–146. https://doi.org/10.1016/B978-0-12-373717-5.00008-7

[84] S.S.G.S. Mylavarapu. 2020. *Context-Aware Quality Assessment of Structured and Unstructured Data*. Ph.D. Dissertation. Oklahoma State University, USA. Advisor(s) T. Johnson.

[85] M.P. Neely. 2005. *The Product Approach to Data Quality and Fitness for Use: A Framework for Analysis*. In *Proceedings of the 2005 Int. Conf. on Information Quality (MIT ICIQ Conference)*, Sponsored by Lockheed Martin, MIT, F. Naumann, M. Gertz, and S. Madnick (Eds.). MIT.

[86] R.C. Nickerson, U. Varshney, and J. Muntermann. 2013. A method for taxonomy development and its application in information systems. *European Journal of Inf. Syst.* 22, 3 (2013), 336–359. https://doi.org/10.1057/EJIS.2012.26

[87] F. Naumann, M. Gertz, and S. Madnick (Eds.). MIT. *Information Quality (MIT ICIQ Conference)*, Sponsored by Lockheed Martin, MIT, F. Naumann, M. Gertz, and S. Madnick (Eds.). MIT.

[88] K. Stefanidis, E. Pitoura, and P. Vassiliadis. 2011. Managing contextual preferences. *Decision support systems* 302–317. https://doi.org/10.1016/J.DSS.2004.12.006

[89] S. Issa, O. Adekunle, F. Handi, S. Si-Said Cherfi, M. Dumontier, and A. Zaveri. 2021. Knowledge Graph Completeness: A Systematic Literature Review. *IEEE Access* 9 (2021), 31322–31339. https://doi.org/10.1109/ACCESS.2021.3056622
In this appendix we list all the PS selected by applying the SLR. For each one we show the assigned reference number, level of formalization of the context definition (reference with underline formal, italic-gray informal, or bold-black no definition), the title along with its authors and year of publication, and a brief summary.

| PS  | Description                                                                 | Summary                                                                                                                                 |
|-----|-----------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------|
| [1] | Evaluating Citizens’ Readiness to Embrace e-Government Services. M. Shakaib Akram, and A. Malik. 2012 | It proposes a conceptual model, considering the relationships among quality characteristics of e-government websites, users’ perceptions about technology, satisfaction, trust and demographic characteristics in e-government context. |
| [2] | Context-aware data quality assessment for big data. D. Ardagna, C. Cappiello, W. Sama, and M. Vitali. 2018 | A methodology to build a DQ adapter module, which selects the best configuration for the DQ assessment based on the user main requirements: minimization of time and budget, and maximization of confidence. |
| [3] | Data Quality in a Big Data Context. F. Arolfo and A. Vaisman. 2018           | In order to capture the new characteristics that Big Data introduces, it dives into this problem, re-defining the DQ dimensions and metrics for a big data scenario, where data may arrive as unstructured documents in real time. |
| [4] | From Big Data to Smart Data: A Data Quality Perspective. M.T. Baldassarre, I. Caballero, D. Caivano, B. Rivas Garcia, and M. Piattini. 2018 | It introduces a methodology to make data smarter, taking as a reference point the quality level of the data itself. |
| [5] | Dependency Discovery in Data Quality. D. Barone, F. Stella, and C. Batini. 2018 | A conceptual framework for the automatic discovery of dependencies between DQ dimensions is described. Dependency discovery consists in recovering the dependency structure for a set of DQ dimensions measured on attributes of a database. |
| [6] | Information Quality in Use. C. Batini and M. Scannapieco. 2016.             | It analyzes how the information processor, be it a human being or an automated process, can manage the fitness for use of the information consumed. |
| [7] | Data Quality Dimensions. C. Batini and M. Scannapieco. 2016.                | It provides a concept of IQ and informally introduces several DQ dimensions, such as accuracy, completeness, currency, and consistency. |
| [8] | Methodologies for Information Quality Assessment and Improvement. C. Batini and M. Scannapieco. 2016 | It addresses DQ methodologies in terms of classifications, typical inputs and outputs, strategies addressed, and typical phases of methodologies. Also, it describes and compares in detail three of the most relevant general purpose methodologies. |
| [9] | Decision Support System for Implementing Data Quality Projects. M. Belthiah, M. Salim Bengtalla, and B. Bounabat. 2015 | It presents a model to identify the opportunities for increased monetary and non-monetary benefits from improved DQ within an enterprise architecture context. |
| [10]| Ontological Multidimensional Data Models and Contextual Data Quality. L. Bertossi and M. Milani. 2018 | It presents a formal model of context for context-based DQ assessment and quality data extraction. For that, they propose ontological contexts, and embedding multidimensional data models in them. |

This table continues on the next page.
| PS | Description | Summary |
|----|-------------|---------|
| [11] | Data Quality Is Context Dependent. L. Bertossi, F. Rizzo, and L. Jiang. 2010 | It focuses on DQ problems caused by this type of semantic discrepancy, considering the intuition and experience that DQ is context dependent. A model of context for DQ assessment is proposed. |
| [12] | Models of Data Quality. Z. Bicevska, J. Bicevskis, and I. Oditis. 2017 | It proposes an approach to DQM presenting three groups of domain-specific language: i) to use the concept of data object in order to describe data to be analysed, ii) to describe DQ requirements, and iii) to describe DQM process. |
| [13] | Domain-specific characteristics of data quality. J. Bicevskis, Z. Bicevska, and I. Oditis. 2017 | It discusses the issue of describing DQ and what should be taken into account when developing a universal DQM solution. It creates quality specifications for each kind of data objects and makes them executable. |
| [14] | The Information Resilience Framework: Vulnerabilities, Capabilities, and Requirements. K. Banahene Blay, S. Yeomans, P. Demian, and D. Murguia. 2020 | It develops and presents an information resilience framework by drawing on the theories of resilience, information quality, and vulnerability. |
| [15] | Visual Interactive Creation, Customization, and Analysis of Data Quality Metrics. C. Bors, T. Gschwandtner, S. Kriglstein, S. Miksch, and M. Pohl. 2018 | It presents an interactive environment for assessing DQ that provides customizable and reusable quality metrics, in combination with immediate visual feedback. |
| [16] | Governing Asset Management Data Infrastructures. P. Brous, P. Herder, and M. Janssen. 2016 | It suggests institutionalization of data governance within an organization as a unifying concept towards the effectiveness and sustainability of data infrastructures, recognizing their inherent complexities. |
| [17] | A Data Quality in Use Model for Big Data. I. Caballero, M.A. Serrano, and M. Piattini. 2014 | It addresses DQ concerns in big data projects. Also, it defines the quality characteristics that the different datasets being used for a specific use should present to fit for such use. |
| [18] | Quality Awareness for a Successful Big Data Exploitation. C. Cappiello, W. Samá, and M. Vitali. 2018 | It focuses on veracity that is related to the presence of uncertain or imprecise data: errors, missing or invalid data can compromise the usefulness of the collected values. |
| [19] | The ABC of Data: A Classifying Framework for Data Readiness. L.A. Casteljuns, Y. Maas, and J. Vanschoren. 2019 | It presents a framework focused on machine learning that encapsulates data cleaning and processing procedures. Datasets are classified within bands, which have associated DQ scores. |
| [20] | Context-Dependent Quality-Aware Source Selection for Live Queries on Linked Data. B. Catania, G. Guerrini, and B. Yaman. 2016 | It presents an approach taking into account context-dependent DQ, according to different dimensions, with the aim of selecting not only the most relevant but also the highest quality sources. |
| [21] | Exploiting Context and Quality for Linked Data Source Selection. B. Catania, G. Guerrini, and B. Yaman. 2019 | It presents how context and quality metadata can be exploited to improve source selection, using context information for interpreting the user needs. |
| [22] | Towards Open Data Quality Improvements Based on Root Cause Analysis of Quality Issues. C. Csáki. 2018 | It argues that while current open DQ frameworks are strong tools to assessing open DQ, or measuring maturity of data released, they are inadequate when it comes to helping public organizations (in how to release their data in better shape, or how to improve quality). |
| [23] | Quality Issues of Public Procurement Open Data. C. Csáki and E. Prier. 2018 | It addresses the gap between theory and practice for researchers of open data by illuminating potential issues and providing applicable solutions. |
| [24] | Big Data Systems: A Software Engineering Perspective. A. Davoudian and M. Liu. 2021 | It provides insight into activities of software engineering (requirements, designing and constructing software to meet the specified requirements, and software/data quality assurance), in the context of big data systems. |

This table continues on the next page.
| PS | Description                                                                 | Summary                                                                                                                                                                                                 |
|----|------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| [25] | Risk-Based Data Validation in Machine Learning-Based Software Systems. H. Foidl and M. Felderer. 2019 | It approaches the validation of data from a software engineering perspective. That is, it focuses on validating data in the situation where a trained machine learning model is implemented and deployed by software engineers. |
| [26] | Data Quality. C. Fürber. 2016                                                  | This chapter analyzes the main concepts and proposals in the DQ area.                                                                                                                                   |
| [27] | Cost and Value Management for Data Quality. M. Ge and M. Helfert. 2013          | It provides an overview of cost and value issues related to DQ. This includes DQ cost and value identification, classification, taxonomy, and evaluation framework, as well as analysis model. |
| [28] | Information quality assessment: Validating measurement. M. Ge, M. Helfert, and D. Jannach. 2011 | It proposes a framework to implement IQ assessment in practice, incorporating a set of valid measurement dimensions and a measurement process.                                                                |
| [29] | Can big data improve firm decision quality? The role of data quality and data diagnosticity. M. Ghasemaghaei and G. Calic. 2019 | This study addressed the impact of the large variety of data, the fast velocity of obtaining data, and the huge volume of generated data on decision quality and the mediating role of DQ and data diagnosticity on this relationship. |
| [30] | An Indicator Function for Insufficient Data Quality-A Contribution to Data Accuracy. Q. Görz and M. Kaiser. 2012 | It investigates how metrics for the DQ dimensions completeness, validity, and currency can be aggregated to derive an indicator for accuracy.                                                                     |
| [31] | A Survey on How to Manage Specific Data Quality Requirements during Information System Development. C. Guerra-Garcia, I. Caballero, and M. Piattini. 2010 | A systematic literature review is applied. It wants to discover how well the management of DQ requirements (at both the methodological and technological levels) is dealt with in specialized literature. |
| [32] | Methodology for linked enterprise data quality assessment through information visualizations. D. Gürdür, J. El-khoury, and M. Nyberg. 2019 | It presents the methodology, which aims to develop a DQ assessment tool—a dashboard—in addition to policies and protocols to DQM.                                                                    |
| [33] | Data Quality Alerting Model for Big Data Analytics. E. Gyalgyulyan, J. Aligon, F. Ravat, and H.V. Astatsryan. 2019 | It proposes a quality model, as the main component of an alert system, which allow to inform users about DQ issues, during their analysis.                                                                |
| [34] | Data quality assessment for improved decision-making: a methodology for small and medium-sized enterprises. L.C. Günther, E. Colangelo, H. Wiendahl, and C. Bauer. 2019 | It presents a methodology that simplifies the execution of DQ evaluations and improves the understandability of its results. One of its main concerns is to make DQ assessment usable to small and medium sized enterprises. |
| [35] | Requirements for Data Quality Metrics. B. Heinrich, D. Hristova, M. Klier, A. Schiller, and M. Szubartowicz. 2018 | Based on a decision-oriented framework, it presents a set of five requirements for DQ metrics. These requirements are relevant for a metric that aims to support an economically oriented management of DQ and decision making under uncertainty. |
| [36] | Data quality in internet of things: A state-of-the-art survey. A. Karkouch, H. Mousannif, H. Al Moatassime, and T. Noël. 2016 | Data properties and their new life cycle in IoT are surveyed. The concept of DQ is defined and a set of generic and domain-specific DQ dimensions, fit for use in assessing IoT’s DQ, are selected. |
| [37] | Assessing Information Quality by Six Sigma Method. S. Hyun Lee and A. Haider. 2012 | It introduces an approach to IQ measurement and employs six-sigma approach to IQ assessment. It focuses on continuous improvement of IQ by a systematic assessment of multiple IQ dimensions. |
| [38] | Data Quality Management and Measurement. X. Mao, B. Gong, F. Su, K. Xu, K. Xian, D. Liu, and H. Guo. 2019 | It proposes a DQM process framework and a DQ problem and measurement model.                                                                                                                               |
| [39] | Contribution of Artificial Neural Network in Predicting Completeness Through the Impact and Complexity of Its Improvement. J. Maqboul and B. Bounabat Jaouad. 2020 | It presents an approach based on qualitative (a survey) and quantitative (to learn from survey data) analysis, to help the decision-makers to target data by its impacts and complexities of process improvement. |

This table continues on the next page.
| PS | Description | Summary |
|---|---|---|
| [40] | Rule-Based Multidimensional Data Quality Assessment Using Contexts. A. Marotta and A. Vaisman. 2016 | It proposes the use of logic rules to assess the quality of measures in a Data Warehouse, accounting for the context in which these measures are considered. |
| [41] | A Data Quality in Use model for Big Data. J. Merino, I. Caballero, B. Rivus, M.A. Serrano, and M. Piattini. 2016 | A DQ-in-use model is presented for the assessment of the quality-in-use of the data from big data solutions. It is composed of contextual adequacy, temporal adequacy and operational adequacy. |
| [42] | Extending contexts with ontologies for multidimensional data quality assessment. M. Milani, L. Bertossi, and S. Ariyan. 2014 | In previous work a context model for the assessment of the quality of a database instance was proposed. Contexts are extended with dimensions (multidimensional contexts are represented as ontologies), making possible a multidimensional assessment of DQ assessment. |
| [43] | Adoption of human metabolic processes as Data Quality Based Models. A. Ngueilbaye, H. Wang, M. Khan, and D. Ahmat Mahamat. 2021 | A model that gets the quality in use levels of the entry data for big data analytics. The adequacies of DQ model based-in-use levels could be comprehended as dependability indicators and adequacy of big data. |
| [44] | Towards a data quality framework for decision support in a multidimensional context. D. Poeppeleman and C. Schultewolter. 2012 | It proposes a framework for DQ assessment based on the contextual factors management level and decision process phase, which allows the decision maker to relate DQ dimensions to values of features of the information demand. |
| [45] | Framework for defining information quality based on data attributes within the digital shadow using LDA. M. Riesener, C. Dölle, G. Schuh, and C. Tönnes. 2019 | It introduces a framework to determine IQ with respect to data-related and system-related attributes. Also, a literature review with focus on IQ and DQ is presented. |
| [46] | Data Quality Issues in Big Data: A Review. F. Ibrahim Salih, S. Adli Ismail, M.M. Hamed, O. Mohd Yusop, A. Azmi, and N. Firdaus Mohd Azmi. 2019 | Research on the quality of large data is reviewed and summarized by exploring the basic characteristics of large data. |
| [47] | Method for the Assessment of Semantic Accuracy Using Rules Identified by Conditional Functional Dependencies. Y.S. Santana and F.S. Lopes. 2019 | It presents a method for assess DQ using the conditional functional dependencies concept to extract quality rules and identify inconsistencies. The quality of a database is evaluated in the semantic accuracy dimension. |
| [48] | Handling Context in Data Quality Management. F. Serra. 2020 | It proposes to model context for DQM, exploiting the contextual nature of data, at each phase of the DQM process. |
| [49] | Data Warehouse Quality Assessment Using Contexts. F. Serra and A. Marotta. 2016 | It presents a study of existing proposals that relate DQ with Data Warehouse Systems and with contexts, and a proposal of a framework for assessing DQ in Data Warehouse Systems. |
| [50] | From Content to Context: The Evolution and Growth of Data Quality Research. G. Shankaranarayanan and R. Blake. 2017 2017 | It identifies core topics and themes of DQ research. According to the authors, in 2017 DQ research was at a critical point due to the growing popularity of social media data and rapidly increasing importance of data analytics in organization. |
| [51] | A secure and quality-aware prototypical architecture for the Internet of Things. S. Sicari, A. Rizzardi, D. Microrandi, C. Cappiello, and A. Ceven-Poristini. 2016 | The design and a prototypical implementation of a distributed middleware layer able to manage heterogeneous data sources, to provide a uniform, consistent data representation, and to evaluate the security and quality level associated to each data unit. |
| [52] | Big Data Quality: A Data Quality Profiling Model. I. Taleb, M. Adel Serhani, and R. Dssouli. 2019 | It proposes a DQ profiling model for big data that involves modules from sampling, profiling, exploratory quality profiling, DQ profile, and a quality profile repository (for managing DQ dimensions and their metrics, among other activities). |

This table continues on the next page.
| PS  | Description                                                                                                                                  | Summary                                                                 |
|-----|----------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------|
| [53] | The Data Quality Framework for the Estonian Public Sector and Its Evaluation. J. Tepandi, M. Laak, J. Linros, P. Raspel, G. Piho, I. Pappel, and D. Draheim. 2017 | It includes a DQ methodology, that comprises a DQ model, a DQ maturity model, and a DQM process, for the public sector organizations and activities for improving DQ. |
| [54] | A Methodology to Evaluate Important Dimensions of Information Quality in Systems. I. Todoran, L. Lecornu, A. Khenchaf, and J. Le Caillec. 2015 | A methodology makes use of existing frameworks and proposes a three-step process capable of tracking the quality changes through the system. |
| [55] | DMN for Data Quality Measurement and Assessment. Á. Valencia-Parra, L. Parody, Á. J. Varela-Yaca, I. Caballero, and M.T. Gómez López. 2019 | It proposes the use of decision model and notation, a declarative language, to facilitate the description of the business rules (for DQ measurements and assessment), as well as their evaluation. |
| [56] | Anatomy of Metadata for Data Curation. L. Visengeriyeva and Z. Abedjan. 2020                                                                   | It investigates the intrinsic connection between metadata and data errors. A taxonomy based on a closed grammar that covers all existing metadata and allows the composition of novel types of metadata is also presented. |
| [57] | An Ontology-Based Quality Framework for Data Integration. J. Wang, N.J. Martin, and A. Poulovassilis. 2011                                  | This framework captures users’ quality requirements in respect of integrated data resources. Also, it is formally represented in description logic. Users’ quality requirements are expressed as logic statements over a variety of quality factors. |
| [58] | Experience: Information Dimensions Affecting Employees’ Perceptions Towards Being Well Informed. M. Zárraga-Rodríguez and M.J. Álvarez. 2015 | It reports the results of a research survey among managers of companies committed to quality management within the framework of a total quality management model. |

Table 5. PS selected through the application of the SLR.