Wirtschaftsinformatik 2022 Proceedings

Track 10: Business Analytics, Data Science & Decision Support

Jan 17th, 12:00 AM

A Functional Taxonomy of Data Quality Tools: Insights from Science and Practice

Marcel Altendeitering
Fraunhofer ISST, Germany, marcel.altendeitering@isst.fraunhofer.de

Martin Tomczyk
TU Dortmund University, Germany, martin.tomczyk@tu-dortmund.de

Recommended Citation

Altendeitering, Marcel and Tomczyk, Martin, "A Functional Taxonomy of Data Quality Tools: Insights from Science and Practice" (2022). *Wirtschaftsinformatik 2022 Proceedings*. 4.
https://aisel.aisnet.org/wi2022/business_analytics/business_analytics/4

A Functional Taxonomy of Data Quality Tools: Insights from Science and Practice

Marcel Altendeitering[1] and Martin Tomczyk[2]

[1] Fraunhofer ISST, Dortmund, Germany
marcel.altendeitering@isst.fraunhofer.de
[2] TU Dortmund University, Chair for Industrial Information Management, Dortmund, Germany
martin.tomczyk@tu-dortmund.de

Abstract. For organizations data quality is a prerequisite for automated decision making and agility. To provide high quality data, numerous tools have emerged that support the different steps of data quality management. Yet, these tools vary in their functional composition and support for current trends, such as AI. There is no common and up-to-date perception of the capabilities a data quality tool should fulfill. In this paper, we develop a functional taxonomy of data quality tools to address this shortcoming and provide a holistic overview of data quality functionalities. We derived the taxonomy through an iterative approach of deductive reasoning by conducting a systematic literature review and inductive reasoning by reviewing existing data quality tools and gaining insights from experts. By applying our taxonomy to 18 commercial data quality tools we aim to provide the reader with a review of data quality tools and reach a functional consensus in the field.

Keywords: Data Quality, Data Quality Tools, Data Management, Taxonomy.

1 Introduction

Managing data quality (DQ) is an important aspect within data management and can help to build competitive advantages and increase a firm's value [1, 2]. At the same time, organizations are increasingly leveraging data to enable automated decision making in the form of machine learning (ML) and artificial intelligence (AI) [3]. To avoid misaligned decision making and secure organizational agility it is necessary to uphold a certain level of DQ [4]. In this sense, DQ is defined as fulfilling the 'fitness for use' desired by data consumers to enable efficient business operations and concise decisions [5, 6]. To achieve this goal in light of distributed data landscapes and quickly changing business requirements, a market for DQ tools emerged. This market is closely integrated with related markets for data integration, metadata management, or master data management solutions [6]. DQ tools, hereby, focus on supporting the different steps of the DQ lifecycle by defining, measuring, analyzing, and improving the quality of data sets [7]. Traditionally, DQ tools had a focus on fulfilling internal compliance guidelines and reducing risks by manually defining DQ rules that new data needs to adhere. Nowadays, these tools are becoming more intelligent and automation and AI are an integral part of DQ tools [8].

17th International Conference on Wirtschaftsinformatik
February 2022, Nürnberg, Germany

However, despite DQ being a $1.77 billion market [9] and many available DQ tools, companies still struggle to leverage high quality data. This causes many organizations to build their own solutions or extend existing DQ tools with customized functionalities (e.g., [3,8,10]). Hereby, it is often difficult to determine what a DQ tool should be capable of and how it differentiates from related solutions in the field of data management. In this paper, we want to reach a unified perception of the functionalities a DQ tool should offer based on insights from science and practice. Managers and practitioners would benefit from a such an overview to inform make-or-buy decisions and build customized solutions. Researchers could build on a structured analysis of DQ tools from both science and practice to formulate potential research gaps and advance the future development of the field. Therefore, we formulated the following research question:

Research Question: What are the functional characteristics of DQ tools described in science and practice?

To answer the proposed research question, we adopted the structured taxonomy development process by Nickerson et al. [11]. A taxonomy offers a suitable way to formulate conceptual knowledge as it helps to "structure or organize the body of knowledge that constitutes a field" [12, p. 65]. It allows to investigate relationships among concepts and classify existing objects. We followed the approach of Nickerson as the combination of inductive and deductive reasoning allowed us to combine scientific and practical knowledge into one artifact [11].

The remainder of this article is structured as follows. First, we outline the theoretical background and related work for our study in section 2. In section 3, we describe the research methodology and taxonomy development process we followed. Our final taxonomy is presented in section 4. In section 5, we offer a discussion on the results of applying our taxonomy to DQ tools. Finally, in section 6, we conclude with implications of our work, its limitations, and paths for future research.

2 Background & Related Work

In the scientific literature high quality data sets are widely considered as an antecedent for enabling organizational agility and securing competitive advantages [1,2,4]. They can, for instance, positively influence the creation of data-driven services [13,14] (e.g., by improving the detection rates of AI services), support the organizational decision making and business intelligence [15] (e.g., by providing more accurate reports), or promote the exchange of data in data platforms [16] (e.g., by raising the trust in data among participants). Following this, DQ is the foundation for every important activity within an organization and a key success factor [17].

To achieve a high level of DQ, it has long been incorporated as an essential building block in data management [1] and several frameworks for managing DQ, such as the Total Data Quality Management (TDQM) framework [7], emerged. Since the day-to-day DQ work can be cumbersome and time-consuming, they are often supported by DQ tools [8].

2.1 Data Quality Tools

There are numerous DQ tools available in science and practice that support the DQ tasks [18]. However, these tools often focus on specific DQ tasks, like duplicate detection, data cleaning, or data validation. Some tools aggregate several DQ functionalities to DQ solutions for a specific area of application. We can distinguish these solutions in data preparation tools [19], data measuring and monitoring tools [18], and general-purpose tools. Besides existing solutions DQ tools are often created as customized solutions [8]. This way, a frictionless integration with existing data management tools and the remaining IT landscape can be achieved.

In this study, we focus on general-purpose DQ tools as they offer the most comprehensive set of functionalities and most commercial DQ tools fall into this category. They can usually be applied in different application domains and are not limited to certain tasks. This way, they can be separated from the two former types, which put an emphasis on the application in data sciene (data preparation tools) or master data sets (measuring and monitoring tools). Against the background of AI, general-purpose DQ tools increasingly use intelligent functionalities to improve DQ, which in turn is needed for the efficient training of AI models.

It can be difficult to distinguish the functionalities of DQ tools from related offerings, such as data catalogs, master data management suites, or data warehouses. For example, to provide data that is ’fit for use’ [5] DQ tools can rely on the metadata that is stored in a data catalog system. In fact, a close integration of DQ tools with related solutions is important to provide trusted data and reduce unnecessary DQ work [6]. Thus, many vendors have related solutions in their product portfolio.

2.2 Related Work

Despite the high importance of DQ for organizations and its rising relevance in view of ever growing amounts of data, DQ remains an issue for organizations [8, 17]. Although a large body of literature is available on the topic, the existing studies are often focused on either technical aspects of DQ [20] or on the management aspect of DQ [1]. There is limited interdisciplinary research available that provides the holistic knowledge required by organizations [8]. With our study we aim to bridge this gap by providing a holistic, interdisciplinary taxonomy of functionalities in DQ tools that is based on findings from science and practice.

To the best of our knowledge such a functional taxonomy does not yet exist. Related taxonomies in the DQ field are either problem-oriented (e.g., [21,22]) or focus on specific DQ tasks and data types (e.g., [23, 24]). The two studies by Ehrlinger et al. [18] and Hameed & Naumann [19] are coming closest to the idee of our paper. In the former study, the authors conducted a comprehensive analysis of DQ tools with an emphasis on monitoring and measurement functionalities. They found that DQ tools are not leveraging ML technologies to its fullest extent and call for further research on the customization and explanation of ML algorithms in DQ tools. The study by Hameed & Naumann presents a structured analysis of the capabilities of data preparation tools and an overview of existing solutions. They highlight the need for further research on automation and interoperability of data preparation and related tools.

3 Research Methodology

3.1 General Approach

To develop the desired taxonomy, we used the methodology introduced by Nickerson et al. [11] as it is well established within the Information Systems (IS) community (e.g. [25–27]) and offers consistency with the guidelines of design science research [28]. The approach includes a total of seven steps. It begins by defining the meta characteristics of the future taxonomy (step one) and by defining the ending conditions of the taxonomy development process (step two). In the steps three to six the user of the method can either follow a conceptual-to-empirical approach, which implies a deductive procedure to derive characteristics and dimensions from theory, or the user choses an empirical-to-conceptual approach, which uses empirical sources, such as DQ solutions at the market, to derive the results inductively. In the seventh step a decision is made whether the ending conditions defined by Nickerson et al. [11] have been met. If not, another iteration using one of the two approaches is invoked. If they have been met and no further changes were made, the method terminates.

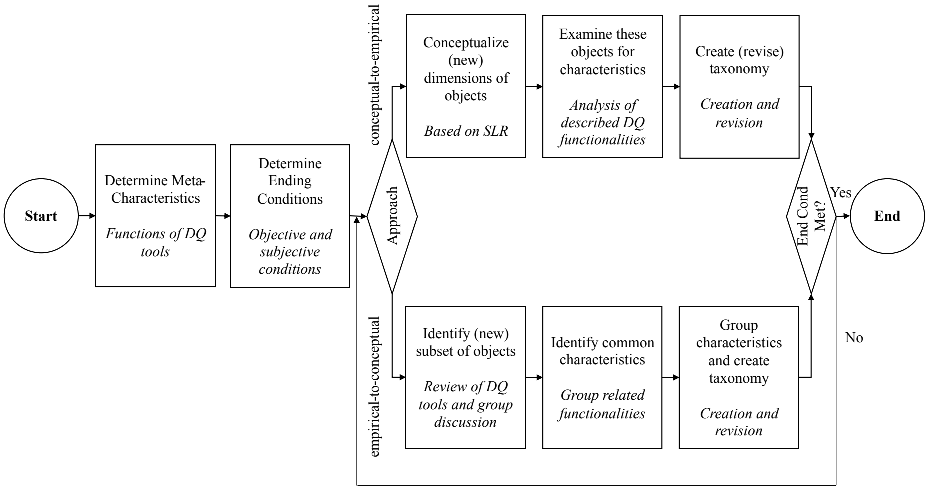
Start
Determine Meta-Characteristics
Functions of DQ tools
Determine Ending Conditions
Objective and subjective conditions
Approach
conceptual-to-empirical
Conceptualize (new) dimensions of objects
Based on SLR
Examine these objects for characteristics
Analysis of described DQ functionalities
Create (revise) taxonomy
Creation and revision
empirical-to-conceptual
Identify (new) subset of objects
Review of DQ tools and group discussion
Identify common characteristics
Group related functionalities
Group characteristics and create taxonomy
Creation and revision
End Cond Met?
Yes
No
End

Figure 1. Taxonomy development process as introduced by Nickerson et al. [11]. Our activities are shown in *italic*.

3.2 Taxonomy Development Process

Meta Characteristics: Following Nickerson et al. [11], the meta-dimensions are guided by the purpose and future utility of the taxonomy. In accordance with the proposed research question, the purpose of our taxonomy is to provide a holistic overview of the functional capabilities of DQ tools. We, therefore, focused our study on capabilities that fulfil functional requirements with regard to DQ. We neglected analyzing aspects that are non-functional or not directly associated with DQ, such as deployment and cost models, security, or usability. While a review of such functionalities would certainly be interesting, it would have been difficult to do in a structured manner as the importance and concrete definition of these requirements varies by users and they often lack documentation and concrete specifications.

First Iteration (conceptual-to-empirical): In the first iteration we used the existing body of literature to obtain a comprehensive theoretical overview of DQ functionalities. We, hereby, followed the literature research guidelines according to vom Brocke et al. [29] and Kuhrmann et al. [30]. To identify relevant scientific literature, we decided to systematically search the Scopus, AISel, IEEE Xplore, ACM DL, and EBSCO databases as these include significant articles from the IS and computer science domains (see Figure 2) [30]. We searched these databases for both *data quality* and *information quality* since they are often used synonymously [31]. We, furthermore, added *tool* and related functional artifacts to the search term to identify implementations of DQ capabilities. Consequently, we applied the following search term: *"data quality" OR "information quality"* ***AND*** *(tool* OR method OR framework OR solution)*. Because of the expected large number of results we decided to limit the search on the article title only.

We identified a total of 677 papers in our initial search. In the next step, we retained all conference papers and journal articles and excluded duplicates and articles not written in English from the result set. This resulted in 221 contributions for further consideration. In the following step, we manually reviewed these papers to identify papers that are relevant to our study. The exclusion of a paper in this step was either due to papers not addressing DQ itself (e.g., methods robust to DQ issues) or a lack of details in the paper (e.g., short papers on DQ). Finally, 46 contributions remained. Based on this set we conducted a forward and backward search as specified by Webster and Watson [32]. This led to a final of 55 papers as the theoretical foundation of our taxonomy.

To derive the relevant characteristics of DQ tools, we systematically screened these 55 papers for DQ functionalities and assigned the identified characteristics to the relevant dimensions using a concept matrix [32]. As a starting point for the decomposition of DQ functionalities in our concept matrix we used the four steps of the TDQM lifecycle [7] as it is a widely accepted framework for DQ management and was often used for classifying solutions we found in literature. However, we observed that the TDQM model could not cover all DQ functionalities described in literature. There were several studies that highlighted the integrability of DQ tools (e.g., [6, 8, 18]). To categorize these characteristics we included a fifth dimension that contains corresponding functionalities.

Second Iteration (empirical-to-conceptual): In the second iteration we aimed to shed light on the empirical perspective on DQ and systematically reviewed DQ tools available on the market. Since there is a large number of DQ tools available (in [18] the authors identified 667 distinct tools) we decided to focus our analysis on commercial DQ tools. We, thus, neglected solutions that provided no enterprise support (e.g., self-hosted open source tools such as MobyDQ [33]), that are not available publicly (e.g., Data Sentinel [10]), or that specialize in certain functionalities (e.g., Tamr [34] for data unification). We chose this discriminator because commercial DQ tools offer a holistic set of functionalities for different DQ tasks and application domains, and helped us to

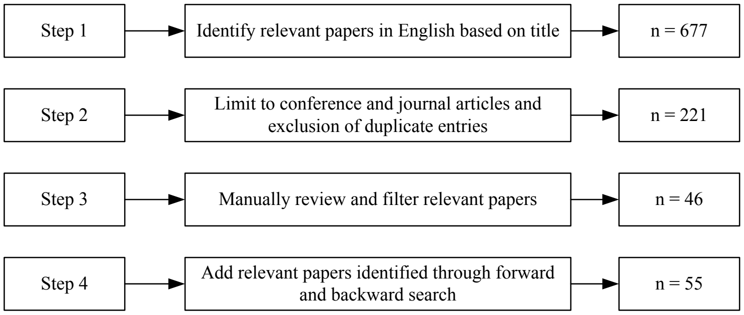
Step 1
Identify relevant papers in English based on title
n = 677
Step 2
Limit to conference and journal articles and exclusion of duplicate entries
n = 221
Step 3
Manually review and filter relevant papers
n = 46
Step 4
Add relevant papers identified through forward and backward search
n = 55

Figure 2. The systematic literature review process we followed.

be in accordance with our research goal. We initialized our search for candidate tools with related surveys [18, 19]. These studies provided a good starting point, but they focus on specific DQ functionalities and required a manual review of the presented tools. Additionally, we included the tools from the 2020 Hype Cycle for DQ solutions by Gartner [6]. Finally, we conducted a Google search to identify further tools that were not identified in the first two steps. We, hereby, limited our search to the first 100 hits, which we examined more closely. This procedure yielded in 18 DQ tools that we included in the second iteration of the taxonomy development process.

We manually reviewed the selected DQ tools with regard to their applied DQ functionalities using the characteristics from the first iteration as a basis. Therefore, wherever possible, we used trial or test versions of the DQ tools to experiment with the offered DQ functionalities. We, furthermore, screened the available documentation for specifications of DQ functionalities. In some cases, the DQ functions were part of a larger solution package. For example, the Ataccama One platform [35] includes several modules for DQ, data cataloging, data integration, etc. We concentrated our analysis on the DQ module in these cases. The review of the tools revealed that our taxonomy was already well developed, comprehensive, and representative. However, we added two characteristics (Job Scheduling and Web Services) and merged or renamed several others. For example, we merged a characteristic *DataOps*, which we derived in the first iteration, with the characteristic *Collaboration* because DQ tools often summarize DataOps functionalities under the larger aspect of collaborative DQ [6].

Third Iteration (empirical-to-conceptual): In the third iteration we conducted a group discussion with five experts (two from science and three from practice) using the taxonomy from the second iteration as a basis. The experts were working in either data science, data engineering, or data management roles and had two to ten years of experience in working with DQ. With the selection of participants we aimed to obtain a broad view on different aspects of DQ. In the group discussion one of the authors presented the characteristics for each dimension and put them up for discussion before continuing with the next dimension. The meeting lasted four hours and was fruitful as the participants could draw on their personal experiences with DQ. The feedback we received helped us to concretize the characteristics for each dimension and supported our findings so far. Specifically, we changed the wording of some characteristics to help the taxonomy become more concise and self-explanatory. Since there were only minor changes in this iteration and we were able to meet all 13 ending conditions specified by Nickerson et al. [11] (see Table 1), we decided to finalize the taxonomy.

Table 1. Ending conditions for each iteration (adapted from Nickerson et al. [11])

	Ending Conditions	*#1*	*#2*	*#3*
Objective	All objects or a representative sample of objects have been examined	-	x	x
	No object was merged with a similar object or split into multiple objects in the last iteration	-	x	x
	At least one object is classified for every characteristics of every dimension	-	x	x
	No new dimensions or characteristics were added in the last iteration	-	-	x
	No dimensions or characteristics were merged or split in the last iteration	-	-	x
	Every dimension is unique and not repeated	x	x	x
	Every characteristic is unique within its dimension	x	x	x
	Each cell (combination of characteristics) is unique and is not repeated	x	x	x
Subjective	Concise	-	-	x
	Robust	-	-	x
	Comprehensive	-	x	x
	Extendible	-	x	x
	Explanatory	-	-	x

4 A Functional DQ Taxonomy

The final taxonomy consists of five dimensions (**D**$_n$) and 25 characteristics (**C**$_{nm}$) (see Table 2). The first four dimensions describe functional capabilities for managing DQ. We divided these dimensions following the TDQM lifecycle, which specifies *Definition*, *Measurement*, *Improvement*, and *Analysis* as the four basic steps in DQ management [7, 36]. The fifth dimension inherits capabilities regarding the integrability with associated data management tools, such as master data management suites.

The dimension **Definition (D$_1$)** includes functionalities that are required for defining what DQ means in a specific context. This is important as DQ requires a 'fitness for use', which varies by user and domain [87]. *Profiling (C_{11})* capabilities offer functionalities to derive relevant metadata (e.g., keywords, common values, etc.) from a data set and use these for quality analysis and building DQ rules. The characteristic *Business Rules (C_{12})* allows users to define custom DQ rules and policies. This way, it facilitates the manifestation of domain specific data knowledge, which can be used for validation purposes. Similarly, *Standard Rules (C_{13})* represent generally applicable DQ rules. These rules are usually pre-defined and used for validating common data types like phone numbers, email addresses, or zip codes [38, 45]. With *Governance (C_{14})* capabilities users can assign managerial and governance aspects, such as data ownership or roles and responsibilities to data items. For example, defining the ownership of a data set is a pre-emptive measure to keep data clean at the source and avoid DQ issues [40].

The dimension **Measurement (D$_2$)** describes functionalities for determining DQ metrics along multiple DQ dimensions [7]. The characteristic *Metrics (C_{21})* includes

Table 2. Final version of our taxonomy and corresponding literature. *(C$_{33}$) and (C$_{56}$) have no corresponding literature as they were identified during the review of DQ tools only.*

Dimension (D$_n$)	*Characteristic (C$_{nm}$)*	*related DQ Literature*
Definition (D$_1$)	Profiling (C$_{11}$)	[8, 18, 19, 37–44]
	Business Rules (C$_{12}$)	[3, 8, 39, 42, 45–59]
	Standard Rules (C$_{13}$)	[38, 45, 48, 49, 59, 60]
	Governance (C$_{14}$)	[37, 40, 47, 61–64]
Measurement (D$_2$)	Metrics (C$_{21}$)	[8, 18, 44, 47, 64–71]
	Real-Time (C$_{22}$)	[3, 8, 18, 44, 45, 48, 50–52, 72, 73]
	Anomalies (C$_{23}$)	[8, 18, 42, 57, 68, 74]
	Validation (C$_{24}$)	[3, 18, 19, 39, 44–47, 50, 52–54, 56–60, 66, 75]
Analysis (D$_3$)	Lineage (C$_{31}$)	[74–78]
	Issue Tracking (C$_{32}$)	[38, 44, 62, 68, 79, 80]
	Job Scheduling (C$_{33}$)	-
	Remediation (C$_{34}$)	[3, 8, 49, 51, 61, 71, 75, 79, 80]
	Cost Saving (C$_{35}$)	[47, 76]
	Collaboration (C$_{36}$)	[37, 47, 57, 61–64, 75, 79]
Improvement (D$_4$)	Auto Correction (C$_{41}$)	[3, 19, 42, 49, 57, 67, 81–84]
	Standardization (C$_{42}$)	[38, 44, 49, 60, 68, 85]
	Deduplication (C$_{43}$)	[18, 19, 42, 44, 49, 60]
	Enrichment (C$_{44}$)	[19, 38, 60, 82, 85, 86]
	Prevention (C$_{45}$)	[72, 81]
Integration (D$_5$)	Master Data (C$_{51}$)	[8, 18, 40, 47, 55]
	Record Linkage (C$_{52}$)	[49, 60, 65, 78]
	Orchestration (C$_{53}$)	[8, 44, 46, 49, 60, 65]
	Import / Export (C$_{54}$)	[8, 18, 44]
	Non-relational (C$_{55}$)	[66, 75]
	Web Services (C$_{56}$)	-

simple data analysis functionalities (e.g., number of null values) that provide visibility into the quality of data products. These are often accompanied with graphical dashboards and scorecards showing DQ trends over time [88, 89]. Sometimes this functionality is extended with a *Real-Time (C$_{21}$)* analysis. The real-time capability detects changes to data and invokes a DQ measurement. With *Anomalies (C$_{23}$)* DQ tools assist users in finding inconsistencies. Hereby, cardinalities (e.g., min/max values), outlier detection (e.g., isolation forest), and distance based (e.g., levenshtein) algorithms are used for labeling data points as outliers [8, 68]. In *Validation (C$_{24}$)*,existing DQ rules are applied to new data to find violations and, therefore, acts as a 'DQ firewall' to the database [88].

Based on the measurements the **Analysis (D$_3$)** capabilities assist to investigate "the root cause" [7, p. 64] of DQ problems. With *Lineage (C$_{31}$)* functionalities DQ tools support users in following the merits of changes to a data set and finding the source of DQ issues [89]. They, thus, prevent the continued downstream propagation of data errors [76]. Following, Chien & Jain [6] *Issue Tracking (C$_{32}$)* provides the possibility to follow up on DQ problems. Specifically, to "identify, quarantine, assign, escalate, resolve and monitor DQ issues" [6, p. 3]. *Job Scheduling (C$_{33}$)* capabilities realize an automated and scheduled execution of DQ analysis jobs. With *Remediation (C_{34})* DQ tools allow to specify workflows and route DQ problems to domain experts or data stewards for resolving the issue. Additionally, the remediation helps to train classifiers for DQ problems based on active learning [3]. *Cost Saving (C_{35})* functionalities provide the possibility to calculate return on investments for specific DQ actions. According to SAP [89] they can help to secure funding and support for DQ initiatives. *Collaboration (C_{36})* functions are related to *Remediation (C_{34})* but have a broader focus on improving the communication and overall workflow between the stakeholders of a data set. It can be seen as a part of the DataOps concept, which influences DQ indirectly [6].

The last stage of the TDQM lifecycle is **Improvement ($\mathbf{D}_4$)**, which summarizes capabilities for fixing issues and increasing DQ. *Auto Correction (C_{41})* describes functionalities for simple automated data cleaning. Typically, these include the automated filling of missing values or deletion of whitespaces [19]. The *Standardization (C_{42})* functionality enables the automated addition of data values based on standardization rules (see C_{13}) and reference data sets [38]. Common examples are data extensions based on zip codes or addresses. Data *Deduplication (C_{43})* refers to the removal of exact and near duplicate entries from a data set. Depending on the data type different algorithms can be used for this task. For example, Ataccama [35] utilizes deterministic matching, fuzzy matching, edit distance, and phonetic algorithms. Following, Chien & Jain, data *Enrichment (C_{44})* specifies "capabilities that integrate externally sourced data to improve completeness and add value" [6, p. 3]. *Prevention (C_{45})* specifies AI based functionalities for avoiding DQ issues. For example, Infogix [90] offers methods for predicting future DQ problems based on historical data characteristics and issue reconciliations.

The dimension **Integration ($\mathbf{D}_5$)** includes capabilities with regard to the interoperability and integrability of DQ solutions and the data they contain. The characteristic *Master Data (C_{51})* creates an integration with master data management suites to propagate high quality data to a central hub [91]. *Record Linkage (C_{52})* uses entity resolution techniques to establish links between data items that can be associated with the same real-life entities. It helps to consolidate data sets and prevent duplicate entries [88]. With data *Orchestration (C_{53})* functionalities a DQ tool enables an integration with data pipelines that are managed in other data management tools (e.g., Extract-Transform-Load (ETL) pipelines). This is important to avoid an isolation of DQ measures and spread high quality data [8]. Furthermore, DQ tools should offer *Import / Export (C_{54})* options to enable the simple sharing of file based results [8]. Handling relational data is a standard functionality for DQ tools. Lately, including *Non-relational (C_{55})* data in DQ management is becoming increasingly popular. Examples include the analysis of data streams [92] or text documents, such as PDF [93]. Lastly, *Web Services (C_{56})* enable the creation of DQ as a service. Hereby, DQ processes and workflows are offered as realtime web services that can be called on demand [88].

5 Taxonomy Application & Discussion

To demonstrate the applicability and conduct an evaluation of our taxonomy, we applied the final version to the 18 DQ tools that we identified in the second iteration of the taxonomy development process (see Table 3). The classification of empirical objects

Table 3. Application of the final taxonomy to DQ solutions at the market (ordered alphabetically). **Infogix was acquired by Precisley and will be included in their portfolio. Since the acquisition was not complete at the time of writing we still included Infogix in this overview.*

Dimension (D_n)	*Characteristic (C_{nm})*	*Ataccama* [35]	*Experian* [96]	*Human Inference* [97]	*IBM* [98]	*Infogix* [90]*	*Informatica* [99]	*InfoZoom* [100]	*Innovative Systems* [101]	*Melissa Data* [102]	*MIOSoft* [92]	*nModal Solutions* [103]	*Oracle* [88]	*Precisely* [93]	*Redpoint* [104]	*SAP* [89]	*SAS* [91]	*Syniti* [105]	*Talend* [106]
Definition (D_1)	Profiling (C_{11})	x	x	x	x	x	x	x	x		x	x	x	x	x	x	x		x
	Business Rules (C_{12})	x	x	x	x	x	x	x			x	x	x	x		x	x	x	x
	Standard Rules (C_{13})	x		x	x		x			x	x		x			x	x	x	x
	Governance (C_{14})	x		x	x	x	x								x	x		x	
Measurement (D_2)	Metrics (C_{21})	x	x	x	x	x	x	x	x		x	x	x		x	x	x		x
	Real-Time (C_{22})	x	x	x	x	x	x	x	x		x	x		x	x		x	x	
	Anomalies (C_{23})	x	x	x	x				x					x	x		x		x
	Validation (C_{24})	x	x	x	x	x	x	x	x	x	x	x	x		x	x	x	x	x
Analysis (D_3)	Lineage (C_{31})															x	x		
	Issue Tracking (C_{32})	x									x		x	x			x	x	x
	Job Scheduling (C_{33})		x	x			x	x			x	x	x				x		
	Remediation (C_{34})				x	x	x	x			x		x			x	x	x	x
	Cost Saving (C_{35})					x										x			
	Collaboration (C_{36})	x	x	x	x	x	x							x		x	x	x	x
Improvement (D_4)	Auto Correction (C_{41})	x	x	x	x		x	x	x		x			x	x	x	x	x	x
	Standardization (C_{42})	x	x	x	x	x	x		x	x			x	x	x		x		x
	Deduplication (C_{43})	x	x	x	x		x	x	x	x	x	x	x	x	x		x		x
	Enrichment (C_{44})	x	x	x			x		x	x	x			x	x				x
	Prevention (C_{45})	x			x	x	x							x					
Integration (D_5)	Master Data (C_{51})	x	x		x			x	x				x	x	x	x	x	x	x
	Record Linkage (C_{52})	x	x	x				x	x		x		x	x	x		x		x
	Orchestration (C_{53})	x	x	x	x	x		x			x	x	x	x	x	x	x	x	x
	Import / Export (C_{54})	x	x	x			x	x			x	x	x	x	x	x	x	x	x
	Non-relational (C_{55})										x			x	x				
	Web Services (C_{56})	x		x					x				x	x					

that differ in their domain, functional focus, and maturity supports the usefulness and validity of our taxonomy [11]. Additionally, the application of the taxonomy provides a state-of-the-art overview of commercial DQ tools, which can help to reveal potential market gaps and highlight paths for future developments. It is obvious that our taxonomy does not offer "mutually exclusive and collectively exhaustive" [11, p. 5] characteristics. However, similar studies (e.g., [94, 95]) show that only some or no characteristics can provide mutually exclusiveness. DQ tools rarely feature mutually exclusive functions, which is represented in our developed taxonomy.

On a general level we observed that most DQ tools focus on the *Definition (D_1)*, *Measurement (D_2)*, and *Improvement (D_4)* of DQ. The *Analysis (D_3)* of root causes and collaborative handling of DQ issues is developed to a lesser extent. The same applies to the *Integration (D_5)* of DQ tools with other solutions across the organizational IT landscape. This finding is tallied with the current trend for interdisciplinary DQ, which sees DQ as "much less of a specialized IT task" [6, p. 2]. It has rather become an integral effort that involves different specialists from across the organization who create high quality data products [6, 8]. We, therefore, argue that existing integrative capabilities such as *Remediation (C_{34})*, *Collaboration (C_{36})*, or *Orchestration (C_{53})* that support the collaborative handling of DQ issues will become increasingly important. We can imagine that new integrative solutions and capabilities will emerge across all steps of the DQ lifecycle (like CoClean has proposed for data cleaning [107]). For example, instead of centralizing the definition and measurement of DQ using rule based approaches, data users (e.g., data analysts) could specify the quality they need for their respective use case. This way use cases of varying DQ can be realized and the users receive data that is just right for their desired task. Since DQ is a context-dependent concept [5] and the amount of data produced is continuously increasing a decentralization of DQ efforts is required. In this sense, future DQ solutions should incorporate capabilities for solving DQ issues at the source and a collaboration with other fields if necessary [17]. Such functionalities could include the simple remediation of DQ problems, incentivize and nudging for high quality data, or gamification approaches.

Additionally, we can confirm a clear trend towards more intelligent and automated DQ processes. Similar to DataOps for ML, DQ tools increasingly automate different parts of the DQ lifecycle. Traditionally, DQ tools realized high DQ by offering functionalities for the manual specification of *Business Rules (C_{12})* and *Standard Rules (C_{13})* and validating data against these rules (*Validation (C_{24})*. Lately, these capabilities are increasingly supported by AI and ML. For example, through the automated generation of DQ rules using unsupervised ML algorithms (e.g., [8, 46]). The same applies to the intelligent automation of resolving bad data (i.e., *Auto Correction (C_{41})*), *Deduplication (C_{43})*, *Record Linkage (C_{52})*, and *Enrichment (C_{43})* of data sets. However, we also found that no DQ tool implements state-of-the-art and innovative ML technologies like contextual ML. DQ tools from the scientific community (e.g., Holoclean [20]) are more advanced in this regard but often focus on a subset of DQ tasks. Future DQ tools should follow these examples and incorporate DQ knowledge from multiple sources to realize a holistic, intelligent DQ solution.

Some functionalities we identified are beyond the standard offerings of DQ tools. These are implemented by less than a third of the tools and represent novel and innovative functionalities that could become a standard offering in the future. One of these functionalities is the handling of *Non-relational (C_{55})* data, which is a current trend in the DQ domain [108]. Hereby, new DQ metrics and definitions are needed that match the special characteristics of non-relational data sets. As mentioned above, offering *Web Services (C_{56})* for integrability and *Prevention (C_{45})* are becoming increasingly important for DQ tools. However, it was often unclear what algorithms are applied and how parameters can be tuned to achieve better results. To overcome this issue DQ tools should invest in a more detailed explanation of AI functionalities [8, 18].

6 Conclusion

In our study we developed a taxonomy of DQ tools through an iterative process of deductive and inductive reasoning as specified by Nickerson et al. [11]. We, therefore, used a combination of a systematic literature review, an analysis of commercial DQ tools, and the input of DQ experts. With the design and application of our taxonomy we can provide a holistic and state-of-the-art overview of DQ functionalities and answer the initially proposed research question.

From our results we can derive several **managerial implications**. As DQ is a contextual and user-centric concept, customized DQ solutions are important for organizations. Our taxonomy can assist in building such customized DQ tools by providing a holistic and comprehensive understanding of the functionalities a DQ tool should offer. This way, it helps to extend existing tools with novel functionalities and inform make-or-buy decisions. Towards this end, the application of our taxonomy provides a comparison of commercial DQ tools and the functionalities they offer. Our findings, furthermore, emphasize the interdisciplinary and collaborative nature of DQ and can help organizations to set up new DQ initiatives that incorporate these aspects. Thus, our taxonomy can support the decision-making for better DQ from single data analytics projects to corporate data strategies.

Besides managerial implications our work offers **scientific implications**. Through a rigorous combination of deductive and inductive research approaches we created a resolute and sound DQ taxonomy. This way, we contribute to the existing body of literature on DQ and offer researchers a conceptual structure for analyzing and classifying DQ tools. It can, thus, facilitate further research on the composition and role of DQ tools with regard to current trends, such as automation and collaboration. In particular, our study advances the area of *Business Analytics, Data Science and Decision Support* by highlighting that DQ tools should become more collaborative, understandable, and automated. Improving DQ tools in these areas can help to ensure the provisioning of high quality data, which is vital for data science.

Despite applying a high level of rigor our research is subject to several **limitations**. Most importantly, we cannot rule out subjectivity in parts of our research. Specifically, this is the case for the manual review of relevant literature and the group discussions we conducted with DQ experts. Other researchers might derive different characteristics and reach other conclusions. With our research method we tried to overcome this problem and avoid personal bias. Moreover, our taxonomy represents a current state of DQ functionalities in research and practice and might be of limited stability for the future. Especially functionalities that describe current trends (e.g., *Non-relational (C_{55})*) might be further generalized in the future. The taxonomy should, therefore, be critically updated and scrutinized to ensure that the characteristics remain relevant and up-to-date.

Besides future improvements of the taxonomy, **future work** can include a more extensive review of DQ tools (e.g., including open source tools). This could provide a more detailed analysis of the differences between theoretical and commercial implementations of DQ functionalities and serve as a basis for the development of archetypical patterns [12]. Moreover, it would be interesting to further investigate how DQ characteristics change over time from traditional to more intelligent methods.

References

1. Otto, B.: Quality and value of the data resource in large enterprises. Information Systems Management 32(3), 234–251 (2015)
2. Côrte-Real, N., Ruivo, P., Oliveira, T.: Leveraging internet of things and big data analytics initiatives in european and american firms: Is data quality a way to extract business value? Information & Management 57(1), 103–141 (2020)
3. Shrivastava, S., Patel, D., Zhou, N., Iyengar, A., Bhamidipaty, A.: Dqlearn : A toolkit for structured data quality learning. In: 2020 IEEE International Conference on Big Data (Big Data). pp. 1644–1653. IEEE (2020)
4. Gür, I., Guggenberger, T.M., Altendeitering, M.: Towards a data management capability model. AMCIS 2021 Proceedings (2021)
5. Wang, R.Y., Strong, D.M.: Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems 12(4), 5–33 (1996)
6. Chien, M., Jain, A.: Gartner magic quadrant for data quality solutions (2020), `https://www.gartner.com/en/documents/3988016/magic-quadrant-for-data-quality-solutions`, (Accessed: 19.05.2021)
7. Wang, R.Y.: A product perspective on total data quality management. Communications of the ACM 41(2), 58–65 (1998)
8. Altendeitering, M., Guggenberger, T.M.: Designing data quality tools: Findings from an action design research project at boehringer ingelheim. ECIS 2021 Research Papers (2021)
9. Dayley, A., Dsilva, V., Jain, A., Chien, M., Menon, S.: Market share: Data quality tools, worldwide, 2019 (2020), `https://www.gartner.com/en/documents/3984707/market-share-data-quality-tools-worldwide-2019`, (Accessed: 01.07.2021)
10. Swami, A., Vasudevan, S., Huyn, J.: Data sentinel: A declarative production-scale data validation platform. In: IEEE 36th International Conference on Data Engineering (ICDE). pp. 1579–1590. IEEE (2020)
11. Nickerson, R.C., Varshney, U., Muntermann, J.: A method for taxonomy development and its application in information systems. European Journal of Information Systems 22(3), 336–359 (2013)
12. Glass, R.L., Vessey, I.: Contemporary application-domain taxonomies. IEEE Software 12(4), 63–76 (1995)
13. Amadori, A., Altendeitering, M., Otto, B.: Challenges of data management in industry 4.0: A single case study of the material retrieval process. In: Abramowicz, W., Klein, G. (eds.) Business Information Systems. pp. 379–390. Springer International Publishing, Cham (2020)
14. Tebernum., D., Altendeitering., M., Howar., F.: Derm: A reference model for data engineering. In: Proceedings of the 10th International Conference on Data Science, Technology and Applications - DATA,. pp. 165–175. INSTICC, SciTePress (2021)
15. Davenport, T., Harris, J.: Competing on analytics: Updated, with a new introduction: The new science of winning. Harvard Business Press (2017)
16. Guggenberger, T., Möller, F., Boualouch, K., Otto, B.: Towards a unifying understanding of digital business models. In: PACIS (2020)
17. Redman, T.C.: To improve data quality, start at the source, `https://hbr.org/2020/02/to-improve-data-quality-start-at-the-source`, (Accessed: 13.08.2021)
18. Ehrlinger, L., Rusz, E., Wöß, W.: A survey of data quality measurement and monitoring tools. arXiv preprint arXiv:1907.08138 (2019)
19. Hameed, M., Naumann, F.: Data preparation: A survey of commercial tools. ACM SIGMOD Record 49(3), 18–29 (2020)

20. Rekatsinas, T., Chu, X., Ilyas, I.F., Ré, C.: Holoclean: Holistic data repairs with probabilistic inference. arXiv preprint arXiv:1702.00820 (2017)
21. Oliveira, P., Rodrigues, F., Henriques, P.R.: A formal definition of data quality problems. In: ICIQ (2005)
22. Bosu, M.F., MacDonell, S.G.: A taxonomy of data quality challenges in empirical software engineering. In: 2013 22nd Australian Software Engineering Conference. pp. 97–106. IEEE (2013)
23. de Almeida, W.G., de Sousa, R.T., de Deus, F.E., Nze, G.D.A., de Mendonça, F.L.L.: Taxonomy of data quality problems in multidimensional data warehouse models. In: 2013 8th Iberian Conference on Information Systems and Technologies (CISTI). pp. 1–7. IEEE (2013)
24. Cockcroft, S.: A taxonomy of spatial data integrity constraints. GeoInformatica 1(4), 327–343 (1997)
25. Paukstadt, U., Strobel, G., Eicker, S.: Understanding services in the era of the internet of things: A smart service taxonomy. ECIS 2019 Proceedings (2019)
26. Rosian, M., Hagenhoff, P., Otto, B.: Towards a holistic cloud computing taxonomy: Theoretical & practical findings. AMCIS 2021 Proceedings (2021)
27. Möller, F., Stachon, M., Hoffmann, C., Bauhaus, H., Otto, B.: Data-driven business models in logistics: A taxonomy of optimization and visibility services. HICSS 2020 Proceedings pp. 5379–5388 (2020)
28. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design science in information systems research. MIS Quarterly pp. 75–105 (2004)
29. Brocke, J.v., Simons, A., Niehaves, B., Niehaves, B., Reimer, K., Plattfaut, R., Cleven, A.: Reconstructing the giant: On the importance of rigour in documenting the literature search process. ECIS 2009 Proceedings (2009)
30. Kuhrmann, M., Fernández, D.M., Daneva, M.: On the pragmatic design of literature studies in software engineering: an experience-based guideline. Empirical software engineering 22(6), 2852–2891 (2017)
31. Madnick, S.E., Wang, R.Y., Lee, Y.W., Zhu, H.: Overview and framework for data and information quality research. Journal of Data and Information Quality (JDIQ) 1(1), 1–22 (2009)
32. Webster, J., Watson, R.T.: Analyzing the past to prepare for the future: Writing a literature review. MIS Quarterly 26(2), 13–23 (2002)
33. Ubisoft: Mobydq, `https://github.com/ubisoft/mobydq`, (Accessed: 09.08.2021)
34. Tamr: Tamr, `https://www.tamr.com/`, (Accessed: 19.05.2021)
35. Ataccama: Enterprise data quality fabric (2021), `https://www.ataccama.com/`, (Accessed: 19.05.2021)
36. Kovac, R., Lee, Y.W., Pipino, L.: Total data quality management: The case of iri. In: IQ. pp. 63–79 (1997)
37. Krogstie, J.: Capturing enterprise data integration challenges using a semiotic data quality framework. Business & Information Systems Engineering 57(1), 27–36 (2015)
38. Asghari, M., Sierra-Sosa, D., Elmaghraby, A.: A semi-automatic system for data management and cleaning. In: 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT). pp. 393–397. IEEE (2018)
39. Debattista, J., Auer, S., Lange, C.: Luzzu—a methodology and framework for linked data quality assessment. Journal of Data and Information Quality 8(1), 1–32 (2016)
40. Falge, C., Otto, B., Österle, H.: Towards a strategy design method for corporate data quality management. In: Wirtschaftsinformatik. pp. 801–815 (2013)
41. Yeoh, W., Wang, T.W., Verbitskiy, Y.: Describing data quality problem through a metadata framework. AMCIS 2012 Proceedings pp. 1–10 (2012)

42. Ali, K., Warraich, M.A.: A framework to implement data cleaning in enterprise data warehouse for robust data quality. In: 2010 International Conference on Information and Emerging Technologies. pp. 1–6. IEEE (2010)
43. Robbert, A., Senne, L.: Data quality: An attribute framework for large systems. AMCIS 2006 Proceedings (2006)
44. Barateiro, J., Galhardas, H.: A survey of data quality tools. Datenbank-Spektrum 14, 15–21 (2005)
45. Wang, Z., Talburt, J.R., Wu, N., Dagtas, S., Zozus, M.N.: A rule-based data quality assessment system for electronic health record data. Applied clinical informatics 11(4), 622–634 (2020)
46. Peng, M., Lee, S., D'Souza, A.G., Doktorchik, C.T.A., Quan, H.: Development and validation of data quality rules in administrative health data using association rule mining. BMC medical informatics and decision making 20(1), 75 (2020)
47. Cichy, C., Rass, S.: An overview of data quality frameworks. IEEE Access 7, 24634–24648 (2019)
48. Karkouch, A., Mousannif, H., Moatassime, H.A., Noel, T.: A model-driven architecture-based data quality management framework for the internet of things. In: 2016 2nd International Conference on Cloud Computing Technologies and Applications (CloudTech). pp. 252–259. IEEE (2016)
49. Pulla, V.S.V., Varol, C., Al, M.: Open source data quality tools: Revisited. In: Information Technology: New Generations, pp. 893–902. Springer (2016)
50. Khayyat, Z., Ilyas, I.F., Jindal, A., Madden, S., Ouzzani, M., Papotti, P., Quiané-Ruiz, J.A., Tang, N., Yin, S.: Bigdansing. In: Sellis, T., Davidson, S.B., Ives, Z. (eds.) Compilation proceedings of the 2015 ACM Symposium on Principles of Database Systems, ACM SIGMOD International Conference on Management of Data, and SIGMOD/PODS. pp. 1215–1230. ACM (2015)
51. Weber, D., Leone, S., Norrie, M.: Constraint-based data quality management framework for object databases. ECIS 2013 Completed Research pp. 1–12 (2013)
52. Geisler, S., Weber, S., Quix, C.: Ontology-based data quality framework for data stream applications. 16th International Conference on Information Quality pp. 1–15 (2011)
53. Alkharboush, N., Li, Y.: A decision rule method for data quality assessment. ICIQ pp. 1–12 (2010)
54. Gao, J., Koronios, A.: Snap-on data quality enhancement and verification tools (deva) for asset management. ICIQ pp. 1–12 (2010)
55. Pham Thi, T.T., Helfert, M.: Discovering dynamic integrity rules with a rules-based tool for data quality analyzing. In: Rachev, B., Smrikarov, A. (eds.) Proceedings of the 11th International Conference on Computer Systems and Technologies CompSysTech '10. p. 89. ACM Press, New York, New York, USA (2010)
56. Liu, B., Pan, J.H., Liu, P.S.: Rule evaluation method and data quality mining system. Computer Integrated Manufacturing Systems 15, 1436–1441 (2009)
57. Luebbers, D., Grimmer, U., Jarke, M.: Systematic development of data mining-based data quality tools. In: Proceedings 2003 VLDB Conference, pp. 548–559. Elsevier (2003)
58. Hipp, J., Güntzer, U., Grimmer, U.: Data quality mining: Making a virute of necessity. In: DMKD. pp. 1–6 (2001)
59. Svanks, M.I.: Integrity analysis. Information and Software Technology 30(10), 595–605 (1988)
60. Woodall, P., Oberhofer, M., Borek, A.: A classification of data quality assessment and improvement methods. International Journal of Information Quality 3(4), 298–321 (2014)
61. Assaf, A., Senart, A., Troncy, R.: Towards an objective assessment framework for linked data quality. In: Management Association, I.R. (ed.) Information Retrieval and Management, pp. 453–478. IGI Global (2018)

62. Long, J.A., Seko, C.E.: A cyclic-hierarchical method for database data-quality evaluation and improvement. In: Wang, Y.R. (ed.) Information quality, pp. 52–66. Advances in management information systems, Routledge, Abingdon, Oxon (2015)
63. Ofner, M.H., Huener, K.M., Otto, B.: Dealing with complexity: a method to adapt and implement a maturity model for corporate data quality management. AMCIS 2009 Proceedings pp. 491–503 (2009)
64. Batini, C., Cabitza, F., Cappiello, C., Francalanci, C.: A comprehensive data quality methodology for web and structured data. International Journal of Innovative Computing and Applications 1(3), 448–456 (2008)
65. Carvalho, C., Moreira, R.S., Torres, J.M.: Data quality visual analysis (dqva) a tool to process and pinspot raw data irregularities. In: 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC). pp. 1036–1045. IEEE
66. Al Chami, Z., Al Bouna, B., Jaoude, C.A., Chbeir, R.: A real-time multimedia data quality assessment framework. In: Chbeir, R., Manolopoulos, Y., Ilarri, S., Papadopoulos, A. (eds.) Proceedings of the 11th International Conference on Management of Digital EcoSystems. pp. 270–276. ACM (2019)
67. Sulistyo, H.A., Kusumasari, T.F., Alam, E.N.: Implementation of data cleansing null method for data quality management dashboard using pentaho data integration. In: Kusrini, K. (ed.) Exploring the role of artificial intelligence for creative industry 4.0. pp. 12–16. IEEE (2020)
68. Pezoulas, V.C., Kourou, K.D., Kalatzis, F., Exarchos, T.P., Venetsanopoulou, A., Zampeli, E., Gandolfo, S., Skopouli, F., de Vita, S., Tzioufas, A.G., Fotiadis, D.I.: Medical data quality assessment: On the development of an automated framework for medical data curation. Computers in biology and medicine 107, 270–283 (2019)
69. Ehrlinger, L., Werth, B., Wöß, W.: Automated continuous data quality measurement with quaiie. International Journal on Advances in Software 11(3), 400–417 (2018)
70. Jesiļevska, S.: Iterative method for reducing the impact of outlying data points: Ensuring data quality. Statistical Journal of the IAOS 32(2), 257–263 (2016)
71. Bai, L., Meredith, R., Burstein, F.: A data quality framework, method and tools for managing data quality in a health care setting: an action case study. Journal of Decision Systems 27(1), 144–154 (2018)
72. Effendy, M.R., Kusumasari, T.F., Hasibuan, M.A.: Star schema implementation for monitoring in data quality management tool (a case study at a government agency). In: 2019 Fourth International Conference on Informatics and Computing (ICIC). pp. 1–6. IEEE
73. Batini, C., Barone, D., Mastrella, M., Maurino, A., Ruffini, C.: A framework and a methodology for data quality assessment and monitoring. In: ICIQ. pp. 333–346. Citeseer (2007)
74. To, A., Meymandpour, R., Davis, J.G., Jourjon, G., Chan, J.: A linked data quality assessment framework for network data. In: Arora, A., Bhattacharya, A., Fletcher, G. (eds.) Proceedings of the 2nd Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA) - GRADES-NDA'19. pp. 1–8 (2019)
75. Zhang, G.: A data traceability method to improve data quality in a big data environment. In: 2020 IEEE Fifth International Conference on Data Science in Cyberspace (DSC). pp. 290–294. IEEE (2020)
76. Liu, Q., Feng, G., Zhao, X., Wang, W.: Minimizing the data quality problem of information systems: A process-based method. Decision Support Systems 137, 113381 (2020)
77. Sun, Y., Lu, T., Gu, N.: A method of electronic health data quality assessment: Enabling data provenance. In: 2017 IEEE 21st International Conference on Computer Supported Cooperative Work in Design (CSCWD). pp. 233–238. IEEE (2017)
78. Weichselbraun, A., Kuntschik, P.: Mitigating linked data quality issues in knowledge-intense information extraction methods. In: Akerkar, R., Cuzzocrea, A., Cao, J., Hacid, M.S. (eds.) Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics. pp. 1–12. ACM, New York, NY, USA (2017)

79. El Bekri, N., Peinsipp-Byma, E.: Assuring data quality by placing the user in the loop. In: 2016 International Conference on Computational Science and Computational Intelligence (CSCI). pp. 468–471. IEEE (2016)
80. Bargh, M., Mbong, F., van vanDijk, J., Choenni, R.: A framework for dynamic data quality management. In: EC 2015 Proceedings: Proceedings of the IADIS International Conference e-Commerce and Digital Marketing. Hogeschool Rotterdam (2015)
81. Kirschner, L., Soremekun, E., Zeller, A.: Debugging inputs. In: Rothermel, G., Bae, D.H. (eds.) Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering. pp. 75–86. ACM, New York, NY, USA (2020)
82. Pellegrino, M.A.: Methods and techniques for data quality improvement of (linked)(open) data. CEUR Workshop Proceedings (2019)
83. Wang, Z., Fu, Y., Song, C., Ge, W., Qiao, L., Zhang, H.: A data quality improvement method based on the greedy algorithm. In: Zhai, X.B., Chen, B., Zhu, K. (eds.) Machine Learning and Intelligent Communications. pp. 256–266. Springer International Publishing, Cham (2019)
84. De, S., Hu, Y., Meduri, V.V., Chen, Y., Kambhampati, S.: Bayeswipe: A scalable probabilistic framework for improving data quality. Journal of Data and Information Quality 8(1), 1–30 (2016)
85. Cahsai, A., Anagnostopoulos, C., Triantafillou, P.: Scalable data quality for big data: The pythia framework for handling missing values. Big data 3(3), 159–172 (2015)
86. Song, S., Sun, Y., Zhang, A., Chen, L., Wang, J.: Enriching data imputation under similarity rule constraints. IEEE Transactions on Knowledge and Data Engineering 32(2), 275–287 (2020)
87. Strong, D.M., Lee, Y.W., Wang, R.Y.: Data quality in context. Communications of the ACM 40(5), 103–110 (1997)
88. Oracle: Oracle enterprise data quality (2021), `https://www.oracle.com/middleware/technologies/enterprise-data-quality.html`, (Accessed: 28.05.2021)
89. SAP: Sap information steward (2021), `https://www.sap.com/products/data-profiling-steward.html`, (Accessed: 28.05.2021)
90. Infogix: Data 360: Data quality (2021), `https://www.infogix.com/data3sixty/dq/`, (Accessed: 23.06.2021)
91. SAS: Sas data quality (2021), `https://www.sas.com/en_us/software/data-quality.html`, (Accessed: 19.05.2021)
92. MIOSoft: The platform for enterprises that are serious about data quality (2021), `https://www.miosoft.com/solutions/data-quality.html`, (Accessed: 28.05.2021)
93. Precisely: Data quality solutions (2021), `https://www.precisely.com/solution/data-quality-solutions`, (Accessed: 28.05.2021)
94. Azkan, C., Iggena, L., Gür, I., Möller, F., Otto, B.: A taxonomy for data-driven services in manufacturing industries. In: PACIS. pp. 184–198 (2020)
95. Labadie, C., Legner, C., Eurich, M., Fadler, M.: Fair enough? enhancing the usage of enterprise data with data catalogs. In: 22nd Conference on Business Informatics (CBI). pp. 201–210. IEEE (2020)
96. Experian: Data quality management (2021), `https://www.edq.com/data-quality-management/`, (Accessed: 19.05.2021)
97. Human Inference: Datahub, `https://www.humaninference.com/en/solutions/datahub`, (Accessed: 19.05.2021)
98. IBM: Data quality: Cleanse data, manage it and support better decision-making (2021), `https://www.ibm.com/analytics/data-quality`, (Accessed: 19.05.2021)

99. Informatica: Informatica data quality: Deliver strategic value, quickly (2021), https://www.informatica.com/products/data-quality/informatica-data-quality.html, (Accessed: 28.05.2021)
100. InfoZoom: Data analyses, data profiling & data quality management with infozoom (2021), https://www.infozoom.com/en/, (Accessed: 23.06.2021)
101. Innovative Systems: Enlighten data quality (2021), https://www.innovativesystems.com/data-quality, (Accessed: 28.05.2021)
102. Melissa: Data quality solutions (2021), https://www.melissa.com/data-quality, (Accessed: 28.05.2021)
103. datamartist: A fast, easy to use, visual data profiling and transformation tool. (2021), http://www.datamartist.com/, (Accessed: 28.05.2021)
104. Redpoint: Data quality & integration (2021), https://www.redpointglobal.com/data-integration/, (Accessed: 29.05.2021)
105. Syniti: 100% data quality (2021), https://www.syniti.com/solutions/data-quality/, (Accessed: 29.05.2021)
106. Talend: Data quality solutions: Profile, clean, and standardize data across your systems (2021), https://www.talend.com/products/data-quality/, (Accessed: 19.05.2021)
107. Musleh, M., Ouzzani, M., Tang, N., Doan, A.: Coclean: Collaborative data cleaning. In: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data. p. 2757–2760. SIGMOD '20, Association for Computing Machinery (2020)
108. Shankaranarayanan, G., Blake, R.: From content to context: The evolution and growth of data quality research. Journal of Data and Information Quality (JDIQ) 8(2), 1–28 (2017)