

Article

Not peer-reviewed version

An Overview of Current and New Data Quality Dimensions under a Common Framework

[Russell Miller](#) , [Harvey Whelan](#) , Michael Chrubasik , David Whittaker , Paul Duncan , [João Gregório](#) *

Posted Date: 13 September 2024

doi: 10.20944/preprints202409.1076.v1

Keywords: data quality; data model; data quality dimensions; data traceability; confidence in data; data metrology; data uncertainty; data structures; big data; IoT









Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Article

An Overview of Current and New Data Quality Dimensions under a Common Framework

Russell Miller ¹, Harvey Whelan ^{1,2}, Paul Duncan ¹, Michael Chrubasik ¹,
David Whittaker ¹ and João Gregório ^{1,*}

¹ Informatics, Data Science Department, National Physical Laboratory, Glasgow, United Kingdom

² Department of Natural Sciences, University of Bath, Bath, Somerset, United Kingdom

* Correspondence: joao.gregorio@npl.co.uk

Abstract: This paper presents a comprehensive exploration of Data Quality terminology, revealing a significant lack of standardisation in the field. We propose a novel approach to aggregating disparate Data Quality terms used to describe the multiple facets of Data Quality, under common umbrella terms, with a focus on the ISO 25012 standard. Our aim is to design a Data Quality Data Model that serves as a universally applicable framework for Data Quality assessment. We introduce four additional Data Quality dimensions: Governance, Usefulness, Quantity, and Semantics, enhancing specificity, complementing the framework established by the ISO 25012 standard, and understanding of Data Quality aspects. The ISO 25012 standard, while tailored for software development, offers a foundation for the development of our proposed Data Quality Data Model. This is due to the prevalent nature of software development across a multitude of domains. In contrast, frameworks like ALCOA+ that are specific to certain domains lack the ability to be generalised. The model we propose can be seen as a “Rosetta Stone” for Data Quality terminology, facilitating a seamless communication of Data Quality between different domains when collaboration is required to tackle cross-domain projects or challenges.

Keywords: data quality; data model; data quality dimensions; data traceability; confidence in data; data metrology; data uncertainty; data structures; big data; IoT

1. Introduction

In this digital age, we are experiencing an unparalleled increase in the generation of data [1]. There is an increasing dependence on data for informing decision-making processes and influencing future plans within organisations. The use of data is readily apparent across many different sectors, from pharmaceutical manufacturing and healthcare to engineering and education. This is exemplified by the frequent adoption of data-driven initiatives by businesses aiming to sustain a competitive advantage. In healthcare, for example, the strategic use of personalised and detailed patient data enables professionals to design and administer tailored therapies. In manufacturing industry, sensors generate large volumes of data, assuming a vital role in the monitoring and enhancement of manufacturing processes. As we increase our reliance on data to drive processes and decision-making, we must also be vigilant about its quality, since the results obtained from using data-driven methods are dependent on the quality of the data used.

Lower-quality data, such as measurements captured by a sensor with insufficient precision, lacks the necessary requirements to support processes and can have adverse consequences for organisations. In data-driven operations, it is imperative to evaluate the data to determine its suitability for the intended task. This prompts a clear and fundamental question: what is data quality and what methodologies can be used for evaluating it? As outlined in ISO standard 25012 [2], data quality relates to the extent to which data meets the specifications established by an organisation responsible for developing a product. The standard presents a comprehensive data model, in the context of software development, that facilitates the assessment of data quality by considering characteristics such as correctness, completeness, and consistency.

Such standards make it clear that the concept of data quality is not novel. However, it has gained significant interest and risen to prominence in recent years. Despite this, a notable lack of established

methodologies or standardised definitions of data quality persists. There are variations between different sectors in terms of their emphasis on specific dimensions of data quality — the different aspects that determine how good or bad data is — which can be attributed to unique data needs and exploitation goals of each sector. The presence of discrepancies in the definitions of data quality can create difficulties in comparing and implementing efficient data quality practices across diverse domains. Unambiguous definitions are particularly relevant when tackling issues such as the current climate crisis which need collaboration between domains.

Prior studies on data quality have typically focused on specific sectors, as evidenced by existing literature reviews [3–16]. There has been a growing need in the healthcare industry, namely in electronic health records, to define and evaluate important dimensions of data quality [15,17,18]. Research has also been conducted in the field of Architecture, Engineering, and Construction (AEC) [19], the autonomous vehicles industry [20], and in data analytics for smart factories [1]. Although these studies have made valuable contributions towards enhancing comprehension of data quality in specific areas, they have also reinforced the issue of conflicting data quality definitions and terminology across sectors. This is further aggravated by disparities within a given sector. For instance, an investigation conducted on electronic health records revealed the existence of 44 unique dimensions employed across different pockets of the healthcare sector [15].

In a literature review conducted by Ibrahim *et al.* [13], data quality issues affecting the master data — data related to business entities that provides sufficient context for transactions — of businesses and organisations were explored. The methodology used aimed to identify key factors influencing the quality of master data. Analysis revealed that data governance emerged as the most frequently addressed term within business operations, with 11 out of 15 studies highlighting its significance. This heightened emphasis on data governance can be attributed to increased GDPR compliance needs and digitalisation, which requires defined roles and responsibilities in data quality management. Elements that impact master data quality were compared for three distinct sectors: business, accounting, and healthcare. They identified terminology discrepancies in these elements, which impact the quality of the data. It follows that there is a significant need to conduct a more overarching analysis in order to capture the diversity observed throughout the different sectors.

Certain sectors, such as AEC, have not given due attention to data quality metrics. A literature review by Morwood *et al.* [19] highlights the insufficient consideration of data quality in monitoring the energy performance of buildings. This concern prompted an examination of recent literature through the lens of data quality, revealing that only 9 out of 162 articles explicitly addressed the subject, and among those that did discuss data quality, an average of only 3.23 data quality dimensions were mentioned. While specific data quality issues, such as low spatio-temporal granularity, data loss, and high measurement uncertainty were sporadically addressed, the primary observation underscores the fragmented nature of data quality approaches in building energy monitoring studies.

A study conducted by Mansouri *et al.* [14] aimed to identify the dimensions of data quality relevant for Internet of Things (IoT) applications. A constraint encountered in the study was the limited number of data quality dimensions that were explicitly referenced for the IoT domain. The researchers discovered that a lack of agreement existed on the dimensions that could be applied to various data categories. They proposed the incorporation of additional dimensions, informed by insights from specialists in the IoT domain, as a potential solution to tackle this issue. An additional literature review focused on IoT has also been published, containing a comparative analysis of the relationship between data types and data quality dimensions to refine available options for managing IoT data [21]. Findings indicate that the majority of IoT solutions have the capability to process structured or semi-structured data, whereas their ability to handle unstructured data is limited. The dimensions of data quality that have been identified as particularly significant in this context include completeness, correctness, consistency, and timeliness.

Another work, done by Firmani *et al.* [22] focuses on data quality related to Big Data. The complexity of big data, which comprises large volumes of unstructured data, presents unique challenges in

terms of data accessibility, which is itself an aspect of data quality. Firmani's research provides perspective on the subject and adds to the establishment of methodologies for assessing the quality of large datasets. Big data is also connected with machine learning (ML), by leveraging MLs ability to discern data characteristics and enhance data processing. The effectiveness and efficiency of ML techniques are tied to the datasets, including their volume, used in the training process, further emphasising the importance of data quality as a factor influencing model performance [23].

Big Data and IoT are revolutionising how data is used across different sectors. In business, big data analytics informs decision-making and optimises operations, while in healthcare, IoT devices can be used to analyse patient data for creating personalised therapies [8]. The common denominator is the role of data and its quality. As these technologies progress, challenges surface which highlight the need for standardised data quality frameworks.

These reviews are proof of, despite efforts to establish standardised data quality metrics and assessment procedures being pursued, different industries having unique data needs that must be considered. As a result, it is possible for inconsistencies to emerge while attempting to define data quality. These inconsistencies stem from a lack of standardisation, with most sectors struggling to keep up and varying in their understanding and implementation of data quality to improve their processes and decision-making. This lack of standardisation creates a large diversity in existing terminology used to describe data quality, resulting in further communication challenges. The goal of this paper is to conduct a comparative analysis of data quality terminology across different domains and structure it into a hierarchical data model. This structured glossary has the potential to act as a translation layer, enabling different domains to communicate with each other using a common language framework when addressing the issue of data quality when tackling larger issues such as the aforementioned climate challenge.

2. Methods

In this section, we outline the methodologies used in this study to ensure a comprehensive and rigorous analysis of data quality terminology. Our approach is two-fold: first, we conducted a literature search, described in Section 2.1, to identify relevant academic papers that discuss data quality. Second, to align the terms used in these papers with the data quality dimensions specified by the ISO 25012 standard, we performed a terminology mapping exercise, detailed in Section 2.2. This enabled us to capture the broad spectrum of data quality descriptors, thereby enriching our understanding of the subject. This also allowed us to condense the heterogeneous terminology into a structured, hierarchical vocabulary.

2.1. Literature Search

The main literature search was conducted using an independent multiple-review approach and is highlighted in Figure 1. The search was carried out using Web of Science (WoS) as the designated search engine. The search query included "data quality" as both the search term and the only required keyword. The scope of the search was limited to recently published academic papers between 2019 and 2023. This resulted in a total of 10,855 papers being identified.

Only publications written in English were considered and only the first 500 papers from the total found were reviewed, under the assumption that papers found deeper in the search lost relevancy. This cap of 500 ensured that the number of papers to be individually reviewed was manageable.

Each researcher individually reviewed these 500 papers based on title, abstract, and keywords. Subjective reviewing criteria were used to decide whether the paper was relevant or not, in accordance with the independent multiple-review approach. Papers identified as relevant by at least one researcher proceeded through the remainder of the screening phase for a more in-depth review.

This approach highlighted a total of 173 papers, already taking into account the removal of duplicate papers found by multiple researchers. From this total, 57 papers were inaccessible, being

locked behind closed access. This resulted in a total of 116 relevant papers being captured for further review and discussion.

In addition to the primary literature search methodology, additional publications were discovered through casual conversations during the research process. These are discussed on a case-by-case basis and are not captured by Figure 1.

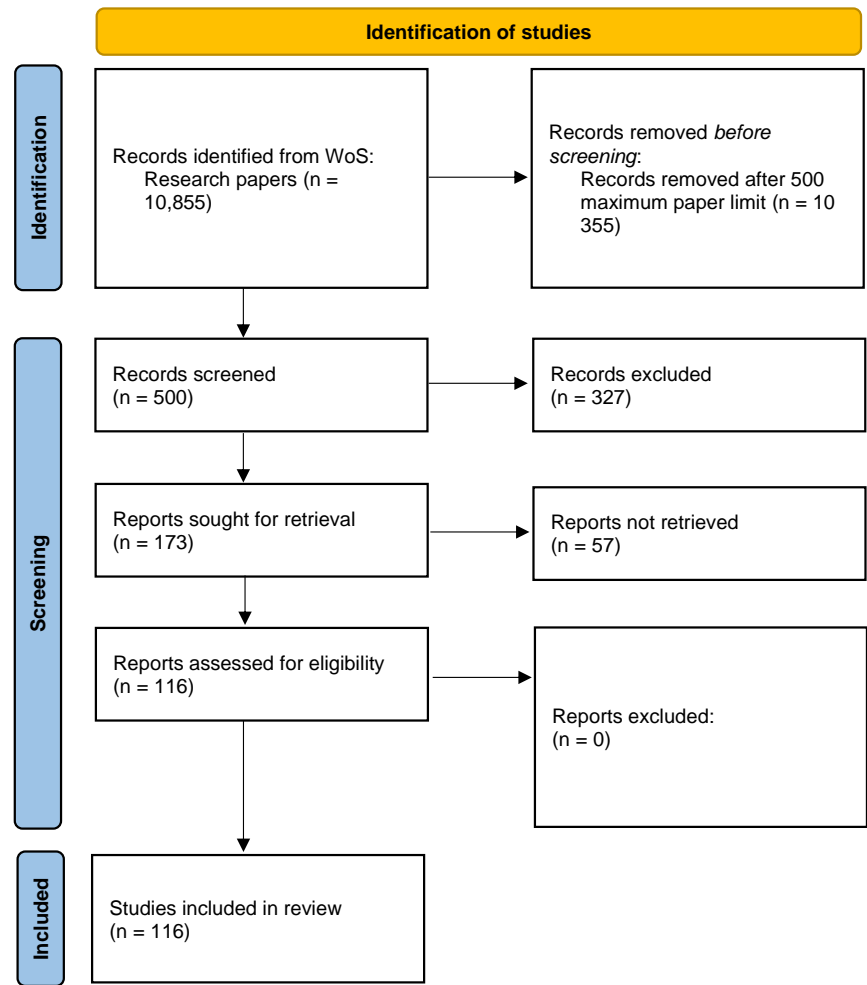


Figure 1. Literature screening process used in this study. Does not account for research papers found outside of the main Web of Science query.

2.2. Terminology Mapping

Each paper included in this study was individually screened for any and all terminology mentioned, described, or used to assess data quality. The terms considered were those that primarily described (or attempted to describe) any of the data quality dimensions specified by the ISO 25012 standard. Terms that did not match the ISO 25012 standard but were used in the context of assessing data quality were also considered.

The outcome was a list of 262 terms used to describe some aspect of data quality or one of its dimensions. The definition of each individual term, as stated in its source material, was used to align the term with one of ISO 25012’s existing dimensions. The list included antonyms, such as the use of “inaccuracy” to describe accurate data. Terms that did not fit into any of the existing dimensions were grouped together based on their similarities to help define new core data quality dimensions, in addition to ISO 25012’s existing terminology.

3. Results

3.1. Data Quality Dimensions

The ISO 25012 standard offers a detailed framework consisting of 15 different dimensions of data quality that determine the features related to the management and improvement of data quality. These dimensions are defined in Table 1 according to their ISO definitions. While this standard was originally developed in the context of software development [2], it has relevance to the field of data science given the close alignment between both practices. They both share a focus on the importance of high-quality, reliable data for successful outcomes, and have cross-applicability over multiple domains.

Table 1. Data quality dimensions as described by ISO 25012. This table presents the dimensions that represent the fundamental properties of data quality.

Dimension	Definition
Accuracy	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.
Completeness	The degree to which subject data associated with an entity has values for all expected attributes and related entity instances in a specific context of use.
Consistency	The degree to which data has attributes that are free from contradiction and are coherent with other data in a specific context of use. It can be either or both among data regarding one entity and across similar data for comparable entities.
Credibility	The degree to which data has attributes that are regarded as true and believable by users in a specific context of use. Credibility includes the concept of authenticity (the truthfulness of origins, attributions, commitments).
Currentness	The degree to which data has attributes that are of the right age in a specific context of use.
Accessibility	The degree to which data can be accessed in a specific context of use, particularly by people who need supporting technology or special configuration because of some disability.
Compliance	The degree to which data has attributes that adhere to standards, conventions or regulations in force and similar rules relating to data quality in a specific context of use.
Confidentiality	The degree to which data has attributes that ensure that it is only accessible and interpretable by authorised users in a specific context of use.
Efficiency	The degree to which data has attributes that can be processed and provide the expected levels of performance by using the appropriate amounts and types of resources in a specific context of use.
Precision	The degree to which data has attributes that are exact or that provide discrimination in a specific context of use.
Traceability	The degree to which data has attributes that provide an audit trail of access to the data and of any changes made to the data in a specific context of use.
Understandability	The degree to which data has attributes that enable it to be read and interpreted by users, and are expressed in appropriate languages, symbols and units in a specific context of use.
Availability	The degree to which data has attributes that enable it to be retrieved by authorised users and/or applications in a specific context of use.
Portability	The degree to which data has attributes that enable it to be installed, replaced or moved from one system to another preserving the existing quality in a specific context of use.
Recoverability	The degree to which data has attributes that enable it to maintain and preserve a specified level of operations and quality, even in the event of failure, in a specific context of use.

While this standard can be interpreted as an all-encompassing framework to assess data quality, its specificity to software development still imposes limitations to its use across different domains. Certain terms could not be associated with existing data quality dimensions. As a result, this work considers the addition of four new dimensions: **governance**, **usefulness**, **quantity**, and **semantics**. These are defined, by this work, in Table 2. The addition of these dimensions expands the ISO 25012 standard into a standardised framework for categorising data quality dimensions more reliability across different domains.

Table 2. Additional Data quality dimensions defined by this work, to complement the ISO 25012 data model.

Dimension	Definition
Governance	The degree to which data has attributes that adhere to the formalised frameworks of authority and accountability that support harmonised data activities across an organisation.
Usefulness	The usefulness of data is determined by the extent to which its attributes meet the specific requirements of users or applications. This includes the data’s adaptability across various contexts, recognising its potential for diverse applicability due to aspects such as reusability and interoperability.
Quantity	The degree to which data has attributes that represent the sufficient amount or volume, providing a comprehensive view of the intended attribute of a concept or event in a specific context of use.
Semantics	The degree to which data accurately and consistently represents the intended meaning, interpretation, and real-world concepts within a specific context of use, ensuring correct semantic understanding by users and applications.

In the process of developing this work, the ALCOA+ framework, a data integrity framework widely adopted in life sciences and endorsed by the US Food and Drug Administration (FDA), was also considered [24,25]. However, compared to ISO 25012, it provides less detail in defining data quality dimensions, which poses challenges for its adaptation into a universal data quality framework. A notable limitation of ALCOA+ is its inability to distinguish between accuracy and precision. While it effectively addresses the data quality needs specific to life sciences, its applicability to other sectors is limited due to its lack of generalisability. This limitation reinforces the rationale for selecting ISO 25012, a software development-oriented data quality framework, as the foundation for the development of the more versatile data quality framework proposed in this study.

3.2. Classification of Data Quality Dimensions

ISO 25012 also assigns each data quality dimension a position in a spectrum that ranges from “Inherent” to “System-dependent”. Inherent dimensions are those that represented the fundamental, intrinsic properties of data that hold true regardless of context or user requirements. Consider a hospital database that records patient information. The accuracy of this data, such as correct names and addresses, is an inherent dimension. If a patient’s name is recorded incorrectly, it could lead to serious errors in patient care.

System-dependent dimensions underscore the role of the system in data quality, and include aspects such as data availability, portability, and recoverability. For example, consider an online banking system. The availability of the system (i.e., the system is up and running when a customer wants to check their account balance) is a system-dependent dimension.

Additionally to this placement, this work also considers dimensions in the middle of this spectrum to be “Contextual”, as they share both inherent and system-dependent characteristics. Contextual dimensions are those that emphasise the importance of the context in which data is used. For instance, in a marketing campaign, customer data needs to be unique, accurate, and consistent across all engagement channels. The relevancy of the data to the specific marketing campaign is a contextual dimension.

Tables 3–5 list all the inherent, contextual, and system-dependent data quality dimensions respectively. These tables include all dimensions from the original ISO 25012 standard as well as the additions proposed by this work. They also list all the 262 associated terms found in the literature search, described in Section 2.1, used to describe each of the core data quality dimensions. Lastly, Figure 2 showcases how the 262 associated terms are organised into a structured glossary for supporting efficient data quality communication across different domains.

Table 3. Inherent Data Quality Dimensions based on ISO 25012. This table presents the dimensions that represent the fundamental, intrinsic properties of data that hold true regardless of the context or user requirements. Each dimension is defined and associated with specific terms from the literature, providing a comprehensive overview of the inherent aspects of data quality. [1,4,7,9,10,12–15,18–22,24–135].

Dimension	Associated Terms
Accuracy	Accuracy, Accurate, Closely match a real-state, Coincidence, Correct, Corrections made, Correctness, Data value out of range, Errors, Free of error, Free of mistakes, Inaccurate, Incorrect, Positive predicted values, Value accuracy
Completeness	Complete, Completeness, Comprehensiveness, Diversity, Entity heterogeneity, Heterogeneity, Incompleteness, Coverage, Min. data capture, Min. sample points, Min. time coverage, Missing values, Representativeness, Study representativeness, Variety, Areas covered, Geographical coverage, Handling of null values, Homogeneity, Missing information, Missingness, Omission, Representativity, Scope, Technological cover, Time-related coverage
Consistency	Coherence, Cohesiveness, Comparability, Consistency, Consistent, Consistent representation, Constant representation, Comparable, Duplication, Incompatibility, Inconsistency, Redundancy, Representational Consistency, Reproducibility, Spatial stability, Structural consistency, Syntactic Accuracy, Thematic accuracy, Variability
Credibility	Agreement, Authenticity, Believability, Bias, Coding Reliability, Confidence, Corroboration, Credibility, Freedom of bias, Impartiality, Incorrect information, Integrity, Misleading, Plausibility, Popularity, Quality, Reliability, Reputability, Reputation, Robustness, Status, Trust, Trustworthiness, Unambiguity, Unbiased, Valid, Validity, Veracity
Currentness	Actuality, Currency, Currentness, Freshness, Outdated Information, Rate of recording, Recency, Timeliness, Timely, Up-to-date, Velocity, Vitality, Volatility, Volatability

Table 4. Contextual Data Quality Dimensions based on ISO 25012. This table presents the dimensions that emphasise the importance of the context in which data is used. Each dimension is defined and associated with terms from the literature, highlighting the multifaceted nature of contextual data quality. The table also introduces two newly added dimensions: Governance and Usefulness. [1,4,7,9,10,12–15,18–22,24–98,100–120,122–142].

Dimension	Associated Terms
Accessibility	Accessibility, Clear definition, Discoverability, Ease of access, Findability
Compliance	Compliance, Concordance, Conformance, Conformity, Licensing, Model conformance, Privacy preservation, Value data type, Appropriate use
Confidentiality	Confidentiality, Data protection, Privacy, Security, Sensitivity, Statistical disclosure control, Vulnerability
Efficiency	Costs effectiveness, Ease of manipulation, Efficiency, Expediency, Minimality, Optimal use of resources, Performance, Viscosity
Governance *	Accountability, Alignment, Auditability, Authority, Authorisation, Enduring, Management, Risks
Precision	Attribute granularity, Brief representation, Concise representation, Conciseness, Detection limit, Distribution bias, Format precision, Imprecise, Intrinsic Approximation, Intrinsic uncertainty, Intrinsic variability, Level of detail, Noisiness, Objectivity, Outliers, Precision, Precision of domains, Representational conciseness, Spatial resolution, Time resolution, Unambiguous, Uncertainty, Variation
Traceability	Attributable, Capture, Contemporaneous, Documentation, Fairness, Identifiability, Lineage, Meta-data, Original, Provenance, Quality of Methodology, Source, Traceability, Translatability, Transparency, Verifiability
Understandability	Characteristic series structure, Clarity, Clean, Complexity, Comprehensibility, Content, Ease of interpretation, Ease of understanding, Format, Formats, Information-to-noise ratio, Informativeness, Interpretable, Interpretability, Legible, Presentation, Presentation quality, Quantitativeness, Readability, Semiotic, Structure, Transformation, Understandability, Understandable, Understanding, Visualisation
Usefulness *	Applicability, Appropriateness, Artificiality, Definition, Essentialness, Expandability, Fitness, Fitness for Purpose, Fitness for use, Flexibility, Importance, Interoperability, Irrelevant, Meaningful, Naturalness, Relevance, Relevancy, Relevant, Reusability, Suitability, Uniqueness, Useableness,Usability, Usefulness, Utility, Valuation, Value, Value-added, Versatility

Table 5. System-dependent Data Quality Dimensions based on ISO 25012. This table details the dimensions that underscore the role of the system in data quality, including aspects related to data availability, portability, recoverability, and quantity. Each dimension is defined and associated with terms from the literature, illustrating how system characteristics can impact data quality. The table also introduces two newly added dimensions: Quantity and Semantics. [4,9,10,12,14,15,18,19,21,24,25,27–30,32,33,35–37,40,43–47,49,51,56–58, 61,62,66,70,75,76,79,83–85,90,95,98,100,101,103,107,115–117,119,120,123,126,127,130–132,134,137].

Dimensions	Associated Terms
Availability	Access security, Adequacy, Attainability, Availability, Available, Obtainability, Visibility
Portability	Controllability, Mobility, Portability, Use of Storage
Quantity *	Amount of data, Appropriate amount, Compactness, Data volume, Scalability, Sufficiency, Suitable amount, Volume
Recoverability	Back-up, Decay, Recoverability
Semantics *	Interlinking, Language, Semantic accuracy, Semantic consistency, Syntactic validity, Syntax

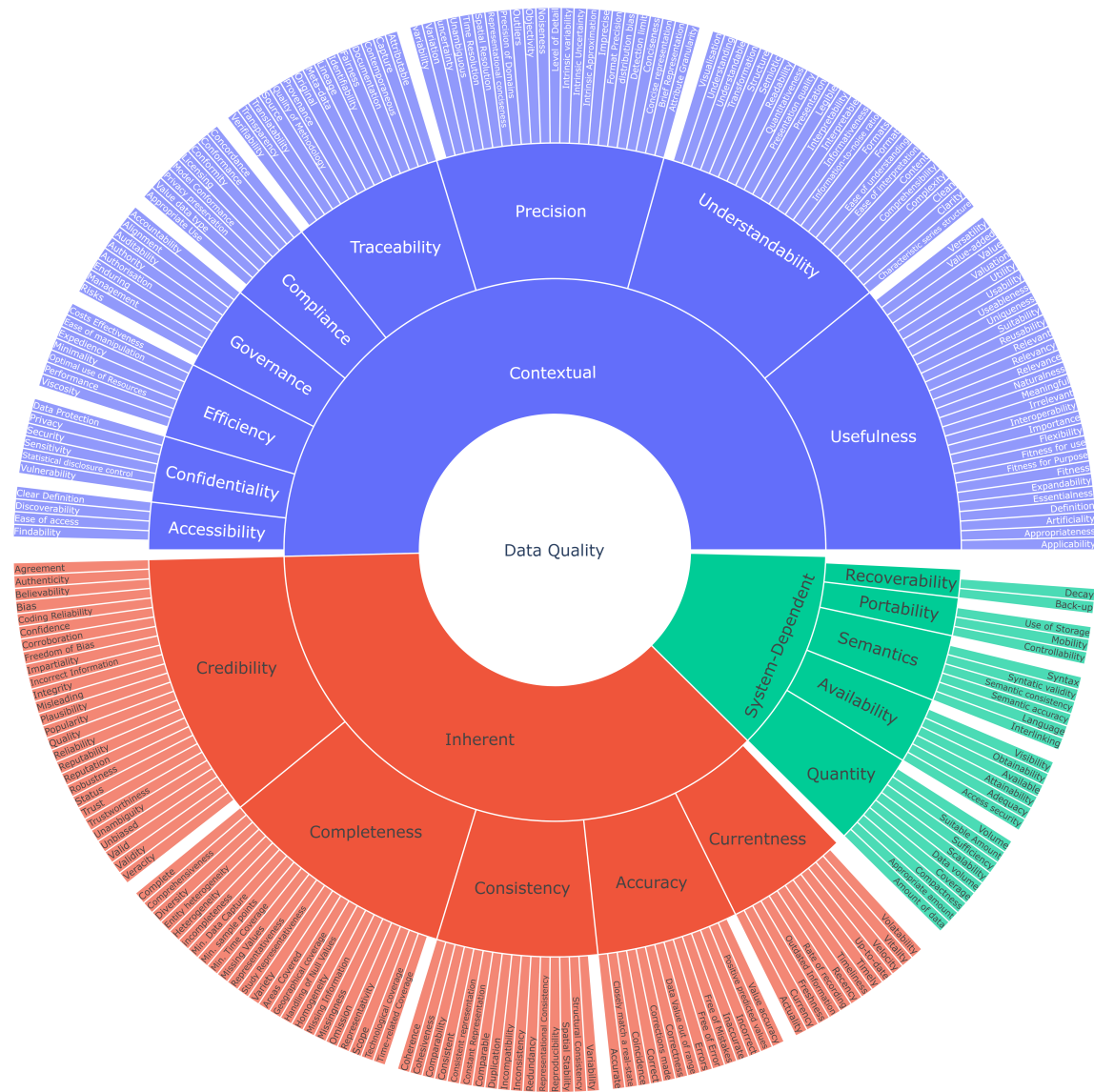


Figure 2. Structured glossary of data quality terminology. The first concentric ring categorises data quality into three domains: inherent, contextual, and system-dependent. Each domain is further split into core data quality dimensions in the second ring. The outermost ring aligns all associated terms found in the literature, and presented in Tables 3–5, with their corresponding core term.

4. Discussion

4.1. Inherent Data Quality

According to ISO 25012, inherent data quality is defined as the degree to which data's quality characteristics have the intrinsic potential to meet stated and implied needs when data is used under specified conditions. It refers to how well the intrinsic attributes and values of data follow constraints and rules that make them fit for use.

The importance of maintaining robust inherent quality cannot be overstated, as the accuracy of downstream processes is only as good as the quality of the source data. Errors in these areas can cause bigger problems later when the data is used for reporting, analytics, and decision-making.

For instance, invalid product codes in a purchase order database (domain values) can result in shipment and accounting complications. Similarly, contradictory order statuses over time (logical consistency) can disrupt plans for fulfillment prioritisation. Additionally, a lack of descriptions for order status code meanings (metadata) can inhibit accurate analysis, leading to shipping errors.

Therefore, it is crucial to use robust validation against business rules, governance oversight, and thorough documentation to tackle quality issues at their root. By examining key downstream data quality dimensions, such as **accuracy**, **completeness**, **credibility** and **consistency**, we can verify that inherent data integrity measures, such as comprehensive validation rules, governance policies, and descriptive metadata, are enabling the data to meet quality expectations across intended usage requirements.

4.1.1. Accuracy

Two closely related dimensions of data quality are **accuracy** and **precision**. While they are often cited together, they serve distinct roles in data quality and should not be confused. Accuracy is defined 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" while precision is defined as "the degree to which subject data associated with an entity has values for all expected attributes and related entity instances in a specific context of use".

Accuracy is one of the most frequently cited data quality dimensions. It specifically relates to how closely data values match the true value concerned. With this formal definition in mind, all **correctness** [7,15,18,30,32,38,48,59,62,66,68,80,99,103,117,135], **free of error** [48,103,119,132], and accuracy-oriented terminology were categorised under the Accuracy dimension. For instance, the term **closely match a real-state** [38], a state depicting the true values of data being represented, logically aligns with accurately reflecting ground truth.

Additionally, out-of-range violations directly contradict defined accuracy bounds, signaling deviation from expectations rather than natural variation [92]. Such breaches can significantly impact accuracy when judgements are based on preset expectations. For example, a sensor fault could result in a temperature reading outside the expected range, leading to a home heating system's controller to set the thermostat's set-point much lower/higher than required. This would compromise both comfort and heating efficiency in the home.

4.1.2. Completeness

Completeness is defined as "The degree to which subject data associated with an entity has values for all expected attributes and related entity instances in a specific context of use." Discussions on the scope of completeness highlighted aspects of heterogeneity, such as **diversity** [73,80] and **variety** [4, 9,47,49,51,66,79,101,103], that both capture the idea of data spanning a broad scope of attributes and characteristics ensuring a comprehensive representation of the domain. **Representativeness** [56,94,127, 130] goes beyond simply filling gaps in the data and involves adequately capturing the full spectrum of variability within a dataset.

Coverage [66,76,103,127] refers to whether the available data encompasses information from all the necessary elements required for a comprehensive overall representation of the concept. For example, survey data could have responses from only a small subset of people in a population. In that case, it has limited coverage to make conclusions about behaviours and traits of the overall population. Sufficient coverage thus requires the data should capture information, metrics and perspectives from key segments that cumulatively contribute to meaningfully conveying the complete phenomenon.

Minimum data capture rates and **coverage** [66,76,103,127] across important sampling dimensions such as time and geography are also directly tied to completeness, as they ensure the necessary representation. The core argument is that true completeness requires not only addressing missing information but also accurately portraying the entire scope of the domain, including the anticipated **diversity**. For example, consider a dataset about customer preferences for a product. Completeness would involve not only ensuring that each customer record has all the necessary attributes filled in (e.g., age, gender, location) but also making sure that the dataset includes a representative sample of customers from various demographics, regions, and preference categories. This way, the dataset can provide a comprehensive and accurate picture of the customer landscape, enabling more reliable insights and decision-making.

4.1.3. Consistency

Consistency is defined as “the degree to which data has attributes that are free from contradiction and are coherent with other data in a specific context of use. It can be either or both among data regarding one entity and across similar data for comparable entities”.

Two terms containing “accuracy” were classified under the dimension of consistency. These are **thematic accuracy** and **syntactic accuracy** [57]. **Thematic accuracy** refers to the correct classification of entities. However, it aligns more with the application of standardised rules for integration than with representing an absolute truth [57]. When datasets are coded or classified for grouped analysis, the accuracy of individual data points may be compromised to some extent. For example, encoding detailed satellite imagery into classified map layers or land cover types for geospatial analysis may result in loss of accuracy for localised regions. A vegetation mapping system that categorises hyper-local flora into broad biome categories like forests, grasslands, deserts, amongst others, may lose precision on individual plant species details. However, this approach enables aggregated analytics about vegetation distribution. The benefits of unified coding frameworks for cross-regional agricultural pattern insights outweigh the localised accuracy loss. In this context, consistency in applying the classification rules holds more importance than the accuracy of individual data points, and that is what is described by **thematic accuracy**.

Building on this idea of prioritising consistency, **syntactic accuracy** requires various datasets to adhere to a common structural schema or standard format for interoperability, such as XML (eXtensible Markup Language) or JSON (JavaScript Object Notation) [57]. These are widely used formats for structuring, exchanging or representing data. This adherence may not perfectly mirror the native representations within each source system, meaning it might not match the original format in which the data was stored or created. Some characteristics that make the data useful or meaningful may need to be compromised to fit uniform syntactical mandates, or rules about how the data should be structured. However, adherence to the schema, the structure and organisation of the data, ensures reliable interchange through expected consistency. This improves consolidation, the process of combining data from multiple sources, and migration capabilities, the ability to move data between systems or formats, despite potential impacts on accuracy for fringe data elements, which are data elements not commonly used.

Just as **thematic accuracy** and **syntactic accuracy** prioritise consistency by generalising observations into shared categorical frameworks, thematic conformity also focuses less on accuracy than enabling unified analysis through abstraction semantics. Abstraction semantics refers to the process of simplifying complex data into more understandable or manageable formats. Similarly, while **syntactic**

accuracy matters for data interchange, shared schema adherence provides consistency but does not inherently ensure the accuracy or truthfulness of the data itself. Inherent truth correspondence refers to data accurately representing the true value of the intended attribute of a concept or event. Mandating compatible structural representations, even if not completely accurate, enables integrating data and deriving unified meaning across different contexts. Hence, both **thematic accuracy** and **syntactic accuracy** qualify as consistency dimensions by prioritising coherent interpretations under standardised constraints.

Though inefficient at first, true duplicates — that is **duplication** [58,81,82], or the replication of identical data or attributes for the same entity — allow for the representations of the same entity to diverge over time, resulting in inconsistency. It is important to note that repeat measurements of the same item, which may vary due to factors such as measurement error or changes in the item over time, would not be considered true duplicates. As such, **duplication** should be considered a potential issue related to consistency.

4.1.4. Credibility

Credibility is defined as “the degree to which data has attributes that are regarded as true and believable by users in a specific context. Credibility includes the concept of authenticity (the truthfulness of origins, attributions, commitments).”

While **integrity**[44] is primarily an ethical attribute, which directly impacts perceived credibility, as it implies researcher honesty and mitigates doubts from questionable practices that would undermine **believability**. Similarly, the absence of bias provides more objective evidence to support credibility, as impartiality aligns with factual accuracy.

Although semantically related, **veracity** [4,47,49,66,79,101,103,122] fundamentally differs from **credibility**. It represents comprehensive quality assurance across datasets, and encompasses attributes of accuracy, completeness, and freedom from distortion. When handling vast volumes of data, there are risks of missing components, inaccurate elements, or an inability to provide meaningful insight if they are left unchecked.

Data **veracity** thus indicates the level of confidence placed on extensive information pools through supplemental reliability checks and controls beyond limited samples. It constitutes an applied practice of instilling trust by promoting centralised policies that enable effective data governance and quality control to enhance the integrity across diverse datasets. In contrast, credibility centers intrinsically on inherent conformity to facts. However, while they are distinct concepts, **veracity** quantification can be considered a measure of the **trustworthiness** [32,36,56,75,78,82,83,89,94,109,127] or **credibility** of a very large data provider.

4.1.5. Currentness

Currentness is defined as “the degree to which data has attributes that are of the right age in a specific context of use”. This definition aligns directly with the focus on **timeliness** [9,12–14,18,19,26,27,30,32,37–40,42–46,48,50–52,54,57,59,64,65,71,76,77,79,80,83–86,88,91,94,97,98,100,102,103,105,107,109,110,112,113,117–119,126–130,132,134]. The term **vitality** [101] encapsulates the concept of maintaining relevance over time, as opposed to becoming obsolete or out-of-date. **Volatility** [38,48,128], on the other hand, signals time sensitivity. Highly volatile data can quickly become outdated if not refreshed promptly, thus necessitating timely maintenance to preserve its usability. By accounting for **volatility** exposure, proactive planning for **currentness** can be achieved, rather than resorting to reactive measures to combat obsolescence.

4.2. Contextual Data Quality

The ISO 25012 standard outlines several dimensions as described in Section 3, we identify a need to further broaden the scope to encompass additional contextual considerations. Consequently, we propose the addition of two new dimensions, Governance and Usefulness under Contextual

dimensions. The rationale behind their introduction stems from the absence of existing dimensions that adequately capture the associated terms identified during our literature review.

Contextual data quality is an important concept that emphasises evaluating the quality of data within the context it is used in. For example, dimensions like Accessibility, Compliance, and Precision should be considered when evaluating contextual data quality. Assessing data along these types of contextual dimensions highlights the multifaceted nature of data quality - whether data is “good” depends significantly on the context of its application.

To illustrate, let us consider the dimension of Accessibility. In a healthcare setting, data must be readily accessible to healthcare providers for timely decision-making, but stringent controls must be in place to prevent unauthorised access, highlighting the interplay between Accessibility and Compliance.

Considering contextual dimensions in a comprehensive manner allows a more complete assessment to determine overall fitness for use. Understanding contextual data quality leads to better data-driven decisions because it highlights how key quality issues relate to the environment data gets used in. Ignoring context can undermine data value by overlooking crucial contextual factors within specific applications; hence, evaluating these factors is crucial for having the right data in the right format at the right time to fully derive value from it.

4.2.1. Accessibility

The dimension of **accessibility** pertains to how readily data can be accessed and obtained within a specific context. This dimension encompasses several related terms identified in the literature review, all of which focus on data access in the face of user constraints dictated by the context. **Discoverability** or **findability** [100], and **clear definition** [43], are closely tied to **accessibility** [10,14,29,32,37,38,43–45, 48,50,57,60,64,69,75,90,94,100,103,116,117,119,120,126–128,130–132]. They highlight the user’s ability to understand and locate data within context-dependent settings.

4.2.2. Compliance

Compliance refers to the degree to which data aligns with externally relevant regulations, rules, standards, and policies within a specific context. This dimension is centered around fulfilling these governance requirements. **Conformance** [15,88,117,135] and **conformity** [18,30,58,123,128] are associated terms that denote the alignment of data with applicable standards or conventions. **Model Conformance** [138] specifically pertains to adherence when dealing with structured or schematic data models.

While there is an overlap between **compliance** and **confidentiality** [30,90,116,131] in the context of Privacy Preservation, compliance specifically necessitates adherence with regulations and policies to restrict access and uphold confidentiality via suitable security controls. Thus, it is more about enforcing confidentiality rather than the concept itself, which is a component of compliance. Lastly, **licensing** [83,127] is about complying with any terms related to data usage access rights.

4.2.3. Confidentiality

The **confidentiality** dimension is concerned with the extent to which data attributes limit access to only authorised users. This dimension involves controls and policies that restrict access permissions to legitimate, authenticated users.

As such **data Protection** [73] examines the safeguards actively used to secure data in accordance with confidentiality constraints. While **privacy** pertains to the appropriate access to any sensitive or personal information, **security** [14,29,40,48,66,83,85,103,117,119,127,132] focuses on technical controls like encryption that prevent unauthorized user access, even in the event of an external system compromise.

Sensitivity [32,103] refers to levels such as high, medium, or low that dictate the degree of confidentiality required. **Vulnerability** [14,101], on the other hand, assesses potential exposures that

could compromise existing access controls if exploited by an unauthorised party. **Statistical disclosure control** [43,60] is associated with the anonymisation of data to minimise the risks of identifying individuals in a dataset.

4.2.4. Efficiency

The **efficiency** dimension addresses whether data attributes allow for the achievement of expected system performance levels and objectives through appropriate resource use. Terms in this category span economic factors, computational efficiency and resource optimisation.

Viscosity refers to the resistance that slows or inhibits the movement and transformation of data. When data is used across different sources or processing pipelines, friction and barriers that reduce flow can introduce inefficiencies. High **viscosity** [101] implies that data does not integrate or stream smoothly to where it needs to go next. Essentially, it describes how efficiently data flows to enable tasks to be completed.

4.2.5. Governance

The proposed dimension of **governance** encompasses organisational structures and policies that guide data activities. The relationship between Governance and Compliance in data management is complex and often overlaps, leading to potential confusion about their distinct roles and objectives. Whilst both concepts are critical for effective data management, it is essential to understand their differences and how they complement each other. Governance focuses on the internal organisational structures, policies, and processes that guide data activities within an enterprise. It establishes the internal rules, responsibilities, and authorities for data management, ensuring that data is handled consistently and in alignment with the goals and values of the organisation. On the other hand, **compliance** is primarily concerned with adhering to external regulations, standards, and policies imposed by regulatory bodies or authorities. It ensures that internal practices and procedures align with external mandates and requirements, such as legal regulations, industry standards, and contractual obligations. Although compliance often necessitates the existence of internal rules and policies, which are aspects of governance, it encompasses a broader range of external requirements beyond enterprise-specific rules. Treating **governance** and **compliance** as separate allows for a clearer definition of their respective scopes and emphasises the importance of considering both internal governance and external compliance requirements in data management. By understanding and addressing both Governance and Compliance, organisations can establish a robust framework for managing data effectively, ensuring internal consistency and alignment with external requirements.

Authority [66] and **authorisation** [37] are manifestations of governance policies that designate permitted data actions across users. These concepts focus on the rules that determine who or what has been officially approved to access and interact with data in specific ways. This is distinct from actual physical controls and system security that enable real-time access, where authorisation is pre-approved based on roles, responsibilities, and data sensitivity. For instance, an employee may be authorised to view sales records but lack the authority to modify them, thus the policy allows read but not write privileges.

Accountability [66] and **management** [20] directly implement governance principles through coordination oversight. **Alignment** [98] falls under governance and assesses behavioural consistency and adherence to centralised policies across an organisation. **Auditability** [37,127] enables accountability monitoring through tracking, serving as a pillar of functional governance. While there is overlap between the dimension of **traceability** and the associated terms of **auditability**, **traceability** refers to the ability to trace the change or state of uniquely identified data points across time in a meaningful sequence.

Auditability specifically pertains to the processes and controls in place to store transactional records of how data has been accessed and modified. For example, retaining details like timestamps, data field changes, and the source user or program making edits supports comprehensive audit trails.

Implementing such **auditability** makes ongoing data usage and alterations transparent. This allows for factual verification of compliance with policies and procedural benchmarks, thereby upholding accountability for proper data handling. Lastly, **risk** [76] assessment is directly tied to governance's risk mitigation duties as stewards of institutional data assets.

4.2.6. Traceability

The **traceability** dimension pertains to the provision of audit trails, capturing data access, and modifications. Although it is a distinct concept from **auditability**, **traceability** aids in providing the necessary infrastructure that auditability uses. At its core, **traceability** involves preserving attribution and documenting events to enable tracing lineage back to original sources. **Lineage** [46] directly maps as a crucial component within the broader scope of traceability, based on its role in preserving and conveying chains of upstream data provenance sources pointing to origination events that led to current states.

Quality of methodology [60] and **transparency** [73] are essential for enabling others to appraise the credibility of tracing procedures. It ensures that the methods used are scientifically justified or internationally recognised. This includes using approved quality procedures, regularly monitoring and enhancing data collection and processing, applying suitable data correction techniques, validating statistical indicators and the conduction of audits. **Transparency** involves providing clear documentation and disclosure of the steps, assumptions, data sources, and tools used.

Identifiability [43] mechanisms link activities directly to the actors involved by relying on policies that avoid obfuscation, violating user rights. Similarly, **verifiability**, which refers to the ability to confirm the accuracy and reliability of traceability information, is doubly essential. It enables post hoc reconstruction following incidents, allowing for the examination and validation of data traceability to understand what happened and why. Moreover, it supports proactive conformity checks, confirming that policies are implemented as intended.

4.2.7. Precision

Precision refers to the exactness and the ability to discern minor variations between data points across a distribution. In ISO 25012, it is referred to as discrimination, which is described as the ability to distinguish subtle differences between phenomena rather than making coarse generalisations [2]. Following the distinction between accuracy and precision, associated terms describing uncertainty can be defined to directly shape data discrimination, aligning with precision's remit regarding the exactness of attributes.

Initially, both presence of **outliers** [116,124,126] (values that significantly differ from the rest of the dataset) and **data values out of range** [92] (values outside the defined range for a dataset) were classified under the Accuracy dimension. However, we later moved **outliers** to the Precision dimension. This is because outliers often represent real-world variation, not necessarily inaccuracy. While it is acknowledged that measurement outliers caused by errors do reduce truth representation, the argument remains that outliers generally relate to distributions. Capturing them ensures that variability in real-life phenomena is represented.

4.2.8. Understandability

The **understandability** dimension emphasises **interpretability** [10,14,18,27,40,43,45,48,50,60,65,81,85,103,119], which is the extent to which users can accurately derive meaning from data attributes within specific contexts.

Visualisation [66,101] plays a crucial role in transforming data into consumable formats, determining whether insights are accessible to target users without requiring expertise in intermediate representation languages. Purposeful symbol and icon selection, along with deliberate graphical arrangement shape the conveyance of meaning, thereby improving clarity and comprehension.

Characteristic series structure [87] examines how when presenting a series of data points (e.g., in a line graph or chart), the way the x-axis values are ordered and the spacing between consecutive ordinal values can influence how well users can perceive patterns, trends, or missing data within that series. This refers to assessing concepts like spacing between points on a visual chart and whether data increments and scales progress evenly, as this impacts how understandable and continuous data trends appear to those interpreting meaning from information flows. For instance, in a line graph tracking expenses over time, having inconsistent gaps between temporal data points or uneven jumps between measurement values on the vertical scale axis (uneven ordinal spacing) would distort the perception of the trend for viewers. Judgment becomes more challenging when formal spacing between sequences or scalar comparisons is misaligned.

Information-to-noise ratio [66] measures the amount of useful, meaningful content compared to irrelevant data. Higher ratios improve **understandability** by allowing users to focus on informative patterns rather than misleading fluctuations. **Semiotic** [117] content is directly related to **understandability** based on its focus on analysing how encoded signs and structural formatting choices either enhance or inhibit accurate audience interpretation.

4.2.9. Usefulness

The **usefulness** dimension represents the ability of data attributes to effectively enable purpose-driven applications and analyses within relevant contexts. Fitness dimensions such as **fitness for use** and **fitness for Purpose** [98,112] were classified under this dimension as they describe whether data aligns with functional requirements. **Versatility** [83,127] involves the adaptable flexibility enabling sustained usefulness across varying constraints and evolving requirements.

The relevance of **interoperability** [52,56,83,91,98,100,109,122,127,129,130,132,140] stems from combined data sources empowering more capable analyses compared to isolated datasets alone; integrating across systems and content types unlocks expanded utility. Reusing data assets for distinct applications beyond their original purpose increases overall usefulness and utility. However, as data assets get reused there is a greater need for governance as it becomes necessary to protect sensitive information as well as uphold dependability.

Cross-purpose use of data requires greater scrutiny to confirm data validity, integrity, and value carry over into broader decision contexts while security and privacy safeguarding also evolves appropriately. Conscientious data governance ensures new uses align to original intent and reliability standards. Therefore, **reusability** [100] is considered an associated term of usefulness. However, it is important to note that the concept of **reusability** in this context differs from that of its namesake provided by the FAIR principles.

Uniqueness [19,30,39,65,72,74,88,103,118,127,128,142] directly enables usefulness by revealing previously inaccessible insights through rare and novel data. Exclusive assets with one-of-a-kind properties intrinsically expand analytical scope into exclusive analytical opportunities others cannot explore. However, highly unique data may form smaller, narrower datasets. This can limit potential scope of applications if analytics require a large amount of data. There is also the concern, of course, that such data could be spurious. So uniqueness possesses an inherent trade-off between novelty and constraints on wide applicability when data subsets become too sparse.

Expandability [103] directly relates to usefulness because data assets that enable additional capacities through potential expansion increase overall utility. Planning for scalable data growth allows meeting evolving objectives over time.

In applications such as machine learning, if data sets are not large enough, “real data” can be used to generate meaningful “fake” data. These data are commonly known as synthetic data. Artificially or synthetically generated data can increase usefulness by enabling scenario modeling, augmentation for underrepresented domains, and privacy-preserving analysis while retaining informative patterns.

However, unlike raw captures of reality, artificial constructs represent simulated approximations of true underlying dynamics. Without safeguards ensuring fidelity and accurate representations in

some applications, use of synthetic data can be misleading and result in ill-informed decisions. Thus, artificial data holds usefulness for situations where representative real data proves inadequate if evaluative rigor ensures accuracy.

4.3. System-Dependent Data Quality

System-dependent data quality refers to the aspects of data quality that are influenced by and specific to the technological or organisational system that collects, stores, manages, and provides access to the data. As formally defined in the ISO 25012 standard, system-dependency includes availability, recoverability and portability - dimensions that can vary substantially based on capabilities and constraints of interconnected data sources, pipelines, and storage options. Two additional dimensions are proposed: Quantity and Semantics.

The complex configuration of different technologies in modern data environments necessitates closer scrutiny and careful governance of these System-dependent Data Quality factors. Both these aspects are critical not just for reliable analytics, but also operational stability, risk minimisation and continuous data quality improvement across interconnected systems, promoting Big Data and IoT.

Consider an organisation with various databases across different regions. When attempting to integrate these databases, the system-dependent data quality becomes apparent. Data availability may be high within each local system, but it may be limited when trying to access it from a different region due to network constraints or data sovereignty laws.

Addressing system-dependent data quality requires a comprehensive understanding of the system landscape. This includes establishing clear data definitions (to ensure consistent semantics), implementing reliable backup and recovery processes (for data recoverability), and designing flexible data architectures (to support data portability).

4.3.1. Availability

The fundamental notion captured by the **availability** dimension is whether users can access the data they need when they require it. This means that the data should be readily obtainable and usable by authorized individuals or systems whenever it is necessary for their specific purpose. The terms assigned under this dimension, including **access security** [10,14,45,85,130], **adequacy** [116], **attainability** [57], **obtainability** [43], **usability** [9,37,66,70,75,76,103,107,123,126,127,137] and **visibility** [127], all impact the degree to which data meets this criteria of availability. This category includes terms which address both the user's ability to find, access and retrieve the data, and the adequacy of the data available to a given user.

4.3.2. Portability

The key requirement captured by the **portability** dimension is preserving utility and meaning of data when moving across storage, software, and hardware environments. **Portability** directly denotes the ease of data transition across different systems and environments without any loss of quality or integrity. It ensure data can be seamlessly moved between multiple platforms and applications whilst retaining its original meaning. Portability enables data to be used and reused across multiple contexts without the need for extensive transformations or adaptations, saving time and effort. It also minimises the risk of data corruption, inconsistencies, or loss during the transition process. By prioritising portability, organisations can ensure that their data remains accessible, usable, and valuable regardless of the specific technology stack or infrastructure in use.

Building on the idea of **portability**, **mobility**, **controllability** and **use of storage** further refine how data **portability** can be effectively managed and optimised across different technical frameworks, towards enhancing the overall utility and integrity of data during transfers. **Mobility** [30] refers to how easily whole data sets can be ported across these frameworks. **Controllability** [30] corresponds to standardised mechanisms governing validation, transport, and backup to enable error-free porting.

Lastly, **use of storage** [43] aligns to leveraging portable storage formats, protocols and abstraction layers that prevent system-dependence and degradation risks.

In this context, abstraction layers refer to functional tiers that hide complex implementation details behind simplified interfaces. Each layer provides services to the layer above it whilst using capabilities from the layer below. This enables modular design by decoupling high-level business needs from low-level technical realisations. Abstraction creates portability across different systems through well-defined Application Programming Interfaces (APIs) - mechanisms that enable multiple software components to communicate - and protocols between abstracted layers instead of actual implementations. It enables interoperability, ease of modification and reuse across diverse deployment environments. In essence, appropriately layered abstractions minimise external dependencies whilst revealing only essential functions.

4.3.3. Quantity

The **quantity** dimension assesses whether the amount and coverage of available data is sufficient to completely and accurately capture information for its intended application. It also considers if existing data assets have enough detail, breadth and granularity in terms of volumes and varieties, to support downstream applications and decisions. **Amount of data** [14,85], **data volume** [71,103,130], and **volume** [4,9,12,14,46,47,49,51,66,79,101,103] directly quantify absolute or proportional magnitude of data.

Scalability [66] examines ability to expand or down-sample data quantities to meet application requirements. **Sufficiency** [27], **suitable amount** [40] and **appropriate amount** [10,14,45,48,80,98,103,132] evaluate if quantities meet adequacy thresholds for intended analytic tasks and decisions. **Compactness** [66] evaluates storage optimisation through compression and minimising redundancy to retain necessary details while maximising efficiency. Quantifying this compactness trade-off enables retaining the most relevant information and reducing data volumes enough to enable efficient large-scale processing and storage. Having appropriate data volumes is crucial for machine learning and big data analytics to uncover insights.

Thus, these associated terms assess if different quantitative needs around comprehensiveness, adequacy, scalability and storage efficiency are fulfilled so that data quantities can provide a satisfactory informational picture.

4.3.4. Recoverability

The **recoverability** dimension requires attributes that enable data to withstand disruptive events and restore original fidelity, operability and utility. The terms **backup** [117], **decay** [76], and **recoverability** [30,90,116,131,134] directly bolster these necessities. Backup refers to maintaining redundant, secondary data copies using archival and snapshot techniques for reinstantiating compromised data post-outages. **Decay** corresponds to safeguarding information integrity over elongated retention cycles against deterioration or distortions over time through aging. Recoverability denotes mechanisms to rapidly invoke fail-safe points, redeploy historical instances, repair corrupted elements via backups to resume services with minimal data loss.

4.3.5. Semantics

The **semantics** quality dimension requires data to have enough contextual attributes to convey interpretable meaning to users or applications within specific contexts. Contextual attributes are additional pieces of information that provide background, descriptive details, or related characteristics about the main data points. These attributes help to clarify the meaning, significance, and implications of the data within a particular setting or use case. **Semantic accuracy** [57,83,127] ensures information objectively represents real-world entities, properties and relationships without distortion or ambiguity. **Semantic consistency** [43,61] requires “persistent”, “unified” and “coherent” meaning per standard

definitions despite usage and modifications over time. Thus, **semantic accuracy and consistency** quantify preservation and stability of meaning over data modification cycles.

For example, consider an electronic health record (EHR) system used in a hospital. **Semantic accuracy** ensures that each patient's demographic information, medical history, diagnoses, and treatment plans are accurately recorded and reflect their real-world health status without any errors or misinterpretations. If the blood type of a patient is recorded as "A+" in one section of the EHR but as "A-" in another, it could lead to serious medical errors and a lack of semantic accuracy.

Building on the same example, **semantic consistency**, in this context, ensures that the meaning and interpretation of the patient data remain the same across different healthcare providers, applications, and time periods. If the "diagnosis" field in the EHR initially represents the primary diagnosis but later includes secondary diagnoses without clear labeling, it could lead to confusion and inconsistencies in understanding the health condition of the patient. Maintaining **semantic consistency** ensures that all healthcare providers can correctly interpret and use the patient data for effective treatment and decision-making.

Syntax and **syntactic validity** govern structuring information elements and validation of schematic rules to enable processing. **Language** [43] denotes representing information per conventions and vocabularies suited for sharing understanding across different user groups. **Interlinking** [83,120,127] establishes explicit linkages across data sources and elements to enrich insights from semantic connections. By governing faithful representation and relation integrity, these associated terms ensure data generates intended meaning and responds reliably during analytic processing.

While the argument could be made that most of the terms the authors have considered as part of the **semantics** Data Quality dimension could be aggregated into existing dimensions, such as **semantic accuracy** and **semantic consistency** under **accuracy** and **consistency** respectively, the fact is that semantic-driven technologies (such as ontologies and knowledge graphs) are becoming increasingly relevant and bring a very specific set of requirements and regulations, which merit the addition of **semantics** as a standalone Data Quality dimension.

For example, consider a knowledge graph used in a biomedical research platform. The knowledge graph integrates data from various sources, such as scientific literature, clinical trials, and patient records, to enable researchers to discover new insights and relationships between diseases, drugs, and genes. In this context, the **semantics** dimension becomes crucial, as it ensures that the data is not only accurate and consistent but also semantically rich and meaningful.

In this scenario, the use of a well-defined ontology that captures the complex relationships between biomedical entities. By leveraging the ontology, the knowledge graph can accurately represent the relationships between a particular gene and its associated diseases, enabling researchers to make informed hypotheses and decisions. The semantic richness provided by the ontology goes beyond simple accuracy and consistency, as it allows for the inference of new knowledge and the identification of previously unknown connections.

On the other hand, a knowledge graph that suffers from semantic inconsistencies and inaccuracies can lead to false conclusions and misdirected research efforts. For instance, if the knowledge graph incorrectly associates a gene with a disease due to conflicting or outdated information from different sources, researchers may draw erroneous conclusions based on this inaccurate semantic representation. Similarly, if the relationships between entities are inconsistently represented across different parts of the knowledge graph, it could lead to confusion and hinder the ability to draw meaningful insights. In these cases, the lack of semantic accuracy and consistency undermines the reliability and usefulness of the knowledge graph, even if the data satisfies other quality dimensions.

5. Conclusions

In this paper, we surveyed recent literature focusing on Data Quality, and have found that a severe lack of standardised terminology exists in the field. This has led us to explore the aggregation of disparate Data Quality terms under common umbrella terms, with an initial focus on the ISO 25012

standard: “Data quality model for Software product Quality Requirements and Evaluation”. The choice of this standard as the starting point for this work is deliberate given Data Science, as an area of research and application, shares many common elements with Software Development.

These umbrella terms, for describing dimensions of Data Quality, can be classified as inherent, contextual, and system-dependent. While the framework proposed in ISO 25012 does a commendable job at capturing these dimensions, we discovered that many terms prevalent in recent literature did not fit into any of the umbrella terms described in the standard. To address this gap, we proposed the addition of four additional Data Quality dimensions: Governance, Usefulness, Quantity, and Semantics. The first two are contextual dimensions, while the latter two are system-dependent. The addition of these four dimensions to those already established by the ISO 25012 standard increases complexity, but also enhances specificity, enabling end-users to fully understand the aspect of Data Quality captured by each dimension.

This creates a consistent representation of the multifaceted aspects of Data Quality and enables the design of a Data Quality Data Model that can serve as a common and generalisable framework for assessing data quality, irrespective of the intended application. In conclusion, our research underscores the need for a more comprehensive, adaptable, and sector-sensitive approach to Data Quality assessment, aiming to facilitate collaboration and communication of Data Quality terminology and assessment across different domains.

6. Future Directions

Moving forward, the next steps in this research will be to operationalise the Data Quality Data Model proposed in this paper. The aim is to develop a mechanism that can quantify data quality according to each dimension. This will involve creating a metric or set of metrics that can accurately measure the quality of data in each dimension.

Another key area of future work will be the introduction of a weighting factor. This factor will be adjustable depending on industry needs, allowing for a more tailored approach to data quality assessment. This will ensure that the model remains flexible and adaptable to various industry contexts.

Finally, we plan to develop a use case to demonstrate the practical application of our model. This will provide a tangible example of this framework can be used in real-world scenarios, further validating its effectiveness and utility.

Author Contributions: **Conceptualization:** João Gregório and Paul Duncan; **methodology:** João Gregório, Russell Miller, and Harvey Whelan; **software:** Russell Miller, and Harvey Whelan; **validation:** João Gregório and Paul Duncan; **formal analysis:** João Gregório, Russell Miller, and Harvey Whelan; **investigation:** João Gregório, Russell Miller, Harvey Whelan, David Whittaker, and Michael Chrubasik; **resources:** João Gregório and Paul Duncan; **data curation:** João Gregório, Russell Miller, and Harvey Whelan; **writing—original draft preparation:** Russell Miller, Harvey Whelan; **writing—review and editing:** Michael Chrubasik, David Whittaker, João Gregório; **visualization:** Russell Miller and Harvey Whelan; **supervision:** João Gregório; **project administration:** Paul Duncan; **funding acquisition:** Paul Duncan. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the UK Government Department for Science, Innovation and Technology through the UK’s National Measurement System.

Data Availability Statement: Data sharing is not applicable.

Acknowledgments: Thanks to Moulham Alsuleman, and Louise Wright for providing feedback on the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Liu, C.; Peng, G.; Kong, Y.; Li, S.; Chen, S. Data Quality Affecting Big Data Analytics in Smart Factories: Research Themes, Issues and Methods. *SYMMETRY-BASEL* **2021**, *13*.
2. International Standards Organization. ISO/IEC 25012:2008. Technical report, International Organization for Standardization, 2008.

3. Chen, H.; Hailey, D.; Wang, N.; Yu, P. A Review of Data Quality Assessment Methods for Public Health Information Systems. *International Journal of Environmental Research and Public Health* **2014**, *11*, 5170–5207.
4. Liu, J.; Li, J.; Li, W.; Wu, J. Rethinking big data: A review on the data quality and usage issues. *ISPRS Journal of Photogrammetry and Remote Sensing* **2016**, *115*, 134–142.
5. Ekegren, C.; Gabbe, B.; Finch, C. Sports Injury Surveillance Systems: A Review of Methods and Data Quality. *Sports medicine* **2016**, *46*, 49–65.
6. Abdullah, M.; Arshah, R. A Review of Data Quality Assessment: Data Quality Dimensions from User's Perspective. *Advanced Science Letters* **2018**, *24*, 7824–7829.
7. Stausberg, J.; Nasseh, D.; Nonnemacher, M. Measuring Data Quality: A Review of the Literature between 2005 and 2013. *Building Capacity for Health Informatics in the Future*. IOS press, 2015, Vol. 210, pp. 712–716.
8. Wang, X.; Williams, C.; Liu, Z.; Croghan, J. Big data management challenges in health research-a literature review. *Briefings in Bioinformatics* **2019**, *20*, 156–167.
9. Ijab, M.T.; Surin, E.S.M.; Nayan, N.M. Conceptualizing big data quality framework from a systematic literature review perspective. *Malaysian Journal of Computer Science* **2019**, pp. 25–37.
10. Liu, G. Data quality problems troubling business and financial researchers: A literature review and synthetic analysis. *Journal of Business & Finance Librarianship* **2020**, *25*, 315–371.
11. Teh, H.; Kempa-Liehr, A.; Wang, K. Sensor data quality: a systematic review. *Journal of Big Data* **2020**, *7*.
12. Salih, F.; Ismail, S.; Hamed, M.; Yusop, O.; Azmi, A.; Azmi, N. Data Quality Issues in Big Data: A Review. Recent Trends in Data Science and Soft Computing: Proceedings of the 3rd International Conference of Reliable Information and Communication Technology (IRICT 2018). Springer, 2019, Vol. 843, pp. 105–116.
13. Ibrahim, A.; Mohamed, I.; Satar, N. Factors Influencing Master Data Quality: A Systematic Review. *International Journal of Advanced Computer Science and Applications* **2021**, *12*, 181–192.
14. Mansouri, T.; Moghadam, M.; Monshizadeh, F.; Zareravasan, A. IoT Data Quality Issues and Potential Solutions: A Literature Review. *The Computer Journal* **2023**, *66*, 615–625.
15. Iturry, M.; Alves-Souza, S.; Ito, M. Data Quality in health records: A literature review. 2021 16th Iberian Conference on Information Systems and Technologies (CISTI). IEEE, 2021.
16. Engsig-Karup, T.; Doupi, P.; Makinen, M.; Launa, R.; Estupinan-Romero, F.; Bernal-Delgado, E.; Kristiansen, N. Review of data quality assessment frameworks experiences around Europe, 2022.
17. Ozonze, O.; Scott, P.; Hopgood, A. Automating Electronic Health Record Data Quality Assessment. *Journal of Medical Systems* **2023**, *47*.
18. Mashoufi, M.; Ayatollahi, H.; Khorasani-Zavareh, D.; Boni, T. Data Quality in Health Care: Main Concepts and Assessment Methodologies. *Methods of Information in Medicine* **2023**, *62*, 5–18.
19. Morewood, J. Building energy performance monitoring through the lens of data quality: A review. *Energy and Buildings* **2023**, *279*.
20. Pradhan, S.; Heyn, H.; Knauss, E. Identifying and managing data quality requirements: a design science study in the field of automated driving. *Software Quality Journal* **2023**.
21. Zhang, L.; Jeong, D.; Lee, S. Data Quality Management in the Internet of Things. *Sensors* **2021**, *21*.
22. Firmani, D.; Mecella, M.; Scannapieco, M.; Batini, C. On the Meaningfulness of Big Data Quality (Invited Paper). *Data Science and Engineering* **2016**, *1*, 6–20.
23. Fenza, G.; Gallo, M.; Loia, V.; Orciuoli, F.; Herrera-Viedma, E. Data set quality in machine learning: consistency measure based on group decision making. *Applied Soft Computing* **2021**, *106*, 107366.
24. Kavasidis, I.; Lallas, E.; Leligkou, H.C.; Oikonomidis, G.; Karydas, D.; Gerogiannis, V.C.; Karageorgos, A. Deep Transformers for Computing and Predicting ALCOA+ Data Integrity Compliance in the Pharmaceutical Industry. *Applied Sciences* **2023**, *13*, 7616.
25. Durá, M.; Sánchez-García, Á.; Sáez, C.; Leal, F.; Chis, A.E.; González-Vélez, H.; García-Gómez, J.M. Towards a computational approach for the assessment of compliance of ALCOA+ Principles in pharma industry. *Studies in Health Technology and Informatics* **2022**, *294*, 755–759.
26. Jaya, I.; Sidi, F.; Ishak, I.; Affendey, L.; A. Jabar, M. A review of data quality research in achieving high data quality within organization. *Journal of Theoretical and Applied Information Technology* **2017**, *95*, 2647–2657.
27. Wand, Y.; Wang, R.Y. Anchoring Data Quality Dimensions in Ontological Foundations. *Commun. ACM* **1996**, *39*, 86–95.

28. Durá, M.; Leal, F.; Sánchez-García, Á.; Sáez, C.; García-Gómez, J.M.; Chis, A.E.; González-Vélez, H. Blockchain for data originality in pharma manufacturing. *Journal of Pharmaceutical Innovation* **2023**, pp. 1–19.
29. Alosert, H.; Savery, J.; Rheume, J.; Cheeks, M.; Turner, R.; Spencer, C.; S. Farid, S.; Goldrick, S. Data integrity within the biopharmaceutical sector in the era of Industry 4.0. *Biotechnology Journal* **2022**, *17*, 2100609.
30. Efimova, O.V.; Igolnikov, B.V.; Isakov, M.P.; Dmitrieva, E.I. Data Quality and Standardization for Effective Use of Digital Platforms. 2021 International Conference on Quality Management, Transport and Information Security, Information Technologies (IT&QM&IS), 2021, pp. 282–285.
31. Arts, D.G.; De Keizer, N.F.; Scheffer, G.J. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. *Journal of the American Medical Informatics Association* **2002**, *9*, 600–611.
32. Weiskopf, N.G.; Weng, C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association* **2013**, *20*, 144–151.
33. Hock, S.C.; Tay, V.; Sachdeva, V.; Wah, C.L. Pharmaceutical Data Integrity: issues, challenges and proposed solutions for manufacturers and inspectors. *Generics and Biosimilars Initiative Journal* **2020**, *9*, 171–183.
34. Boukouvala, F.; Muzzio, F.J.; Ierapetritou, M.G. Predictive modeling of pharmaceutical processes with missing and noisy data. *AIChE journal* **2010**, *56*, 2860–2872.
35. Tabersky, D.; Woelfle, M.; Ruess, J.A.; Brem, S.; Brombacher, S. Recent regulatory trends in pharmaceutical manufacturing and their impact on the industry. *Chimia* **2018**, *72*, 146–146.
36. Leal, F.; Chis, A.E.; Caton, S.; González-Vélez, H.; García-Gómez, J.M.; Durá, M.; Sánchez-García, A.; Sáez, C.; Karageorgos, A.; Gerogiannis, V.C.; others. Smart pharmaceutical manufacturing: Ensuring end-to-end traceability and data integrity in medicine production. *Big Data Research* **2021**, *24*, 100172.
37. Cai, L.; Zhu, Y. The challenges of data quality and data quality assessment in the big data era. *Data science journal* **2015**, *14*, 2–2.
38. Zulkifli, P.; Akshir, E.; Azis, N.; Cox, K. The development of data quality metrics using thematic analysis. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)* **2019**, *8*, 304–310.
39. Hub, G.D.Q. The Government Data Quality Framework. Technical report, Government Digital Service, 2020.
40. Botha, M.; Botha, A.; Herselman, M. Compiling a Prioritized List of Health Data Quality Challenges in Public Healthcare Systems. IST-Africa 2014 Conference Proceedings. IEEE, 2014.
41. Heinrich, B.; Klier, M. Metric-based data quality assessment - Developing and evaluating a probability-based currency metric. *Decision Support Systems* **2015**, *72*, 82–96.
42. Cappiello, C.; Pernici, B.; Villani, L. Strategies for Data Quality Monitoring in Business Processes. Lecture Notes in Computer Science (LNCS). Springer, 2015, Vol. 9051, pp. 226–238.
43. Jesilevska, S. Data quality aspects in latvian innovation system. New Challenges of Economic and Business Development–2016. University of Latvia, 2016, pp. 307–320.
44. Ortega-Ruiz, L.; Caro, A.; Rodriguez, A. Identifying the Data Quality terminology used by Business People. 2015 34th International Conference of the Chilean Computer Science Society (SCCC). IEEE, 2015.
45. Laranjeiro, N.; Soydemir, S.; Bernardino, J. A Survey on Data Quality: Classifying Poor Data. 2015 IEEE 21st Pacific rim international symposium on dependable computing (PRDC). IEEE, 2015.
46. Becker, D.; McMullen, B.; King, T. Big data, big data quality problem. 2015 IEEE international conference on big data (big data). IEEE, 2015, pp. 2644–2653.
47. Rao, D.; Gudivada, V.; Raghavan, V. Data Quality Issues in Big Data. 2015 IEEE International Conference on Big Data (Big Data). IEEE, 2015, pp. 2654–2660.
48. Juddoo, S. Overview of data quality challenges in the context of Big Data. 2015 International Conference on Computing, Communication and Security (ICCCS). IEEE, 2015.
49. Taleb, I.; El Kassabi, H.; Serhani, M.; Dssouli, R.; Bouhaddioui, C. Big Data Quality: A Quality Dimensions Evaluation. 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld). IEEE, 2016, pp. 759–765.

50. Jiang, H.; Liang, L.; Zhang, Y. An Exploration of Data Quality Management Based on Allocation Efficiency Model. *Proceedings of 20th International Conference on Industrial Engineering and Engineering Management: Theory and Apply of Industrial Management*. Springer, 2015, pp. 313–318.
51. Haug, F. Bad Big Data Science. 2016 IEEE International Conference on Big Data (Big Data). IEEE, 2016, pp. 2863–2871.
52. Karkouch, A.; Mousannif, H.; Al Moatassime, H.; Noel, T. A Model-Driven Architecture-based Data Quality Management Framework for the Internet of Things. booktitle=2016 2nd International Conference on Cloud Computing Technologies and Applications (CloudTech). IEEE, 2016, pp. 252–259.
53. Rivas, B.; Merino, J.; Caballero, I.; Serrano, M.; Piattini, M. Towards a service architecture for master data exchange based on ISO 8000 with support to process large datasets. *Computer Standards & Interfaces* **2017**, *54*, 94–104.
54. Aljumaili, M.; Karim, R.; Tretten, P. Metadata-based data quality assessment. *VINE Journal of Information and Knowledge Management Systems* **2016**, *46*, 232–250.
55. Heinrich, B.; Hristova, D.; Klier, M.; Schiller, A.; Szubartowicz, M. Requirements for Data Quality Metrics. *Journal of Data and Information Quality (JDIQ)* **2018**, *9*.
56. Edelen, A.; Ingwersen, W. The creation, management, and use of data quality information for life cycle assessment. *The International Journal of Life Cycle Assessment* **2018**, *23*, 759–772.
57. Fu, Q.; Easton, J. Understanding Data Quality Ensuring Data Quality by Design in the Rail Industry. 2017 IEEE International Conference on Big Data (Big Data). IEEE, 2017, pp. 3792–3799.
58. Hart, R.; Kuo, M. Better Data Quality for Better Healthcare Research Results - A Case Study. *Building Capacity for Health Informatics in the Future*. IOS press, 2017, Vol. 234, pp. 161–166.
59. Lim, Y.; Yusof, M.; Sivasampu, S. Assessing primary care data quality. *INT J HEALTH CARE Q* **2018**, *31*, 203–213.
60. Jesilevska, S.; Skiltere, D. Analysis of deficiencies of data quality dimensions. *New Challenges of Economic and Business Development–2017 Digital Economy (2017)*. University of Latvia, 2017, pp. 236–246.
61. Heinrich, B.; Klier, M.; Schiller, A.; Wagner, G. Assessing data quality - A probability-based metric for semantic consistency. *Decision Support Systems* **2018**, *110*, 95–106.
62. Koltay, T. Data governance, data literacy and the management of data quality. *IFLA journal* **2016**, *42*, 303–312.
63. Liu, C.; Talaie-Khoei, A.; Zowghi, D.; Daniel, J. Data Completeness in Healthcare: A Literature Survey. *Pacific Asia Journal of the Association for Information Systems* **2017**, *9*, 75–100.
64. Cichy, C.; Rass, S. An Overview of Data Quality Frameworks. *IEEE Access* **2019**, *7*, 24634–24648.
65. Gyulgyulyan, E.; Ravat, F.; Astsatryan, H.; Aligon, J. Data Quality Impact in Business Intelligence. 2018 Ivannikov Memorial Workshop (IVMEM). IEEE, 2018, pp. 47–51.
66. Abdallah, M. Big Data Quality Challenges. 2019 International Conference on Big Data and Computational Intelligence (ICBDICI). IEEE, 2019.
67. Rajan, N.; Gouripeddi, R.; Mo, P.; Madsen, R.; Facelli, J. Towards a content agnostic computable knowledge repository for data quality assessment. *Computer Methods and Programs in Biomedicine* **2019**, *177*, 193–201.
68. Bronselaer, A.; Nielandt, J.; Boeckling, T.; De Tre, G. Operational Measurement of Data Quality. *COMM COM INF SC*. Springer, 2018, Vol. 855, pp. 517–528.
69. Barsi, A.; Kugler, Z.; Juhasz, A.; Szabo, G.; Batini, C.; Abdulmuttalib, H.; Huang, G.; Shen, H. Remote sensing data quality model: from data sources to lifecycle phases. *International Journal of Image and Data Fusion* **2019**, *10*, 280–299.
70. Liu, Y.; Wang, Y.; Zhou, K.; Yang, Y.; Liu, Y. Semantic-aware data quality assessment for image big data. *Future Generation Computer Systems* **2020**, *102*, 53–65.
71. Liu, C.; Nitschke, P.; Williams, S.; Zowghi, D. Data quality and the Internet of Things. *Computer* **2020**, *102*, 573–599.
72. Cristalli, E.; Serra, F.; Marotta, A. Data Quality Evaluation in Document Oriented Data Stores. *Advances in Conceptual Modeling: ER 2018 Workshops Emp-ER, MoBiD, MREBA, QMMQ, SCME*, Xi'an, China, October 22–25, 2018, Proceedings 37. Springer, 2019, Vol. 11158, pp. 309–318.
73. Firmani, D.; Tanca, L.; Torlone, R. Ethical Dimensions for Data Quality. *Journal of Data and Information Quality (JDIQ)* **2020**, *12*.

74. Grueneberg, K.; Calo, S.; Dewan, P.; Verma, D.; O’Gorman, T. A Policy-based Approach for Measuring Data Quality. 2017 IEEE International Conference on Big Data (Big Data). IEEE, 2019, pp. 4025–4031.
75. Mustapha, J.C.; Mokhtar, S.A.; Jaffar, J.; Boursier, P. Measurement of Data Consumer Satisfaction with Data Quality for Improvement of Data Utilization. 2019 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS). IEEE, 2019, pp. 1–7.
76. Ceravolo, P.; Bellini, E. Towards Configurable Composite Data Quality Assessment. 2019 IEEE 21st Conference on Business Informatics (CBI). IEEE, 2019, Vol. 1, pp. 249–257.
77. Günther, L.C.; Colangelo, E.; Wiendahl, H.H.; Bauer, C. Data quality assessment for improved decision-making: a methodology for small and medium-sized enterprises. *Procedia Manufacturing* **2019**, *29*, 583–591.
78. Ehrlinger, L.; Haunschmid, V.; Palazzini, D.; Lettner, C. A DaQL to Monitor Data Quality in Machine Learning Applications. Database and Expert Systems Applications: 30th International Conference, DEXA 2019, Linz, Austria, August 26–29, 2019, Proceedings, Part I 30. Springer, 2019, pp. 227–237.
79. Ridzuan, F.; Zainon, W.M.N.W. A Review on Data Cleansing Methods for Big Data. *Procedia Computer Science* **2019**, *161*, 731–738.
80. Li, A.; Zhang, L.; Qian, J.; Xiao, X.; Li, X.; Xie, Y. TODQA: Efficient Task-Oriented Data Quality Assessment. 2019 15th International Conference on Mobile Ad-Hoc and Sensor Networks (MSN). IEEE, 2019, pp. 81–88.
81. Souibgui, M.; Atigui, F.; Zammali, S.; Cherfi, S.; Yahia, S.B. Data quality in ETL process: A preliminary study. *Procedia Computer Science* **2019**, *159*, 676–687.
82. Nikiforova, A. Definition and Evaluation of Data Quality: User-Oriented Data Object-Driven Approach to Data Quality Assessment. *Baltic Journal of Modern Computing* **2020**, *8*, 391–432.
83. Albertoni, R.; Isaac, A. Introducing the Data Quality Vocabulary (DQV). *Semantic Web* **2021**, *12*, 81–97.
84. Mulgund, P.; Sharman, R.; Anand, P.; Shekhar, S.; Karadi, P. Data Quality Issues With Physician-Rating Websites: Systematic Review. *Journal of Medical Internet Research* **2020**, *22*.
85. Valencia-Parra, A.; Parody, L.; Varela-Vaca, A.; Caballero, I.; Gomez-Lopez, M. DMN4DQ: When data quality meets DMN. *Decision Support Systems* **2021**, *141*.
86. Onyeabor, G.; Ta’a, A. A Model for Addressing Quality Issues in Big Data. Recent Trends in Data Science and Soft Computing: Proceedings of the 3rd International Conference of Reliable Information and Communication Technology (IRICT 2018). Springer, 2019, Vol. 843, pp. 65–73.
87. Marev, M.; Compatangelo, E.; Vasconcelos, W. Intrinsic Indicators for Numerical Data Quality. 5th International Conference on Internet of Things, Big Data and Security, IoTBDS 2020. Scipress, 2020, pp. 341–348.
88. Sarafidis, M.; Tarousi, M.; Anastasiou, A.; Pitoglou, S.; Lampoukas, E.; Spetsariasis, A.; Matsopoulos, G.; Koutsouris, D. Data Quality Challenges in a Learning Health System. *Studies in health technology and informatics* **2020**, *270*, 143–147.
89. Musto, J.; Dahanayake, A. Integrating data quality requirements to citizen science application design. Proceedings of the 11th International Conference on Management of Digital EcoSystems. Association for Computing Machinery, 2019, pp. 166–173.
90. Musto, J.; Dahanayake, A. Improving Data Quality, Privacy and Provenance in Citizen Science Applications. Information Modelling and Knowledge Bases XXXI. IOS press, 2020, Vol. 321, pp. 141–160.
91. Weatherburn, C. Data quality in primary care, Scotland. *Scottish medical journal* **2021**, *66*, 66–72.
92. Gadde, M.; Wang, Z.; Zozus, M.; Talburt, J.; Greer, M. Rules Based Data Quality Assessment on Claims Database. In *The Importance of Health Informatics in Public Health during a Pandemic*; IOS press, 2020; Vol. 272, pp. 350–353.
93. Foscarin, F.; Rigaux, P.; Thion, V. Data quality assessment in digital score libraries The GioQoso Project. *International Journal on Digital Libraries* **2021**, *22*, 159–173.
94. Piscopo, A.; Simperl, E. What we talk about when we talk about Wikidata quality: a literature survey. Proceedings of the 15th International Symposium on Open Collaboration. Association for Computing Machinery, 2019.
95. Gualo, F.; Rodriguez, M.; Verdugo, J.; Caballero, I.; Piattini, M. Data quality certification using ISO/IEC 25012: Industrial experiences. *Journal of Systems and Software* **2021**, *176*.
96. Schmidt, C.; Struckmann, S.; Enzenbach, C.; Reineke, A.; Stausberg, J.; Damerow, S.; Huebner, M.; Schmidt, B.; Sauerbrei, W.; Richter, A. Facilitating harmonized data quality assessments. A data quality framework

- for observational health research data collections with software implementations in R. *BMC Medical Research Methodology* **2021**, 21.
97. Kong, X. Evaluation of Flight Test Data Quality Based on Rough Set Theory. 2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). IEEE, 2020, pp. 1053–1057.
 98. Wong, K.; Wong, R. Big data quality prediction informed by banking regulation. *International Journal of Data Science and Analytics* **2021**, 12, 147–164.
 99. Lettner, C.; Stumptner, R.; Fagner, W.; Rauchenzauner, F.; Ehrlinger, L. DaQL 2.0: Measure Data Quality based on Entity Models. *Procedia Computer Science*. Elsevier, 2021, Vol. 180, pp. 772–777.
 100. Kong, L.; Xi, Y.; Lang, Y.; Wang, Y.; Zhang, Q. A Data Quality Evaluation Index for Data Journals. *Lecture Notes in Computer Science (LNCS)*. Springer, 2019, Vol. 11473, pp. 291–300.
 101. Taleb, I.; Serhani, M.; Bouhaddioui, C.; Dssouli, R. Big data quality framework: a holistic approach to continuous quality management. *Journal of Big Data* **2021**, 8.
 102. Akgul, M. Data Quality: Success Factors Emergent Research Forum (ERF). *AMCIS 2021 Proceedings*. Association for Information Systems, 2021.
 103. Juddoo, S.; George, C.; Duquenoy, P.; Windridge, D. Data Governance in the Health Industry: Investigating Data Quality Dimensions within a Big Data Context. *Applied System Innovation* **2018**, 1.
 104. Bronselaer, A. Data Quality Management: An Overview of Methods and Challenges. *Lecture Notes in Artificial Intelligence*. Springer, 2021, Vol. 12871, pp. 127–141.
 105. Bogdanov, A.; Degtyarev, A.; Shchegoleva, N.; Khvatov, V. Data Quality in a Decentralized Environment. *Lecture Notes in Computer Science (LNCS)*. Springer, 2020, Vol. 12251, pp. 58–71.
 106. Valencia-Parra, A.; Parody, L.; Varela-Vaca, A.; Caballero, I.; Gomez-Lopez, M. DMN for Data Quality Measurement and Assessment. *Lecture Notes in Business Information Processing*. Springer, 2019, Vol. 362, pp. 362–374.
 107. Fang, Z.; Liu, Y.; Lu, Q.; Pitt, M.; Hanna, S.; Tian, Z. BIM-integrated portfolio-based strategic asset data quality management. *Automation in Construction* **2022**, 134.
 108. Jain, A.; Patel, H.; Nagalapatti, L.; Gupta, N.; Mehta, S.; Guttula, S.; Mujumdar, S.; Afzal, S.; Mittal, R.; Munigala, V. Overview and Importance of Data Quality for Machine Learning Tasks. *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. Association for Computing Machinery, 2020, pp. 3561–3562.
 109. Shenoy, K.; Ilievski, F.; Garijo, D.; Schwabe, D.; Szekely, P. A study of the quality of Wikidata. *Journal of Web Semantics* **2022**, 72.
 110. Hickey, D.; Connor, R.; McCormack, P.; Kearney, P.; Rosti, R.; Brennan, R. The Data Quality Index: Improving Data Quality in Irish Healthcare Records. 24th International Conference Enterprise Information Systems (ICEIS '21). *ICEIS*, 2021, pp. 625–636.
 111. Talha, M.; Kalam, A. Big Data: Towards a Collaborative Security System at the Service of Data Quality. *International Conference on Hybrid Intelligent Systems*. Springer, 2022, Vol. 420, pp. 595–606.
 112. Ehrlinger, L.; Woess, W. A Survey of Data Quality Measurement and Monitoring Tools. *Frontiers in Big Data* **2022**, 5.
 113. AbuHalimeh, A. Improving Data Quality in Clinical Research Informatics Tools. *Frontiers in Big Data* **2022**, 5.
 114. Azeroual, O. Proof of Concept to Secure the Quality of Research Data. Fourteenth International Conference on Machine Vision (ICMV 2021). *Society of Photo-Optical Instrumentation Engineers (SPIE)*, 2022, Vol. 12084.
 115. Caballero, I.; Gualo, F.; Rodriguez, M.; Piattini, M. BR4DQ: A methodology for grouping business rules for data quality evaluation. *Information Systems* **2022**, 109.
 116. Nakajima, S.; Nakatani, T. AI Extension of SQuARE Data Quality Model. 2021 IEEE 21st International Conference on Software Quality, Reliability and Security Companion (QRS-C). IEEE, 2021, pp. 306–313.
 117. Reda, O.; Zellou, A. SMDQM- Social Media Data Quality Assessment Model. 2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET). IEEE, 2022, pp. 733–739.

118. Mohammed, M.; Talburt, J.; Dagtas, S.; Hollingsworth, M. A Zero Trust Model Based Framework For Data Quality Assessment. 2021 International Conference on Computational Science and Computational Intelligence (CSCI). IEEE, 2021, pp. 305–307.
119. Iyengar, A.; Patel, D.; Shrivastava, S.; Zhou, N.; Bhamidipaty, A. Real-Time Data Quality Analysis. 2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI). IEEE, 2020, pp. 101–108.
120. To, A.; Meymandpour, R.; Davis, J.; Jourjon, G.; Chan, J. A Linked Data Quality Assessment Framework for Network Data. Proceedings of the 2nd Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA). Association for Computing Machinery, 2019.
121. Wurl, A.; Falkner, A.; Haselbock, A.; Mazak, A. Using Signifiers for Data Integration in Rail Automation. Proceedings of the 6th International Conference on Data Science, Technology and Applications. Scipress, 2017, pp. 172–179.
122. Kuban, M.; Gabaj, S.; Aggoune, W.; Vona, C.; Rigamonti, S.; Draxl, C. Similarity of materials and data-quality assessment by fingerprinting. *MRS Bulletin* **2022**.
123. Brajkovic, H.; Jaksic, D.; Poscic, P. Data Warehouse and Data Quality - An Overview. Central European Conference on Information and Intelligent Systems. Faculty of Organization and Informatics Varazdin, 2020, pp. 17–24.
124. Valverde, C.; Marotta, A.; Panach, J.; Vallespir, D. Towards a model and methodology for evaluating data quality in software engineering experiments. *Information and Software Technology* **2022**, 151.
125. Serra, F.; Peralta, V.; Marotta, A.; Marcel, P. Modeling Context for Data Quality Management. Lecture Notes in Computer Science (LNCS). Springer, 2022, Vol. 13607, pp. 325–335.
126. Nesca, M.; Katz, A.; Leung, C.; Lix, L. A scoping review of preprocessing methods for unstructured text data to assess data quality. *International Journal of Population Data Science* **2022**, 7.
127. Ben Hassine, S.; Clement, D. Open Data Quality Dimensions and Metrics: State of the Art and Applied Use Cases. Lecture Notes in Business Information Processing. Springer, 2020, Vol. 394, pp. 311–323.
128. Elouataoui, W.; El Alaoui, I.; El Mendili, S.; Gahi, Y. An Advanced Big Data Quality Framework Based on Weighted Metrics. *Big Data and Cognitive Computing* **2022**, 6.
129. Mashoufi, M.; Ayatollahi, H.; Khorasani-Zavareh, D.; Boni, T. Data quality assessment in emergency medical services: an objective approach. *BMC Emergency Medicine* **2023**, 23.
130. Buelvas, J.; Munera, D.; Tobon, V.; Aguirre, J.; Gaviria, N. Data Quality in IoT-Based Air Quality Monitoring Systems: a Systematic Mapping Study. *Water, Air, & Soil Pollution* **2023**, 234.
131. Guerra-Garcia, C.; Nikiforova, A.; Jimenez, S.; Perez-Gonzalez, H.; Ramirez-Torres, M.; Ontanon-Garcia, L. ISO/IEC 25012-based methodology for managing data quality requirements in the development of information systems: Towards Data Quality by Design. *Data and Knowledge Engineering* **2023**, 145.
132. Krishna, C.; Ruikar, K.; Jha, K. Determinants of Data Quality Dimensions for Assessing Highway Infrastructure Data Using Semiotic Framework. *Buildings* **2023**, 13.
133. Mirzaie, M.; Behkamal, B.; Allahbakhsh, M.; Paydar, S.; Bertino, E. State of the art on quality control for data streams: A systematic literature review. *Computer Science Review* **2023**, 48.
134. Bertrand, Y.; Van Belle, R.; De Weerd, J.; Serral, E. Defining Data Quality Issues in Process Mining with IoT Data. Lecture Notes in Business Information Processing. Springer, 2023, Vol. 468, pp. 422–434.
135. Lewis, A.; Weiskopf, N.; Abrams, Z.; Foraker, R.; Lai, A.; Payne, P.; Gupta, A. Electronic health record data quality assessment and tools: a systematic review. *Journal of the American Medical Informatics Association* **2023**.
136. Arden, N.S.; Fisher, A.C.; Tyner, K.; Lawrence, X.Y.; Lee, S.L.; Kopcha, M. Industry 4.0 for pharmaceutical manufacturing: Preparing for the smart factories of the future. *International Journal of Pharmaceutics* **2021**, 602, 120554.
137. Perez-Castillo, R.; Carretero, A.G.; Rodriguez, M.; Caballero, I.; Piattini, M.; Mate, A.; Kim, S.; Lee, D. Data Quality Best Practices in IoT Environments. 2018 11th International Conference on the Quality of Information and Communications Technology (QUATIC). IEEE, 2018, pp. 272–275.
138. Huser, V.; Li, X.; Zhang, Z.; Jung, S.; Park, R.W.; Banda, J.; Razzaghi, H.; Londhe, A.; Natarajan, K. Extending Achilles Heel Data Quality Tool with New Rules Informed by Multi-Site Data Quality Comparison. In *MEDINFO 2019: Health and Wellbeing e-Networks for All*; IOS Press, 2019; pp. 1488–1489.

139. Heine, F.; Kleiner, C.; Oelsner, T. A DSL for Automated Data Quality Monitoring. *Lecture Notes in Computer Science (LNCS)*. Springer, 2020, Vol. 12391, pp. 89–105.
140. Montana, P.; Marotta, A. Data Quality Management oriented to the Electronic Medical Record. 2021 XLVII Latin American Computing Conference (CLEI). IEEE, 2021.
141. Strozyna, M.; Filipiak, D.; Wecel, K. Data Quality Assessment - A Use Case from the Maritime Domain. *Lecture Notes in Business Information Processing*. Springer, 2020, Vol. 394, pp. 5–20.
142. Ji, R.; Hou, H.; Sheng, G.; Jiang, X. Data Quality Assessment for Electrical Equipment Condition Monitoring. 2022 9th International Conference on Condition Monitoring and Diagnosis (CMD). IEEE, 2022, pp. 259–262.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.