Understanding the differences across data quality classifications
a literature review and guidelines for future research

Haug, Anders

Published in:
Industrial Management & Data Systems

DOI:
10.1108/IMDS-12-2020-0756

Publication date:
2021

Document version:
Accepted manuscript

Citation for pulished version (APA):
Haug, A. (2021). Understanding the differences across data quality classifications: a literature review and guidelines for future research. *Industrial Management & Data Systems*, *121*(12), 2651-2671. https://doi.org/10.1108/IMDS-12-2020-0756

Understanding the differences across data quality classifications: A literature review and guidelines for future research

Anders Haug
Department of Entrepreneurship and Relationship Management, University of Southern Denmark,
Universitetsparken 1, DK-6000 Kolding, Denmark.
Email: adg@sam.sdu.dk

Abstract

Purpose – Numerous data quality (DQ) definitions in the form of sets of DQ dimensions are found in the literature. The great differences across such DQ classifications (DQCs) imply a lack of clarity about what DQ is. To provide an improved foundation for future research, this article clarifies the ways in which DQCs differ and provides guidelines for dealing with this variance.

Design/methodology/approach – A literature review identifies DQCs in conference and journal articles, which are analyzed to reveal the types of differences across these. On this basis, guidelines for future research are developed.

Findings – The literature review found 110 unique DQCs in journals and conference articles. The analysis of these articles identified seven distinct types of differences across DQCs. This gave rise to the development of seven guidelines for future DQ research.

Originality – The literature review did not identify articles, which, based on systematic searches, identify and analyze existing DQCs. Thus, this article provides new knowledge on the variance across DQCs, as well as guidelines for addressing this.

Research limitations/implications – By identifying differences across DQCs and providing a set of guidelines, this article may promote that future research, to a greater extent, will converge around common understandings of DQ.

Practical implications – Awareness of the identified types of differences across DQCs may support managers when planning and conducting DQ improvement projects.

Keywords Data Quality, Information Quality, Data Management, Information Management, Data Quality Dimensions, Information Quality Dimensions

Paper type Research paper

1. Introduction

Data quality (DQ) has become one of the most significant challenges for companies in the digital age with regard to ensuring efficiency and quality in processes and products (Experian, 2016; Experian, 2019; IBM, 2016; KPMG, 2017). In this context, a recent global cross-industrial survey found that 95% of companies experience negative impacts from poor DQ (Experian, 2019), while another survey found that 84% of the CEOs interviewed are concerned about their DQ (KPMG, 2017). The impacts of these DQ issues are enormous. Specifically, IBM (2016) estimated that in 2016 in the US alone, poor DQ cost the US economy $3.1 trillion, an amount close to the total US federal expenditure that year ($3.9 trillion). Despite improved technologies for managing data, the economic costs of DQ problems seem to be increasing (Experian, 2016; Experian, 2019; O'Brien, 2015).

Given the magnitude of the economic costs of poor DQ, one might assume that DQ problems are systematically addressed by companies. However, this does not seem to be the case, as, for example, illustrated by a survey of German companies, which showed that 63% determine their DQ manually and ad hoc, while not employing a long-term DQ management strategy (Schäffer and Beckmann, 2014). Similarly, an Experian survey showed that almost 80% of companies lack structured and systematic approaches for ensuring DQ (Experian, 2016). Based on a review of previous surveys, O'Brien (2015)

argues that evidence suggests that many organizations are unaware of the extent of their DQ problems, while others either ignore or do not prioritize such issues.

For the reasons described above, significant attention has been given to DQ in academic and practitioner literature since the mid-90s (as illustrated by the subsequent literature review). In contemporary information systems research, DQ topics are often addressed within its subfields, such as data improvement methods (e.g., Arazy *et al.*, 2017; Asghari *et al.*, 2020), social media (Ferretti *et al.*, 2018; Moravec *et al.*, 2019), mobile apps (Leon, 2018; Zhou, 2013), healthcare (Li and Qin, 2017; Savitza *et al.*, 2020), big data (Ghasemaghaei *et al.*, 2018; Popovič *et al.*, 2018), and enterprise systems (Jayawickrama *et al.*, 2017; Laumer *et al.*, 2017; Ram *et al.*, 2013; Reinking *et al.*, 2020). The research on DQ has provided many valuable insights, but it has also resulted in a plethora of different definitions of DQ. In other words, despite the intensive research and ongoing discussions on DQ, there is no consensus about which DQ dimensions constitute DQ (Liaw *et al.*, 2013; Lóscio *et al.*, 2013; Sebastian-Coleman, 2012). There is, however, there is agreement that DQ is a multidimensional concept (Shamala *et al.*, 2017), which means that DQ is perceived as a set of DQ dimensions that each describes a particular characteristic of DQ (sometimes grouped under DQ categories). In the remainder of this article, the term "data quality classification" (DQC) is used to highlight that it is such multidimensional DQ definitions that are in focus.

The great differences across DQCs have been explained by the fact that data can represent various aspects of real-world phenomena, and, for this reason, different DQCs are suitable for different areas of application (e.g., Batini and Scannapieco, 2006). At best, however, this only offers a partial explanation, since although DQCs focus on the same application area, in many cases, such DQCs vary greatly (as the literature review in this article later demonstrates). Thus, a more fundamental explanation may, as argued by Shamala *et al.* (2017), be that most DQ frameworks (hereunder DQCs) are ad hoc, intuitive and incomplete, and may therefore not produce robust and systematic measurement models. Consequently, the DQ dimensions included in DQCs often overlap, are vaguely defined, and do not have a solid basis in theory (Forsgren *et al.*, 2016; Zhang *et al.*, 2019). In this context, Jayawardene *et al.* (2015) argue that "while it is important to embrace the diversity of views of DQ, it is equally important for the DQ research and practitioner community to be united in the consistent interpretation of this foundational concept."

The numerous DQCs in the literature would appear to make it challenging for practitioners to determine which DQ dimensions to apply. Furthermore, if they manage to choose a suitable set of DQ dimensions, they then face the problem of a lack of clarity about how to map the dimensions and metrics from DQ research to practical implementations (Batini *et al.*, 2014, p.70; Liaw *et al.*, 2013). In fact, according to Sebastian-Coleman (2012), a central challenge for DQ practitioners when engaging in DQ improvement is determining how to measure DQ. For researchers, the differences in DQCs and meanings attributed to their DQ dimensions make it hard to compare the results of different studies and build on existing research. To address this issue, some attempts have been made to consolidate the DQ literature.

One of the first more extensive literature reviews of DQ dimensions was carried out by Wang *et al.* (1995b). Their results include a count of how frequently each of the 26 identified DQ dimensions was mentioned in the literature (Wand and Wang, 1996). Since then, much DQ research has been produced, rendering this type of study much more extensive. Maybe, for this reason, more contemporary literature reviews have mainly focused on DQ within subfields of information systems research. These include reviews of DQ dimensions in relation to the Internet of Things (Liu *et al.*, 2020), e-government websites (Rasool and Warraich, 2018), electronic health records (Weiskopf and Weng, 2013), and sensor data quality (Teh *et al.*, 2020). Besides focusing only on certain contexts, these studies do not analyze differences across DQCs.

Discussions of DQCs are, however, not entirely absent from the literature, but such articles do not identify DQCs based on systematic searches, and they include only a fraction of all DQCs. Specifically, the literature review of this article did not identify previous attempts to provide an extensive account of the different DQCs or to investigate the reasons for the variance across these. Furthermore, the literature review did not identify a structured set of guidelines to guide DQC development. Thus, this article raises two questions:

RQ1: What are the differences across DQCs in the academic literature?
RQ2: Which aspects should be considered in future DQC development?

These questions are addressed through a literature review with the aim of identifying differences across DQCs. On this basis, a set of guidelines for future DQ research is outlined. Hereby, this article contributes to future research in three ways. First, the findings of this article may support researchers in deciding on which DQCs to build upon in studies involving DQ. Second, the guidelines proposed in this article may support research aiming to develop new DQCs for particular contexts. Finally, the article provides a foundation for research focusing on consolidating existing DQCs within certain contexts.

2. Methodology

The purpose of the present literature review was to identify existing DQCs in the academic literature as a basis for identifying types of differences in the dimensions included in DQCs. The literature review followed the methodological approach of Leidner and Kayworth (2006) of developing (1) a strategy for searching prior literature, (2) criteria for selecting literature to be included in the review, and (3) a scheme for documenting and analyzing selected literature.

2.1. Database searches

To identify relevant literature on DQCs, the search strategy consisted of two types of searches: (1) searches in titles, abstracts, and keywords in research databases, and (2) identification of additional articles using the references in the identified articles. For the research database search, the following search string was used:

> *data/information quality dimensions/framework/categorization/classification OR dimensions/framework/categorization/classification of data/information quality*

Searches were conducted using the Web of Science and Scopus databases. These databases were chosen since they include the top information systems journals and conferences, and because they offer flexible search functionalities (as opposed to Google Scholar). In other words, it could be expected to find the most relevant DQCs through these searches, either in the articles identified or through the references in these articles. The searches were conducted in January 2021. A filter was used to include only peer-reviewed and English-language journal and conference articles. The search in Web of Science returned 231 results, while the search in Scopus returned 439 results. This amounts to 670 articles in total and 468 after removing duplicates.

2.2. Selection criteria

The identified 468 articles were inspected to determine if they involved DQC development. The applied criteria concerned if DQCs include novel combinations of DQ dimensions. Thus, first, the articles that directly used existing DQCs were discarded. Next, DQCs with references to previous literature were compared to their sources, and the ones without novelty were discarded. In this context, it should be noted that, besides the articles that directly focus on the creation of new DQCs, some authors define a novel

combination of categories and/or dimensions to answer other questions. Such authors typically include selections of between three and six common dimensions used to study or explain a certain phenomenon, while to varying extent recognizing that other dimensions could also be relevant. Thus, compared to articles whose main focus is on the creation of a DQC, in some cases, these have built-in reservations with regard to their completeness. Nevertheless, such sets of DQ dimensions represent unique DQCs in the sense that they include a unique subset of DQ dimensions. These were therefore included in the count, which amounted to 37 DQCs.

2.3. Identification of additional articles

The second part of the search aimed at identifying additional DQCs, which had not been identified by the database search. This was done by inspecting the references in the identified 468 articles to determine if they referred to articles involving DQC development. This analysis resulted in 257 additional articles. Using the selection criteria described above, the analysis identified an additional 72 DQCs in journal and conference articles.

2.4. Analysis of results

In total, the searches identified 110 unique DQCs in journal and conference articles. The 110 articles identified were analyzed to understand their foundation, assumptions, focus, delimitations, and content to make comparisons and generalizations across DQCs. This analysis was done using a coding scheme to interpret data and to develop categories for detailed discussions. Specifically, the analysis involved "open coding", followed by "axial coding" (Strauss, 1998; Flick, 2009). In the open coding process, parts of the articles related to definitions of the proposed DQCs were identified, while the axial coding process organized the identified concepts into categories describing types of differences across DQCs. These results are described in the following sections.

3. Descriptive findings

Until 1993, only sporadic contributions were identified. Specifically, from 1978 to 1992, only five DQCs were identified. Thus, Figure 1 only shows publications from 1993 to 2020.

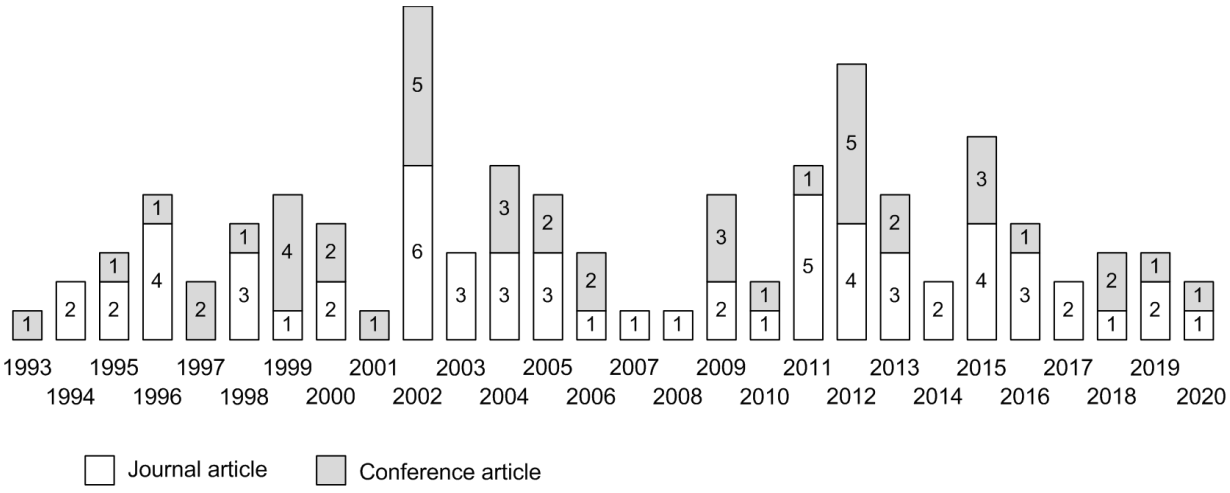

Figure. 1. Yearly distribution of DQCs 1993-2020.

The DQCs from the academic literature involve different foundations for the extraction of DQ categories and dimensions. In this context, five distinct approaches were identified: (1) empirical study; (2) ontological approach; (3) systematic literature review; (4) literature references; and (5) limited or no

arguments. The DQCs based on empirical studies involve surveys of students, managers, or employees regarding their perceptions of DQ dimensions or encounters with DQ issues. The results of such surveys are processed by producing a candidate list of DQ dimensions that is analyzed and organized into a set of DQ dimensions (e.g., Wang and Strong, 1996). Secondly, the ontological approach involves deriving DQ dimensions based on a set of assumptions about reality. This approach is known from the work of Wand and Wang (1996), who derived DQ dimensions based on the identification of possible problematic relationships between real-world aspects and information system representations thereof. Thirdly, structured literature reviews (however, limited to a small subset of journals) involve explicit accounts of literature searches for DQ dimensions, on which basis the identified dimensions are organized into a DQC (e.g., Jayawardene *et al.*, 2015). Fourthly, some articles refer to parts of the DQ literature while not considering other parts. For such studies, the level of argumentation provided includes: (1) providing limited, if any, argumentation for the selection of references to the literature used involving the DQ dimensions included in their DQCs, (2) using highly cited DQCs or dimensions as a basis for DQC construction, and (3) using reviews carried out by others as a basis. Fifthly, in some cases, selections of DQC dimensions or categories are proposed without much argumentation or literature references.

To give an impression of the variety across DQCs and to lay a foundation for the subsequent discussions, some of the most cited DQCs that attempt to provide a relatively complete account of DQ dimensions are shown in Table 1 and 2, divided into ones having a general and more specific focus. It should be noted that, although more recent DQCs exist (e.g., Gürdür *et al.*, 2019; Huang, 2018; Liu *et al.*, 2020; Rajan *et al.*, 2019; Rasool and Warraich, 2018; Teh *et al.*, 2020), the ones shown are still frequently cited in current DQ research. In the tables, the final two columns show the number of Scopus (Sc) citations and the number of Google Scholar (GS) citations (citations counts were conducted on August 14, 2020). Web of Science citations are not included, as less than half of the identified articles are indexed here.

Table 1. General information systems DQCs.

Author	DQ categories	DQ dimensions	Sc	GS
Wang *et al.* (1995a)	Accessible	Available	147	375
	Interpretable	Syntax, Semantics		
	Useful	Relevant, Timely (current, non-volatile)		
	Believable	Complete, Consistent, Credible source, Accurate		
Miller (1996)	None	Relevance, Accuracy, Timeliness, Completeness, Coherence, Format, Accessibility, Compatibility, Security, Validity	135	351
Wang and Strong (1996)	Intrinsic	Accuracy, Objectivity, Believability, Reputation	2162	4745
	Contextual	Value-added, Relevancy, Timeliness, Completeness, Appropriate amount of data		
	Representational	Interpretability, Ease of understanding, Representational consistency, Concise representation		
	Accessibility	Accessibility, Access security		
Wand and Wang (1996)	Intrinsic	Completeness, Unambiguousness, Meaningfulness, Correctness	819	1855
	Extrinsic	(not in focus)		
Wang (1998)	None	Free of error, Objectivity, Reputation, Believability, Relevance, Value-added, Timeliness, Completeness, Appropriate amount, Interpretability, Understandability, Consistent representation, Concise representation, Ease of manipulation, Accessibility, Security	558	1204
Naumann and Rolker (2000)	Subject	Believability, Concise representation, Interpretability, Relevancy, Reputation, Understandability, Value-added	-	438
	Object	Completeness, Customer support, Documentation, Objectivity, Price, Reliability, Security, Timeliness, Verifiability		
	Process	Accuracy, Amount of data, Availability, Consistent representation, Latency, Response time		

Kahn *et al.* (2002)	Soundness (product)	Free of error, Concise representation, Completeness, Consistent representation	455	1030
	Usefulness (product)	Appropriate amount, Relevancy, Understandability, Interpretability, Objectivity		
	Dependability (service)	Timeliness, Security		
	Usability (service)	Believability, Accessibility, Ease of operation, Reputation, Value-added		
Pipino *et al.* (2002)	(none)	Accessibility, Appropriate amount of data, Believability, Completeness, Concise/consistent representation, Ease of manipulation, Free-of-error, Interpretability, Objectivity, Relevancy, Reputation, Security, Timeliness, Understandability, Value-added	942	1919
Lee *et al.* (2002)	None	Accessibility, Appropriate amount, Believability, Completeness, Concise representation, Consistent representation, Ease of operation, Free-of-error, Interpretability, Objectivity, Relevancy, Reputation, Security, Timeliness, Understandability	895	1826
Bovee *et al.* (2003)	Accessibility	Accessibility	109	324
	Interpretability	Intelligible, Meaningful		
	Relevance	Age, Volatility, Other user-specified criteria		
	Integrity	Accuracy, Completeness, Consistency, Existence		
Stvilia *et al.* (2007)	Intrinsic	Accuracy/validity, Cohesiveness, Complexity, Semantic consistency, Structural consistency, Currency, Informativeness/redundancy, Naturalness, Precision/completeness	254	497
	Relational/contextual	Accuracy, Accessibility, Complexity, Naturalness, Informativeness/redundancy, Relevance/aboutness, Precision/completeness, Security, Semantic consistency, Structural consistency, Verifiability, Volatility		
	Reputational	Authority		
Gorla *et al.* (2010)	Content	Accurate, Complete, Concise, Useful, Relevant	341	806
	Format	Appearance/format, Comparable/consistency, Easy to understand		

Table 2. DQCs with other focuses.

Author (Focus)	**DQ categories**	**DQ dimensions**	**Sc**	**GS**
Jarke *et al.* (1999) (Data warehousing)	Scheme quality	Correctness, Completeness, Minimality, Traceability, Interpretability	97	265
	Metadata evolution	Metadata evolution		
	Accessibility	System, Transactional, Security		
	Usefulness	Interpretability, Timeliness, Volatility, Responsiveness		
Katerattanakul and Siau (1999) (Websites)	Intrinsic	Accuracy/errors of content, Accurate/workable/relevant hyperlinks	-	356
	Contextual	Provision of author's information		
	Representational	Organization, Visual settings, Typographical features, Consistency, Vividness, Attractiveness, Confusion of the content		
	Accessibility	Navigational tools provided		
Zhu and Gauch (2000) (Internet search systems)	Availability	Accessibility, Timeliness	127	293
	Usability	Credibility		
	Reliability	Accuracy, Consistency, Integrity, Completeness		
	Relevance	Fitness		
	Presentation	Readability		
Chae *et al.* (2002) (Mobile internet services)	Connection	Stability, Responsiveness	-	447
	Contextual	Timeliness, Promptness		
	Interaction	Structure, Presentation, Navigation		
	Content	Objectivity, Believability, Amount		
Rieh (2002) (Web information)	None	Source, Content, Format, Presentation, Currency, Accuracy, Speed of loading	444	988

Xu *et al.* (2002) (ERP systems)	Intrinsic	Multiple sources of same data, Questionable believability, Judgement involved in data production, Questionable objectivity, Poor reputation, Little added value	123	270
	Accessibility	Lack of computing resources, Poor accessibility, Access security, interpretability and understandability, Concise/consistent representation, Amount of data, Timeliness		
	Contextual	Operational data production problems, Changing data consumer needs, Incomplete data, Poor relevancy, Distributed computing, Inconsistent representation, Little value added		
Nelson *et al.* (2005) (Data warehousing)	Intrinsic/Accuracy	Correct, Unambiguous, Meaningful, Believable, Consistent	493	997
	Contextual	Completeness, Currency (up-to-date, precision)		
	Representational	Format (understandable, representational interpretable)		
Zaveri *et al.* (2012) (Linked data on the web)	Accessibility	Availability, Licensing, Interlinking, Security, Performance	236	411
	Intrinsic	Accuracy, Consistency, Conciseness, Completeness		
	Contextual	Relevancy, Trustworthiness, Understandability, Timeliness		
	Representational	Representational Conciseness, Interoperability, Versatility		
Weiskopf and Weng (2013) (E-health records)	Completeness	Accessibility, Accuracy, Availability, Missingness, Omission, Presence, Quality, Rate of recording, Sensitivity, Validity	392	697
	Correctness	Accuracy, Corrections made, Errors, Misleading, Positive predictive value, Quality, Validity		
	Concordance	Agreement, Consistency, Reliability, Variation		
	Plausibility	Accuracy, Believability, Trustworthiness, Validity		
	Currency	Recency, Timeliness		
Cai and Zhu (2015) (Big data)	Availability	Accessibility, Timeliness	192	423
	Usability	Credibility		
	Reliability	Accuracy, Consistency, Integrity, Completeness		
	Relevance	Fitness		
	Presentation	Readability		

Besides using DQCs from the journal and conference articles in their studies, many studies also apply DQCs from books. In this context, some of the most oft-used DQCs include the ones from Redman and Blanton (1997), English (1999), Loshin (2001), Redman (2001), Wang *et al.* (2002), and Eppler (2006). Furthermore, some articles also use DQCs found in standards (e.g., ISO/IEC, 2015) and reports (e.g., Jayawardene *et al.*, 2015).

4. Analysis of data quality definitions

The analysis of the 110 DQCs led to the identification of seven distinct types of differences across DQCs. These are shown in Table 3, together with some examples from the open coding process. Subsequently, the seven types of differences across DQCs are discussed.

Table 3. Analysis of DQCs in the identified articles.

No.	DQ categories (axial coding)	Examples from the open coding process
1	DQC focus	"This framework consists of big data quality dimensions…" (Cai and Zhu, 2015)
		"…dimensions of data quality assessment in the context of electronic health record (EHR) data…" (Weiskopf and Weng, 2013)
2	DQ evaluation perspective	"…the definitions take either an intrinsic or a contextual view of information quality." (Nelson *et al.*, 2005)
		"Intrinsic IQ measurements reference general cultural norms and conventions. Relational measures reference the immediate context…" (Stvilia *et al.* (2007)
3	DQ category set	"…we first categorize the relevant data warehouse quality dimensions…" (Jarke *et al.*, 1999)

		"We chose to use an iterative process to label and categorize the dimensions of data quality…" (Weiskopf and Weng, 2013)
4	DQ dimension set	"…literature to identify IoT DQ dimensions…" (Liu *et al.*, 2020)
		"Various data quality methodologies are based on identifications of data quality dimensions…" (Zhang *et al.*, 2019)
5	Data level focus	"…most of the methods are applicable at the triple or graph level and to a lesser extend on the dataset level." (Zaveri *et al.*, 2012)
		"This paper focuses on the tabular datasets…" (Even and Shankaranarayanan, 2005)
6	Data structure focus	"…social media data are usually unstructured, and their consistency and integrity are not suitable for evaluation." (Cai and Zhu, 2015)
		"In the field of data quality, most authors either implicitly or explicitly distinguish three types of data… [structured, unstructured and semi-structred]" (Batini *et al.*, (2009)
7	DQ dimension definitions	"The accuracy dimension is the most straightforward and is merely the difference between the correct value and that actually used." (Ballou and Pazer, 1985)
		"Completeness: the extent to which information is not missing and is of sufficient breadth and depth for the task at hand" (Kahn *et al.*, 2002)

4.1. DQC focus

The first type of difference across DQCs concerns their focus. As described in the previous section, some DQCs have a general focus, while others focus on specific applications. With regard to the DQCs with a general focus, these most often do not discuss the limitations of their applications. This could be interpreted as these suggest that they are generally applicable, thus, to some extent, arguing against the need for special focus DQCs.

With regard to the DQCs focusing on particular applications, the problem is not as much that DQCs are unclear about their focus, as a lack of argumentation for why this particular focus requires a special set of DQ dimensions. While it for some of the DQCs seems reasonable that they would involve a special set of DQ dimensions (e.g., "DQ of conceptual views" by Levitin and Redman, 1995), for some other DQCs, it is not as obvious. One example is the ERP (enterprise resource planning) system focused DQCs, which lack argumentation for why such systems require a special set of DQ dimensions – for example, as compared to CRM (customer relationship management) systems. By not demanding justification for why a certain focus requires a special set of DQ dimensions, the door becomes open to a high number of different DQCs, which may have limited value.

4.2. DQ evaluation perspective

The second type of difference across DQCs concerns the DQ evaluation perspective, i.e., the evaluation of DQ through objective or subjective measures. In this context, an often-applied understanding of DQ is its "fitness for use" (Batini and Scannapieco, 2006; Wang and Strong, 1996; Wang *et al.*, 2002). Specifically, this understanding defines DQ as to how well the data serves a certain purpose in operations, decision-making, or planning. Another perspective on DQ concerns the agreement between the data and the real world, i.e., "data-reality–correspondence" (Heinrich *et al.*, 2009; Orr, 1998; Wand and Wang, 1996). Such DQ is sometimes referred to as "intrinsic DQ", which is defined as quality aspects that can be evaluated without any additional data besides the data set, i.e., a comparison between data and the real-world aspect it describes (Bovee *et al.*, 2003; Haug *et al.*, 2009; Mocnik *et al.*, 2018; Wand and Wang, 1996). "Intrinsic DQ" can be distinguished from "extrinsic DQ", which concerns not only aspects that relate to the correspondence between data and the real world, but also aspects that can be viewed as extrinsic to this (e.g., user perceptions) (Bovee *et al.*, 2003).

Several of the DQCs identified include distinctions similar to the discussed intrinsic-extrinsic distinction. These include English (1999), who distinguishes between "inherent" DQ ("how accurately

data represents the real world facts") and "pragmatic" DQ ("how well information enables knowledge workers (the information customers) to accomplish business objectives effectively"); Watts *et al.* (2009), who distinguish between "objective measurements of quality" (derived purely from the data set itself) and "contextual assessments" (quality assessments moderated by the characteristics of the decision-maker and characteristics of the decision task); and Piro *et al.* (2014, p. 34), who distinguish between "hard dimensions" (those that can be measured objectively) and "soft dimensions" (those that can only be assessed using subjective evaluation).

An example of relatively objectively measurable data could be, for example, the "correctness" (Wand and Wang, 1996) of a registered number of a certain item type at a certain stock location. On the other hand, some DQ dimensions cannot be as objectively evaluated, such as "relevance", "understandability", and "applicability" (Eppler and Helfert, 2004; Lee *et al.*, 2009; Wang and Strong, 1996). Thus, while extrinsic dimensions typically imply a need for stakeholder involvement, intrinsic dimensions do not. This, however, does not mean that evaluations of intrinsic dimensions are unproblematic. In order for the intrinsic DQ dimension to make sense, there is a need for agreement on how measures of DQ are evaluated. In other words, there is a need to define the requirements with regard to the required digits, error tolerances, or similar. If there are no predefined agreements as to measurement requirements for some data, evaluations could very well depend on the context.

The distinction between intrinsic end extrinsic DQ explains some of the misunderstandings with regard to the meanings of DQ dimensions. Specifically, dimensions such as "completeness" and "meaningfulness" have been presented as both intrinsic and extrinsic qualities (compare, for example, Wand and Wang, 1996; Wang and Strong, 1996; Bovee *et al.*, 2003; Nelson *et al.*, 2005). In fact, they can be both, but with very different meanings depending on which of the DQ qualities they concern. While "intrinsic completeness" would make sense, for example, with regard to the number of missing addresses for registered customers, "extrinsic completeness" makes sense with regard to, for example, individual users' perception of the amount of detail in which some data explains a certain phenomenon or task. Similarly, while "intrinsic meaningfulness" concerns registered objects that do not correspond to the real world (Wand and Wang, 1996), "extrinsic meaningfulness" concerns the question of whether or not the user understands the data (Bovee *et al.*, 2003). Another example is "believability" and "reputation", which Wang and Strong (1996) define as intrinsic dimensions, while other authors argue that these dimensions concern subjective evaluation (e.g., Naumann and Rolker, 2000; Ge and Helfert, 2008, pp. 382-383).

4.3. DQ category set

The third type of difference across DQCs concerns the included DQ categories. In this context, most DQCs agree on a category that corresponds to the intrinsic category, but as previously mentioned, sometimes using terms such as "hard", "objective", "inherent", "soundness" or "integrity" (English, 1999; Kahn *et al.*, 2002; Piro *et al.*, 2014; Wand and Wang, 1996; Watts *et al.*, 2009). On the other hand, with regard to the extrinsic categories, there is great variance. Some only operate with two overall categories, as well as a mixture thereof. This includes ISO/IEC (2015), who makes a distinction between "inherent DQ" (referring to data itself) and "system-dependent DQ" (depending on the technological domain in which data is used) and has a category for dimensions placed in both. Another example is Loshin (2011), who describes a "qualitative" DQ category, which is a combination of intrinsic and contextual DQ categories.

A more common approach than a bipartite division of DQ categories is to define a set of categories that involves different extrinsic qualities – as is the case with the DQC of Wang and Strong (1996), who operate with four categories: intrinsic, contextual, representational, and accessibility. The argument for

this sort of categorization as compared to the narrower two-tier version is that extrinsic DQ involves dimensions of a very different nature, relating to different types of processes. The logic of including these four categories, as described by Wang and Strong (1996), is illustrated in Figure 2. Specifically, firstly the user must be able to access the data. Secondly, the user must be able to interpret the data. Subsequently, the user can determine the relevance of the data. If it is relevant, the intrinsic quality can be estimated.

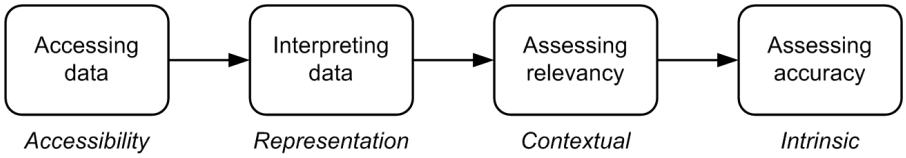

Figure 2. The four-tier division of DQ categories.

Similar to the review by Jayawardene *et al.* (2015), the present study observed that many DQCs are influenced by the one by Wang and Strong (1996), to which they make some modifications of included DQ dimensions or make smaller DQ category adjustments. In this context, the set of DQ categories included in a DQC, to a large extent, depends on which of the five identified approaches for identifying DQ categories is applied (as described in Section 3).

4.4. DQ dimension set

The fourth type of difference across DQCs concerns the selection of DQ dimensions included in the DQCs. Specifically, while almost all the DQCs can agree upon including dimensions that concern accuracy, completeness, timeliness, and usefulness, other types of dimensions, e.g., reputation, consistency, and traceability, are only mentioned by the minority. As for DQ categories, the selection or development of DQ dimensions, to a large extent, depend on which of the five approaches for identifying DQ categories (as described in Section 3) is applied. In this context, some DQCs only focus on a subset of the most common DQ dimension, which is why these DQCs do not, in principle, exclude the non-mentioned dimensions. However, in many cases, DQCs also exclude commonly applied DQ dimensions without any argumentation or reservations.

The opposite situation occurs with DQCs that include a high number of dimensions. Specifically, while some DQCs include fewer than ten dimensions (e.g., Cai and Zhu, 2015), some include more than 30 dimensions (e.g., Jayawardene *et al.*, 2015). Although focusing on different application contexts may, to some extent, explain some differences across DQCs, this is not the full explanation. Rather, there is a need for more critical evaluations of new DQCs to determine if their DQ dimensions are adequately comprehensive for capturing relevant characteristics of DQ, and if they include adequately relevant DQ dimensions only.

4.5. Data level focus

The fifth type of difference across DQCs concerns the level at which data is evaluated (i.e., granularity levels). In this context, different ways of distinguishing and naming data level dimensions can be found in the literature (e.g., Even and Shankaranarayanan, 2005; Gürdür *et al.*, 2019; Pipino *et al.*, 2002; Zaveri *et al.*, 2012). On this basis, a relatively detailed distinction between data levels could include:

- Data item (i.e., data or value) (e.g., item price for a particular product)
- Data record (i.e., tuple or row) (e.g., data in a particular order line)
- Data field (i.e., attribute or column) (e.g., customer addresses)
- Data set (e.g., tables/relations or views) (e.g., customer data)

- Database (e.g., ERP system database)
- Database collection (e.g., ERP and CRM system databases)

The majority of the DQCs identified are not clear on this matter. This is, however, a relevant aspect to clarify, as different evaluation perspectives can imply different meanings of DQ dimensions (Pipino *et al.*, 2002) and influence their relevance. Specifically, DQ dimensions such as "redundancy" or "duplicates" make little sense at the data item level, while they are relevant at the data field and higher abstraction levels. DQ dimensions can also change their meaning, depending on whether the focus is on the data item level or on higher levels. An example is the dimension "believability", which at the data item level would concern the credibility of the particular data, while at higher abstraction levels, it would also involve the question of whether the data items in the data set support or contradict each other.

4.6. Data structure focus

The sixth type of difference across DQCs concerns the structure of the data in focus, i.e., the distinction between structured, semi-structured, and unstructured data. Structured data typically refers to particular characteristics of entities and events, while with unstructured data, there is not as strict a mapping between the data and what it describes. In this context, most of the DQCs identified are not explicit about their focus in this regard.

Clarity about data structure focus is, however, important, as DQ dimensions can change relevance and meaning depending on whether the focus is on structured or unstructured data. Specifically, dimensions such as "conciseness", "compactness", and "simplicity" have little relevance in the context of structured data, as the format for this type of content is typically predefined (e.g., customer number, customer group, customer name, etc.). If applying these dimensions to evaluate structured data, their meaning would relate to the particular combination of fields rather than to their contents.

4.7. DQ dimension definitions

The seventh type of difference across DQCs concerns the definitions of individual DQ dimensions. Specifically, the dimensions included in DQCs are sometimes used in different meanings, if defined at all. As argued earlier, different focuses with regard to data level, data structure, and data evaluation principles can produce confusion with regard to meanings of DQ dimensions. However, even with similar perspectives in this regard, different meanings are still attributed to DQ dimensions. One example of the differences in definitions attributed to DQ dimensions is "timeliness". According to Klein and Lehner (2009), timeliness has been used in two different types of meanings, namely with regard to the actual age of the data (e.g., Naumann, 2002), and with regard to the suitability of the age of data in relation to a use context (e.g., Wang and Strong, 1996). As “timeliness” is typically used in the latter meaning, the former would be better described by the term “age”.

Another example of different definitions being attributed to a dimension is “accuracy”. Specifically, Naumann (2002) defines accuracy as “the percentage of data without errors”; Ballou and Pazer (1985) define accuracy as “the difference between the correct value and that actually used”; and ISO/IEC (2008) defines accuracy as “the degree to which data has attributes that correctly represent the true value of the intended attribute of a concept or event in a specific context of use”. In this context, there are two potential issues. First, the definitions refer to different data levels, i.e., Ballou and Pazer (1985) and ISO/IEC (2008) refer to the data item level, while Naumann (2002) refers to data field or higher levels. Second, there is also a potential problem in that the first two definitions prescribe a certain way of measuring the accuracy in their definition (as opposed to just focusing on what it means), which might not be appropriate in all cases.

Even if the problem concerning clarity with regard to the meaning of individual dimensions is solved, problems still persist with regard to overlaps. Specifically, across DQCs, there are several dimensions that overlap in terms of their meaning. One such example of such overlaps is "correctness" and "accuracy". In this context, the latter is linguistically broader in the sense that something may be inaccurate without being deemed to be incorrect. A similar example concerns dimensions such as "interpretability", "clearness", "readability", and "comprehensibility", whose meanings overlap but which do not refer to exactly the same phenomenon.

Finally, some of the DQCs identified apply DQ dimensions with limited or no explanation of the meaning in which they are used. Although it may appear to be relatively obvious what a DQ dimension refers to (e.g., accuracy, timeliness, or completeness), as shown by the discussion above, such DQ dimensions have, in fact, in many cases been attributed different meanings in the literature.

5. Guidelines for future data quality research

To justify the development of a new DQC, it needs to offer something (such as a particular focus) that is not provided by the existing literature, as it otherwise would seem pointless to develop it. However, DQC contributors rarely discuss or test how their DQCs offer better accounts of DQ for the context in focus than existing ones, but rather seem to be driven by an "if it works" standard. The problem with this approach is that limited justification for the introduction of new DQCs is provided, as it seems that almost DQC would fulfill this criterion. Specifically, applying any random DQC is likely to provide some level of insight into the DQ of the situation in focus. On the other hand, it may be too extensive a demand to suggest that the criterion for proposing a new DQC is tests that demonstrate its superiority over existing DQCs pertaining to the context in focus. A more realistic path could, therefore, be to employ a basic set of standards to provide justification for the introduction of new DQCs. In this context, the identified types of differences across DQCs may be converted into seven guidelines for the development or consolidation of DQCs:

1) DQC focus: clarify the areas of application of the DQC, hereunder the limitations for its application
2) DQ dimensions: select a set of clearly distinguishable DQ dimensions that cover the relevant DQ characteristics in focus
3) DQ categories: select a set of clearly distinguishable categories that cover the DQ themes in focus
4) DQ evaluation perspectives: clarify which type of DQ evaluations the framework focuses on, i.e., objective and/or subjective evaluations
5) Data level focus: clarify the data level focus(es) for each DQ dimension, i.e., clarification of at which data levels the DQ dimension should be applied
6) Data structure focus: clarify the data structure focus for each DQ dimension, i.e., clarification of the DQ dimension focuses on structured, semi-structured, and/or unstructured data
7) DQ dimension definitions: provide clear definitions of the included DQ dimensions while considering the relationships with existing definitions

In relation to the first guideline, some DQCs are explicitly aimed at specific contexts (e.g., ERP or CRM systems), while other DQCs are not clear about their focus (as discussed in Section 4.1). To ensure that researchers and practitioners will apply a DQC as intended, there is a need to clarify the contexts in which it should be applied. This may be done in terms of IT systems, business processes, company types, or similar.

With regard to the second guideline, DQC creators should strive to include DQ dimensions that are clearly distinguishable and cover all the relevant DQ characteristics in relation to the DQC's focus. To achieve this, two types of comparative analyses may be carried out. The first analysis concerns the

identification of potential conceptual overlaps between the DQ dimensions included in DQC, and if this is the case, adjusting of these. The second analysis concerns ensuring that all relevant DQ characteristics are included in the DQC. This can be done by comparing the developed DQC to existing DQCs, which may lead to the identification of overlooked aspects. In this context, the overview of DQCs provided by this article (Table 1 and 2) may be used. The same two kinds of analyses can be conducted for the identified DQ categories to ensure their distinguishability and ability to include all relevant DQ themes in relation to the DQC's focus (guideline 3). In relation to these analyses, Wand and Weber's (1993; 1995) work (which builds on Bunge's (1977) ontological model) on analyzing and evaluating modeling constructs may be consulted, as well as more general literature on classification and categorization principles (see Jacob, 2004).

Next, guidelines 4, 5, and 6 concern that the DQC creator clarifies each included DQ dimension's evaluation perspective, data level focus, and data structure focus. In this context, the distinctions described in sections 4.2, 4.5, and 4.6 may be applied. Finally, the seventh guideline concerns that DQC creators should be clear about the meanings of the included DQ dimensions. As shown by this article, the existing literature has attributed different meanings to certain DQ dimensions, for which reason only stating the name of a DQ dimension may lead to misunderstandings. Besides providing clear definitions of DQ dimensions, DQC creators should also consider the meanings previously attributed to the DQ dimension to avoid adding additional meanings to DQ dimensions and so that readers may intuitively understand the dimensions' meaning.

The seven guidelines described above aim to promote that DQ research converges around fewer DQCs so that it, to a larger extend, becomes possible to compare and build on existing research. This should, however, not be understood as a goal of ending up with just one universal DQC, as different DQCs may have value in different contexts. Nevertheless, there seem to be DQ dimensions that are relevant independent of the focus area, for example, “accuracy”, “completeness”, and “timeliness”. Thus, a distinction may be employed between general DQ dimensions, which are relevant in almost all contexts, and context-specific dimensions, which are only relevant in certain contexts. By clarifying which of these two categories their DQ dimensions belong to, special focus DQCs would improve the clarity regarding the unique characteristics of their focus and, to a greater extent, allow for (at least partially) cross-study comparisons.

A central issue, as reflected by several of the identified differences across DQCs, is that existing literature often only to a limited extent is consulted. However, research needs to "acknowledge the stream of logic on which they are drawing and to which they are contributing" (Sutton and Staw, 1995: 372), as well as describe the logical connections between the proposed and existent constructs (Bacharach, 1989). This lack of drawing on exiting DQ literature may be a consequence of the extensiveness of this literature. Nevertheless, aiming for such justification could be a way of avoiding the development of new DQCs with limited value. Specifically, authors developing new DQCs need to account for the shortcomings of exiting DQCs to justify the development of a new DQC, as well as the value added by using the proposed DQC.

6. Discussion and conclusions

As shown by the literature review of this article, the DQ dimensions included in DQCs vary greatly. Specifically, this article identified 110 unique DQCs in journal and conference articles. These include more than 300 uniquely named DQ dimensions (an exact count would be debatable, as different terms sometimes refer to similar DQ dimensions, and some DQ names are used in different meanings). If the search were to be extended to additional academic outlets and included books, reports, and other kinds of practitioner literature, it seems likely that many more DQCs would be found. Previous literature has

explained the differences across DQCs by arguing that data can represent various aspects of real-world phenomena (Batini and Scannapieco, 2006). However, as shown by the literature review of this article, there is great variance even across DQCs with a similar focus. Seven types of differences across DQCs were identified, which indicates that a lack of methodological rigor is a more fundamental explanation. To address this issue, seven guidelines for future development and consolidation of DQCs were proposed.

6.1. Implications for future research

The literature review of this article identified several articles that analyze DQ dimensions, from the early work in the mid-1990s (Wand and Wang, 1996; Wang and Strong, 1996) to more contemporary literature (e.g., Jayawardene *et al.*, 2015; Liu *et al.*, 2020; Rasool and Warraich, 2018; Teh *et al.*, 2020; Weiskopf and Weng, 2013). On the other hand, the review did not identify articles, which, based on systematic searches, identify, and analyze DQCs. Thus, this article aimed to provide insights on this topic to advance information systems research with a focus on DQ, which was done by identifying variance types across DQCs and defining seven guidelines for future DQ research.

Theory development is a central goal of information systems research, and past research has, from an overall perspective, been relatively successful in advancing theories within different subject domains (Straub, 2012; Weber, 2003). As the information systems field has matured, the theoretical focus on explanation and prediction (Gregor, 2006; Gregor and Klein, 2014), to an increasing degree, moves towards theory extension (Grover and Lyytinen, 2015). This involves a focus on enriching explanations and enhancing predictions through research that challenges the assumptions of existing theories in significant ways (Alvesson and Sandberg, 2011, p. 247). A lack of coherence within a research area, however, hinders such a process in unfolding. Specifically, the continuous production of new DQCs, without justifying their contribution, has resulted in a plethora of different definitions of DQ, making it difficult to compare studies and build on existing research. This may, in fact, be a major barrier to the progression of research in this field.

By identifying seven types of differences across DQCs and formulating seven guidelines for DQC development and consolidation, this article contributes to three forms of future DQ research: (1) research investigating DQ while building on existing DQCs, (2) research developing new DQCs for particular contexts, and (3) research consolidating existing DQCs within certain contexts.

6.2. Implications for practice

Although this article has its main focus on advancing DQ research, it is not without value for practice. Specifically, managers may consider the identified seven types of variance across DQCs to reduce potential issues in DQ improvement projects. This includes ensuring clear definitions of which data (and systems) are in focus; to which extent DQ assessments should include subjective DQ evaluations; which DQ categories should be evaluated; which DQ dimensions should be evaluated; the data levels at which analyses are carried out; to which extent analyses should include semi-structured and unstructured data; and how such DQ dimensions are defined.

6.3. Limitations and future research

The present study has two potential limitations, which concern the search and analysis strategies. In relation to the search strategy, as mentioned previously, a more extensive literature study is likely to have identified additional DQCs. On the other hand, it seems unlikely that the applied search approach did not identify the most influential DQCs. Thus, an extended search seems to be of limited value. With regard to the analysis strategy, the well-established approach of using open and axial coding was employed (Strauss, 1998; Flick, 2009). However, in spite of the rigor of this approach, it involves a subjective

element with regard to determining which text elements should be included and how these elements are converted into categories. This, however, does not take away the validity of the categories identified. In other words, another study with the same focus and method would likely identify similar issues, albeit naming and organizing them differently.

Future research may utilize the guidelines offered by this article when developing new DQCs aimed at special contexts to ensure the quality of these. Furthermore, future research may work on establishing more common understandings of DQ by clarifying definitions of DQ dimensions and consolidating DQCs within different contexts. By identifying seven types of differences across DQCs and developing seven guidelines for future research, this article offers a foundation for this work.

References

Alvesson, M. and Sandberg, J. (2011), "Generating research questions through problematization", *Academy of Management Review,* Vol. 36, No. 2, pp. 247-271.

Arazy, O., Kopak, R. and Hadar, I. (2017), "Heuristic Principles and Differential Judgments in the Assessment of Information Quality", *Journal of the Association for Information Systems*, Vol. 18, No. 5, pp. 403-432.

Asghari, M., Sierra-Sosa, D. and Elmaghraby, A.S. (2020), "A topic modeling framework for spatio-temporal information management", *Information Processing & Management*, Advance online publication. doi.org/10.1016/j.cmpb.2019.05.017

Bacharach, S.B. (1989), "Organizational theories: Some criteria for evaluation", *Academy of Management Journal,* Vol. 14, No. 4, pp. 496–515.

Ballou, D. and Pazer, H. (1985), "Modeling data and process quality in multi-input, multi-output information systems", *Management Science,* Vol. 31, No. 2, pp. 150-162.

Batini, C., Cappiello, C. Francalanci, C. and Maurino, A. (2009). Methodologies for data quality assessment and improvement, *ACM Computing Surveys*, Vol. 41, No. 3, pp. 1-52.

Batini, C. and Scannapieco, M. (2006), *Data Quality: Concepts, Methodologies and Techniques*, Springer, Basel, Switzerland.

Batini, C., Palmonari, M. and Viscusi, G. (2014), "Opening the closed world: A survey of information quality research in the wild", Floridi, L. and Illari, P. (Eds), *The Philosophy of Information Quality,* Springer, Cham, Switzerland, pp. 43–73.

Bovee, M., Srivastava, R.P. and Mak, B. (2003), "A conceptual framework and belief-function approach to assessing overall information quality", *International Journal of Intelligent Systems*, Vol. 18, No. 1, pp. 51–74.

Bunge, M. (1977), *Treatise on Basic Philosophy: Vol. 3: Ontology I: The Furniture of the World*. Reidel, Boston, MA.

Cai, L. and Zhu, Y. (2015), "The challenges of data quality and data quality assessment in the big data era", *Data Science Journal*, Vol. 14, No. 2, pp. 1-10.

Chae, M., Kim, J., Kim, H. and Ryu, H. (2002), "Information quality for mobile internet services: A theoretical model with empirical validation", *Electronic Markets*, Vol. 12, No. 1, pp. 38–46.

English, L.P. (1999), *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits,* Wiley Computer Publishing, New York, NY.

Eppler, M. and Helfert, M. (2004), "A classification and analysis of data quality costs", Chengular-Smith, S., Raschid, L., Long, J. and Seko, C (Eds.), *Proceedings of the 9th International Conference on Information Quality* (ICIQ), MIT IQ Publishing, Cambridge, MA, pp. 311-325.

Eppler, M.J. (2006), *Managing Information Quality*, Springer, Belin/Heidelberg, Germany.

Even, A. and Shankaranarayanan, G. (2005), "Value-driven data quality assessment", Naumann, F., Gertz, M. and Madnick, S.E. (Eds.), *Proceedings of the 2005 International Conference on Information Quality* (ICIQ), MIT IQ Publishing, Cambridge, MA, pp. 265–279.

Experian (2016), *The 2016 Global Data Management Research Benchmark Report,* Experian, London, UK.

Experian. (2019), *Benchmark Report: 2019 Global Data Management Research: Taking Control in the Digital Age*, Experian, London, UK.

Ferretti, E., Cagnina, L., Paiz, V., Donne, S.D:, Zacagnini, R. and Errecalde, M. (2018), "Quality flaw prediction in Spanish Wikipedia: A case of study with verifiability flaws", *Information Processing & Management*, Vol. 54, No. 6, pp. 1169-1181.

Flick, O. (2009), *An Introduction to Qualitative Research* (4th ed.), London, UK: Sage.

Floridi, L. (2011), *The Philosophy of Information*, Oxford University Press, Oxford, UK.

Forsgren, N., Durcikova, A., Clay, P.F. and Wang, X. (2016), "The integrated user satisfaction model: Assessing information quality and system quality as second-order constructs in system administration", *Communications of the Association for Information Systems,* Vol. 38, No. 1, pp. 803-839.

Ge, M. and Helfert, M. (2008), "Data and information quality assessment in information manufacturing systems", Abramowicz, W. and Fensel, D. (Eds.), *Business Information Systems Proceedings of the 11th International Conference,* Springer, Berlin, Germany, pp. 380-389.

Ghasemaghaei, M., Ebrahimi, S. and Hassanein, K. (2018), "Data analytics competency for improving firm decision making performance", *The Journal of Strategic Information Systems,* Vol. 27, No. 1, pp. 101-113.

Gorla, N., Somers, T.M. and Wong, B. (2010), "Organizational impact of system quality, information quality, and service quality", *The Journal of Strategic Information Systems*, Vol. 19, No. 3, pp. 207-228.

Gregor, S. (2006), "The nature of theory in information systems", *MIS Quarterly*, Vol. 30, No. 3, pp. 611-642.

Gregor, S. and Klein, G. (2014), "Eight Obstacles to Overcome in the Theory Testing Genre", *Journal of the Association for Information Systems*, Vol. 15, No. 11, pp. i-xix.

Grover, V. and Lyytinen, K. (2015), "New state of play in information systems research: The push to the edges", *MIS Quarterly,* Vol. 39, No. 2, pp. 271-296.

Gürdür, D., El-Khoury, J. and Nyberg, M. (2019), "Methodology for linked enterprise data quality assessment through information visualizations", *Journal of Industrial Information Integration*, Vol. 15, No. Sep., pp. 191–200.

Haug, A., Arlbjørn, J.S. and Pedersen, A. (2009), "A classification model of ERP system data quality", *Industrial Management & Data Systems,* Vol. 109, No. 8, pp. 1053-1068.

Heinrich, B., Klier, M. and Kaiser, M. (2009), "A procedure to develop metrics for currency and its application in CRM", *Journal of Data and Information Quality*, Vol. 1, No. 1, pp. 1-28.

Heinrich, B., Hristova, D., Klier, M., Schiller, A. and Szubartowic, M. (2018), "Requirements for data quality metrics", *Journal of Data and Information Quality*, Vol. 9, No. 2, pp. 1-29.

Huang, H. (2018), "Big data to knowledge: Harnessing semiotic relationships of data quality and skills in genome curation work", *Journal of Information Science*, Vol. 44, No. 6, pp. 785–801.

IBM (2016), *Extracting Business Value from the 4 V's of Big Data*. IBM Big Data and Analytics Hub. Retrieved Oct. 7, 2019, from http://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data.

ISO/IEC (2008). ISO/IEC 25012: *Software Engineering — Software Product Quality Requirements and Evaluation (SQuaRE) — Data Quality Model*, International Organization for Standardization, Geneva.

ISO/IEC (2015), *ISO/IEC 25024: 2015: Systems and Software Engineering: Systems and Software Quality Requirements and Evaluation (SQuaRE) – Measurement of Data Quality,* International Organization for Standardization, Geneva, Switzerland.

Jacob, E.K. (2004), ""Classification and Categorization: A Difference that Makes a Difference", *Library Trends*, Vol. 52, No. 3, pp. 515-540.

Jarke, M., Jeusfeld, M.A., Quix, C. and Vassiliadis, P. (1999), "Architecture and quality in data warehouses: An extended repository approach", *Information Systems*, Vol. 24, No. 3, pp. 229-253.

Jayawardene, V., Sadiq, S. and Indulska, M. (2015), *An analysis of data quality dimensions* (ITEE Technical Report No. 2015-02), University of Queensland, Brisbane, Australia.

Jayawickrama, U., Liu, S. and Smith, M.H. (2017), "Knowledge prioritisation for ERP implementation success: Perspectives of clients and implementation partners in UK industries", *Industrial Management & Data Systems,* Vol. 117, No. 7, pp. 1521-1546.

Kahn, B.K., Strong, D.M. and Wang, R.Y. (2002), "Information quality benchmarks: Product and service performance", *Communications of the ACM,* Vol. 45, No. 4, 184-192.

Katerattanakul, P. and Siau, K. (1999), "Measuring information quality of web sites: Development of instrument", De, P. and De Gross, J.I. (Eds.), *Proceedings of the 20th International Conference on Information Systems (ICIS),* Association for Information Systems, Atlanta, GA, pp. 279-285.

Klein, A. and Lehner, W. (2009), "Representing data quality in sensor data streaming environments", *Journal of Data and Information Quality*, Vol. 1, No. 2, pp. 1–28.

KPMG (2017), *Disrupt and Grow: 2017 Global CEO Outlook*. Retrieved October 7, 2019, from: https://assets.kpmg.com/content/dam/kpmg/xx/pdf/2017/06/2017-global-ceo-outlook.pdf

Laumer, S., Maier, C. and Weitzel, T. (2017), "Information quality, user satisfaction, and the manifestation of workarounds: a qualitative and quantitative study of enterprise content management system users", *European Journal of Information Systems*, Vol. 26, No. 4, pp. 333-360.

Lee, Y.W., Strong, D.M., Kahn, B.K. and Wang, R.Y. (2002), "AIMQ: A methodology for information quality assessment", *Information Management*, Vol. 40, No. 2, pp. 133–146.

Lee, Y.W., Pipino, L.L., Funk, J.D. and Wang, R.Y. (2009), *Journey to Data Quality*. MIT Press, Cambridge, MA.

Leidner, D.E. and Kayworth, T. (2006), "A review of culture in information systems research: Toward a theory of information technology culture conflict", *Management Information Systems Quarterly*, Vol. 30, No. 2, pp. 357-399.

Leon, S. (2018), "Service mobile apps: A millennial generation perspective", *Industrial Management & Data Systems,* Vol. 118, No. 9, pp. 1837-1860.

Levitin, A. and Redman, T. (1995), "Quality dimensions of a conceptual view", *Information Processing & Management*, Vol. 31, No. 1, pp. 81-88.

Li, X.-B. and Qin, J. (2017), "Anonymizing and Sharing Medical Text Records", *Information Systems Research,* Vol. 28, No. 2, pp. 332–352.

Liaw, S.T., Rahimi, A., Ray, P., *et al.* (2013), "Towards an ontology for data quality in integrated chronic disease management: A realist review of the literature", *International Journal of Medical Informatics*, Vol. 82, No. 1, pp. 10–24.

Liu, C., Nitschke, P., Williams, S.P. and Zowghi, D. (2020), "Data quality and the Internet of Things", *Computing,* Vol. 102, No. 2, pp. 573–599.

Loshin, D. (2001), *Enterprise Knowledge Management; The Data Quality Approach.* Morgan Kauffman, San Diego, CA.

Loshin, D. (2011), *The Practitioner's Guide to Data Quality Improvement*. Morgan Kaufman, Burlington, MA.

Lóscio, B.F., Batista, M.C.M. and Souza, D. (2013), "Using information quality for the identification of relevant web data sources: A proposal", Taniar, D., Pardede, E., Steinbauer, M. and Khalil, I. (Eds.), *Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services*, ACM, New York, NY, pp. 36–44.

Miller, H. (1996), "The multiple dimensions of information quality", *Information Systems Management*, Vol. 13, No. 2, pp. 79-83.

Mocnik, F.-B., Mobasheri, A., Griesbaum, L., Eckle, M., Jacobs, C. and Klonner, C. (2018), "A grounding-based ontology of data quality measures", *Journal of Spatial Information Science*, Vol. 16, No. 1, pp. 1–25.

Moravec, P., Minas, R. & Dennis, A.R. (2019), "Fake news on social media: People believe what they want to believe when it makes no sense at all", *MIS Quarterly,* Vol. 43, No. 4, pp. 1343-1360.

Naumann, F. and Rolker, C. (2000), "Assessment methods for information quality criteria", Klein, B.D. and Rossin, D.F. (Eds.), *Proceedings of the 5th International Conference on Information Quality (IQ'00)*, MIT, Cambridge, MA, pp. 148-162.

Naumann, F., (2002), *Quality-Driven Query Answering for Integrated Information Systems*. Springer, Berlin/Heidelberg, Germany.

Nelson, R.R., Todd, P.A. and Wixom, H. (2005), "Antecedents of information and system quality: An empirical examination within the context of data warehousing", *Journal of Management Information Systems*, Vol. 21, No. 4, pp. 199-235.

O'Brien, T. (2015), "'Accounting' for data quality in enterprise systems", *Procedia Computer Science,* Vol. 64, No. Dec., pp. 442–449.

Orr, K. (1998), “Data quality and systems theory”, *Communications of the ACM,* Vol. 41, No. 2, pp. 66–71.

Pipino, L.L., Lee, Y.W. and Wang, R.Y. (2002), “Data quality assessment”, *Communications of the ACM*, Vol. 45, No. 4, pp. 211–218.

Piro, A, Rohweder, J.P., Möller, F., Pickert, L. and Klingenberg, C. (2014), “Einleitung [Introduction]”, Piro, A. (ed.), *Informationsqualität Bewerten: Grundlagen, Methoden, Praxisbeispiele*, Symposion Publishing, Düsseldorf, Germany.

Popovič, A., Hackney, R., Tassabehji, R. and Castelli, M. (2018), “The impact of big data analytics on firms’ high value business performance”, *Information Systems Frontiers*, Vol. 20, No. 2, pp. 209-222.

Rajan, N.S., Gouripeddi, R., Mo, P., Madsen, R.K. and Facelli, J.C. (2019), “Towards a content agnostic computable knowledge repository for data quality assessment”, *Computer Methods and Programs in Biomedicine,* Vol. 177, No. Aug., pp. 193–201.

Ram, J., Corkindale, D. and Wu, M.L. (2013), “Examining the role of system quality in ERP projects”, *Industrial Management & Data Systems,* Vol. 113, No. 3, pp. 350-366.

Rasool, T. and Warraich, N.F. (2018), “Does Quality matter: A systematic review of information quality of e-government websites”, Kankanhalli, A. Ojo, A. and Soares, D. (Eds.), *ICEGOV '18: Proceedings of the 11th International Conference on Theory and Practice of Electronic Governance*, Association for Computing Machinery, New York, NY, pp. 433–442.

Redman, T.C. and Blanton, A. (1997), *Data Quality for the Information Age*. Artech House, Norwood, MA.

Redman, T.C. (2001), *Data Quality: The Field Guide*, Digital Press, Boston, MA.

Reinking, J., Arnold, V. and Sutton, S. (2020), Synthesizing enterprise data to strategically align performance: the intentionality of strategy surrogation, *International Journal of Accounting Information Systems*, Vol. 36, Advance online publication.

Rieh, S. (2002), “Judgment of information quality and cognitive authority in the web”, *Journal of the American Society for Information Science and Technology*, Vol. 53, No. 2, pp. 145–161

Savitz, S.T., Savitz, L.A., Fleming, N.S., Shah, N.D. and Go, A.S. (2020), “How much can we trust electronic health record data?”, *Healthcare*, Vol. 8, No. 3, Advance online publication. doi.org/10.1016/j.accinf.2019.100444

Schäffer, T. and Beckmann, H. (2014), *Trendstudie Stammdatenqualität 2013: Erhebung der aktuellen Situation zur Stammdatenqualität in Unternehmen und daraus Abgeleitete,* Steinbeis-Edition, Stuttgart, Germany.

Sebastian-Coleman, L. (2012), *Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework*, Elsevier, Waltham, MA.

Shamala, P., Ahmad, R., Zolait, A. and Sedek, M. (2017), “Integrating information quality dimensions into information security risk management (ISRM)”, *Journal of Information Security and Applications*, Vol. 36, No. C, pp. 1–10.

Sneath, P.H.A. (1973), *Numerical Taxonomy. The Principles and Practice of Numerical Classification,* W.H. Freeman and Company, San Francisco, CA.

Straub, D. (2012), “Editor’s comments: Does MIS have native theories”, *MIS Quarterly,* Vol. 36, No. 2, pp. iii-xii.

Strauss, A. (1998), *Qualitative Analysis for Social Scientists,* Cambridge University Press, New York, NY.

Strong, D.M., Lee, Y.W. and Wang, R.Y. (1997), “Data quality in context”, *Communications of the ACM*, Vol. 40, No. 5, pp. 103-110.

Stvilia, B., Gasser, L., Twidale, M.B. and Smith, L.C. (2007), “A framework for information quality assessment”, *Journal of the American Society for Information Science & Technology*, Vol. 58, No. 12, pp. 1720–1733.

Sutton, R.I. and Staw, B.M. (1995), “What theory is not”, *Administrative Science Quarterly,* Vol. 40, No. 3, pp. 371–384.

Teh, H.Y., Kempa-Liehr, A.W. and Wang, K.K.-I. (2020), “Sensor data quality: a systematic review”, *Journal of Big Data,* Vol. 7, No. 11, Advance online publication. doi.org/10.1186/s40537-020-0285-1

Wand, Y. and Weber, R. (1993), "On the ontological expressiveness of information systems analysis and design grammars", *Journal of Information Systems*, Vol. 3, No. 4, pp. 217–237.

Wand, Y. and Weber, R. (1995), "On the deep structure of information systems", *Journal of Information Systems*, Vol. 5, No. 3, pp. 203–223.

Wand, Y. and Wang, R.Y. (1996), "Anchoring data quality dimensions in ontological foundations", *Communications of the ACM*, Vol. 39, No. 11, pp. 86–95.

Wang, R.Y. and Strong, D.M. (1996), "Beyond accuracy: What data quality means to data consumers", *Journal of Management Information Systems*, Vol. 12, No. 4, pp. 5–33.

Wang, R.Y., Reddy, M.P. and Kon, H.B. (1995a), "Toward quality data: An attribute-based approach", *Decision Support Systems*, Vol. 13, No. 3-4, pp. 349-372.

Wang, R.Y., Storey, V.C. and Firth, C.P. (1995b), "A Framework for analysis of data quality research", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 7, No. 4, pp. 623-640.

Wang, R.Y., Ziad, M, and Lee, Y.W. (2002), *Data Quality,* Kluwer Academic Publishers, New York, NY.

Wang, R.Y. (1998), "A product perspective on total data quality management", *Communications of ACM*, Vol. 41, No. 2, pp. 58-65.

Watts, S., Shankaranarayanan, G. and Even, A. (2009), "Data quality assessment in context: A cognitive perspective", *Decision Support Systems,* Vol. 48, No. 1, pp. 202–211.

Weber, R. (2003), "Editor's comment: Theoretically speaking", *MIS Quarterly,* Vol. 27, No. 3, pp. iii-xii.

Weiskopf, N.G. and Weng, C. (2013), "Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research", *Journal of the American Medical Informatic Association*, Vol. 20, No. 1, pp. 144–151.

Xu, H., Nord, J.H., Brown, N. and Nord, G.D. (2002), "Data quality issues in implementing an ERP", *Industrial Management & Data Systems*, Vol. 102, No. 1, pp. 47-58.

Zaveri, A., Rula, A. Maurino, A., Pietrobon, R. and Lehmann, J. (2012), "Quality assessment for linked data: A survey", *Semantic Web–Interoperability Usability Applicability*, Vol. 7, No. 1, pp. 63–93.

Zhang, R., Indulska, M. and Sadiq, S. (2019), "Discovering data quality problems: The case of repurposed data", *Business Information Systems Engineering*, Vol. 61, No. 5, pp. 575–593.

Zhou, T. (2013), "Understanding continuance usage of mobile sites", *Industrial Management & Data Systems*, Vol. 113, No. 9, pp. 1286-1299.

Zhu, X. and Gauch, S. (2000), "Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web", Yannakoudakis, E.J. Belkin, N.J. Ingwersen, P. and Leong, M.-K. (Eds.), *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, New York, NY, pp. 288-295.