Trust the Data You Use: Scalability Assurance Forms (SAF) for a Holistic Quality Assessment of Data Assets in Data Ecosystems

Keywords: Knowledge Graphs, Data Asset Quality, AI Systems Integration, Scalability Assurance Forms (SAF)

Abstract: Companies generate terabytes of raw, unstructured data daily, which requires processing and organization to become valuable data assets. In the era of data-driven decision-making, evaluating these data assets' quality is crucial for various data services, users, and ecosystems. This paper introduces "Scalability Assurance Forms" (SAF), a novel framework to assess the quality of data assets, including raw data and semantic descriptions, with essential contextual information for cross-domain AI systems. The methodology includes a comprehensive literature review on quality models for linked data and knowledge graphs, and previous research findings on data quality. The SAF framework standardizes data asset quality assessments through 31 dimensions and 10 overarching groups derived from the literature. These dimensions enable a holistic assessment of data set quality by grouping them according to individual user requirements. The modular approach of the SAF framework ensures the maintenance of data asset quality across interconnected data sources, supporting reliable data-driven services and robust AI application development.The SAF framework addresses the need for trust in systems where participants may not know or historically trust each other, promoting the quality and reliability of data assets in diverse ecosystems.

1 Introduction

In the context of the exponential growth of *Artificial Intelligence* (AI) and big data, the effective organisation and presentation of vast amounts of knowledge have become crucial. Across various domains and applications, the quality of data and its linked (meta) data descriptions are essential for making well-informed, data-driven decisions. This is evidenced by different findings (Günther et al., 2019; Loh et al., 2020; McCausland, 2021), who states that due to different data processing approaches, it cannot be assumed that the quality and applicability of the data is uniform in different organisations and applications.

High-quality research and analysis depend on reliable data (Arias et al., 2020), a concept epitomized by the adage "garbage in, garbage out" (Kilkenny and Robinson, 2018). Although in literature, discussions on *Data Quality* (DQ) appear relatively recent , the concern with DQ is as longstanding as the practice of data collection itself, even once termed "a key issue of our time" during an era of less digitization (Naroll et al., 1961; Jensen et al., 1986).

Knowledge graphs (KGs) have been demonstrated to be highly effective tools for collecting and articulating knowledge about the real world in the form of graph data. Their capacity to represent complex information is widely recognized, and they are rapidly gaining traction in both academia and industry (Peng et al., 2023). *Linked Data* (LD), the foundation of KGs, promotes the publication and linking of data in a machine-readable format using web standards. This approach enables the linking and reuse of data across organizational silos and facilitates interoperability between different institutions (Radulovic et al., 2017). Furthermore, these technologies facilitate a structured and organized collection of data (NIST, 2020), which we refer to as a *Data Asset* (DA) in the course of this work. DA are used for business monitoring and decision-making, such as databases or even an Excel spreadsheet, as opposed to unorganized raw data that has no immediate use. To illustrate, a database containing information on the entry of identified persons into a room can be considered a DA, whereas the raw data generated by a key card reader is not considered a DA.

Data Indicators vs. Semantic Indicators. Building on this foundation, the subsequent sections of this paper will elaborate on a holistic approach to *Data Asset Quality* (DAQ) assessment, categorized into *Data Indicators* and *Semantic Indicators*. These categories are devised to provide a comprehensive framework for evaluating the robustness of datasets within KG and across AI systems.

- *Data Indicators (DI)* focus on the intrinsic quality of raw data, assessing aspects such as accuracy, completeness, and consistency.

For example, in a healthcare dataset, a DI might evaluate the precision of diagnostic codes and the presence of patient records without missing values. This level of scrutiny ensures that the foundational data used in AI algorithms is reliable and robust, mitigating risks associated with poor DQ.

- *Semantic Indicators (SI)* pertain to the semantic descriptions of datasets or applications, encompassing the structured interlinking and contextual relevance of data. These indicators evaluate how effectively data is described and linked, akin to the metadata or LD standards used to enhance data discoverability and usability. An instance of this could be assessing the adequacy of annotations in a scholarly database, where the clarity and correctness of metadata directly influence the ease of data integration and retrieval across different academic platforms

As exemplified in the public transport domain by bus departure time datasets: the use of DI would involve rigorous verification of the accuracy of the dataset, its temporal completeness, and the synchronicity of schedules across different transit routes. A robust data indicator would check that departure timestamps are not only accurate to the minute but also consistently formatted and complete for each route, with no missing or ambiguous entries. Such care ensures the reliability of the dataset, a critical factor in the development of AI systems for route optimization and predictive modeling in urban mobility.

SI, in this context, would delve into the semantic richness of the dataset, ensuring that each departure time is adequately described with contextually relevant metadata. This may include *Resource Description Framework* (RDF) annotations linking each timestamp to corresponding route identifiers, bus capacities, accessibility features, or integration with real-time traffic conditions. By embedding this semantic layer, the dataset goes beyond simple planning to provide a comprehensive set of information that can integrate seamlessly with smart city infrastructures and deliver insightful, actionable information to end users. In the remainder of this paper, we refer to the manifestations of DI and SI as dimensions. Illustrative examples of these dimensions are presented in Figure 1. Together, these indicators form the backbone of our methodology, addressing the dual aspects of DQ (Kilkenny and Robinson, 2018; Hassenstein and Vanella, 2022; Batini and Rula, 2021) and semantic richness (Zaveri et al., 2015; Wang et al., 2021) to enhance the utility and reliability of data-asset-driven systems (Radulovic et al., 2017). This integrated assessment approach not only aligns with the strategic goals of semantic interoperability but also ensures that both data and its contextual framework are optimized for cross-domain applications.

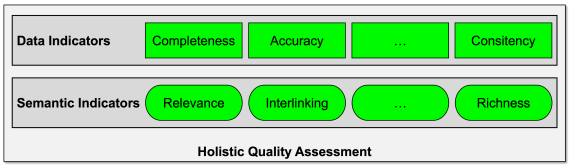

Figure 1: Conceptual model of the Holistic Quality Assessment framework, delineating the DI (square shapes) and SI (round shapes). DI like *Completeness*, *Accuracy*, and *Consistency* evaluate raw DQ, while SI like *Relevance*, *Interlinking*, and *Richness* assess semantic aspects of data within KGs. This ensures comprehensive quality assessment for robust AI systems.

ISO Standard 25012 Within the ISO Standard 25012 dimensions are defined as distinct aspects of DQ that can be measured and assessed independently. By differentiating these aspects, the standard delineates a general DQ model for data held in a structured format within a data-driven system, emphasizing quality dimensions for target data used by humans and systems. It categorizes DQ requirements and measures aligned with these dimensions, enabling an evaluation process to analyze data independently from other components of the computer system. Our approach adopts these established dimensions as a template to guide our investigation, ensuring that our methodology aligns with recognized standards and provides a robust basis for assessing DQ in KGs and AI systems. This strategic focus on dimensions, rather than diving deeply into individual metrics, positions our research as a foundational reference point, facilitating subsequent detailed studies aimed at refining these quality assessments. Building upon the foundation of holistic DQ assessment through DI and SI, it is crucial to note that quality in this context is measured using specific dimensions, which are qualified through various metrics. In our work, we focus on these dimensions to lay the groundwork for future research, as they are commonly defined at the dimension level in existing literature and ISO standards. Examining the various metrics that can be employed to quantify the different dimensions or to describe how to measure the different dimensions for different DAs is outside the scope of this study.

Research Questions (RQs). The goal of this research is to analyze existing methods for assessing the quality of structured data in order to identify needed data in an opaque ecosystem. To achieve this goal, we aim to answer how we can holistically evaluate data by including DI and SI.

Thereby, we formulated the subsequent *RQs*:

- **RQ1:** What are the common quality dimensions between raw data and Knowledge Graphs?
- **RQ2:** How can these dimensions be used to holistically and individually assess existing data assets?

By answering the formulated *RQs*, we formulate *Scalability Assurance Forms* (SAF), a novel framework to holistically assess the quality of data assets that include common data quality dimensions as formulated by ISO 25012 and KG-specific quality dimensions. Thereby, our contributions are four-fold:

- Introduction of SAF as a novel framework for orchestrating and assessing DAQ for raw data and knowledge graphs.
- Development of a holistic evaluation approach for DI and SI to ensure the quality and scalability of AI systems.
- Facilitation of integration and analysis through standardized DAQ assessments, reducing redundancy and ensuring data integrity.
- Provision of customization to individual user requirements, which is particularly important in interconnected data ecosystems to support the reliability of data-driven services.

In the following, we will first provide information on the required theoretical background (Chapter 2) on data quality standards, data ecosystems, linked data, and knowledge graphs. Subsequently, we describe our methodology (Chapter 3) and resulting SAF (Chapter 4). We conclude in Chapter 5 by discussing and recapitulating our study.

2 Theoretical Background

A broadly used definition for DQ is the "fitness for use" principle (Juran et al., 1974), which states that "Data are of high quality if they are fit for their intended uses, by customers, in operation, decision making, and planning" (Redman, 2001). More precise definitions, which make clear that DQ depends on the use case and its requirements, are provided by standards such as ISO 25012 (ISO25012, 2008) and DAMA (Kwaliteit, 2023). The ISO standard 25012 defines DQ as the "degree to which the characteristics of data satisfy stated and implied needs when used under specified conditions" (ISO25012, 2008).

Data Quality Standards. Existing DQ dimensions and standards, such as *ISO 25012* and *ISO 8000-2*, play a crucial role in the evaluation and assurance of DQ in various contexts. *ISO 25012*, titled "Data Quality Model", provides a framework for assessing the quality of data based on fifteen key dimensions, including accuracy, completeness, consistency, and timeliness. These dimensions describe various attributes of data that collectively determine the overall quality. For instance, accuracy pertains to the correctness of data, completeness refers to the extent to which expected data is present, consistency ensures data is free from contradictions, and timeliness addresses the relevance of data at a given time. By differentiating these aspects, *ISO 25012* provides a comprehensive framework for evaluating the multifaceted nature of DQ within structured data systems. However, its limitations lie in its generality, as it is not tailored specifically to the complexities of Knowledge Graphs (KGs) or Linked Data (LD), which involve intricate relationships and semantic structures. *ISO 8000-2*, known as the "Data Quality: Vocabulary" standard, focuses on defining terms and concepts related to DQ, aiming to create a common understanding and language for discussing DQ issues. While it provides valuable terminological clarity, it does not offer specific guidelines for implementing quality assessments in dynamic and interconnected data ecosystems. Both standards, while foundational, do not fully address the unique challenges posed by the rapidly evolving fields of AI and big data, where DQ needs to be evaluated in a holistic and scalable manner, especially in federated and distributed environments.

To assess DQ, a data quality model (or framework) is typically established, defined by *ISO 25012* as a "defined set of characteristics which provides a framework for specifying data quality requirements and evaluating data quality". These characteristics (dimensions) encompass both quantitative and qualitative assessments. *ISO 25012* distinguishes between inherent DQ, which refers to the intrinsic potential of data to meet quality needs, and system-dependent DQ, which is influenced by the technological environment. *ISO 8000* defines three meta-characteristics: syntactic quality, which pertains to conformity to specified syntax; semantic quality, which concerns the accurate representation of entities; and pragmatic quality, which relates to conformance to usage-based requirements. These standards provide a foundational basis for DQ assessment, yet they fall short in addressing the specific needs of emerging data architectures (Zhang et al., 2021).

Data Ecosystems. An example of such distributed environments are data ecosystems, a concept that is rapidly materializing, particularly in Europe, embodying a transformative approach to data management and use (Otto et al., 2022). These ecosystems are designed to give individuals and organizations greater sovereignty over their data, embodying the principles of empowerment and control. Within these federated environments, data from multiple sources is brought together, facilitating the creation of interoperable applications that harness the collective power of shared information. The anticipated value of such ecosystems lies in their potential to streamline collaboration, drive innovation, and improve the efficiency of services across sectors (Theissen-Lipp et al., 2023). This new paradigm aims to transcend traditional data silos and promote an open and dynamic exchange of data that is securely accessible and usable within the broader digital economy. As these ecosystems evolve, they are expected to become key pillars in the realization of a unified digital marketplace, fostering economic growth and digital autonomy (Otto et al., 2022). This requires trust not only in the inherent quality of the data but also in the descriptions, context, and semantics that accompany the data (Theissen-Lipp et al., 2023). Therefore, there is a growing need for a holistic approach to assessing the quality of data sets and data-driven applications, particularly in the context of the Semantic Web, where understanding the structure depends on distinguishing between LD and KGs.

Linked Data and Knowledge Graphs. LD is a set of best practices for distributing structured data on the Web (Zaveri et al., 2015). It uses Uniform Resource Identifiers (URIs) and RDF to make data both machine-readable and interoperable (Ji et al., 2022). KGs build on the principles of LD by forming an intricate network of data that encodes knowledge in a machine-understandable format, often supported by inference capabilities (Ban et al., 2024). This structure allows KGs to exploit these semantic connections, creating an advanced, integrated data architecture that is central to sophisticated analytics and AI applications (Pan et al., 2017). In its formal sense, a KG is a graph-based knowledge base consisting of interconnected entities characterized by specific types and attributes. Emerging from the foundations of semantic networks and the principles of LD (Quillian, 1967), KGs have demonstrated their versatility across multiple sectors and are recognized as essential components of modern industrial information systems (Li et al., 2021). Within AI systems, the importance of KGs is underlined by their contribution to metadata quality. This is critical for providing accurate descriptions of data sets, which is essential for developing interoperable systems across domains. By improving metadata quality, KGs mitigate the risk of misinformation, thereby enhancing the trustworthiness and reliability of AI systems (Pan et al., 2024). This synergy of LD and KGs provides a solid foundation for evaluating the quality of data sets and data-driven applications, which is crucial for domain-spanning interoperable systems.

While there are several quality models for KGs and LD, they tend to be broad in scope. However, despite the fundamental nature of LD quality, efforts to standardize quality tracking and assurance are scarce, with a notable lack of consensus on the definition of quality dimensions and metrics (Zaveri et al., 2015; Radulovic et al., 2017). Given the open nature of LD, the diversity of information it contains, and the limitless and dynamic number of autonomous data sources and publishers, conventional methods are inadequate. These challenges call for innovative assessment methods that can accommodate the dynamic and open characteristics of LD and support the automated and scalable exchange of high-quality data across different systems (Zaveri et al., 2015).

3 Methodology

Our methodology is based on a Structured Literature Review (SLR) with a subsequent analysis and synthesis of existing frameworks. To derive new high-level sets of DAQ dimensions for both DI and SI, we first anchored our clustering process to the pre-existing structure provided by the ISO 25012 framework (ISO25012, 2008). This approach ensured that our categorization was in line with recognized standards and provided a solid basis for our analysis.

To mitigate researcher bias, two scientists (RS1 and RS2) from different institutions independently conducted this SLR according to the literature review process as described by (Moher et al., 2010; Kitchenham, 2004). We chose a systematic literature review as a reasonable methodology to: (i) identify open issues and (ii) contribute to a common conceptualization that encompasses the various approaches developed in a field. Thereby, we aim to summarize the established dimensions from the ISO 25012 standard with various methods for evaluating KGs and LD. As a result, we propose a novel framework to enable individual assessments of various data assets within a data ecosystem.

Title	Source
A compendium and evaluation of taxonomy quality attributes	(Unterkalmsteiner and Abdeen, 2024)
A comprehensive quality model for Linked Data	(Radulovic et al., 2017)
A Data Quality Framework for Graph-Based Virtual Data Integration Systems	(Li et al., 2022)
A Data Quality Scorecard to Assess a Data Source's Fitness for Use	(Grillo, 2018)
A Quality Framework for Data Integration	(Wang, 2012)
A Quality Model for Linked Data Exploration	(Cappiello et al., 2016)
A Quality Model for Mashups	(Cappiello et al., 2011)
A Review on Data Quality Dimensions for Big Data	(Ramasamy and Chowdhury, 2020)
A Semiotic Approach to Investigate Quality Issues of Open Big Data Ecosystems	(Krogstie and Gao, 2015)
Architecture and quality in data warehouses	(Jarke et al., 1999)
Big Data Quality Models: A Systematic Mapping Study	(Montero et al., 2021)
Classification of Knowledge Graph Completeness Measurement Techniques	(Issa et al., 2021)
Data Infrastructures for Asset Management Viewed as Complex Adaptive Systems	(Brous et al., 2014)
Data Quality Management in the Internet of Things	(Zhang et al., 2021)
DQ Tags and Decision-Making	(Price and Shanks, 2010)
EPIC: A Proposed Model for Approaching Metadata Improvement	(Tarver and Phillips, 2021)
Evolution of quality assessment in SPL: a systematic mapping	(Martins et al., 2020)
Exploiting Linked Data and Knowledge Graphs in Large Organisations	(Pan et al., 2017)
Information quality dimensions for the social web	(Schaal et al., 2012)
KGMM - A Maturity Model for Scholarly Knowledge Graphs Based on Intertwined Human-Machine Collaboration	(Hussein et al., 2022)
Knowledge Graph Quality Management: a Comprehensive Survey	(Xue and Zou, 2022)
Knowledge Graphs: A Practical Review of the Research Landscape	(Kejriwal, 2022)
Prioritization of data quality dimensions and skills requirements in genome annotation work	(Huang et al., 2012)
Quality assessment for Linked Data: A Survey: A systematic literature review and conceptual framework	(Zaveri et al., 2015)
Quality Evaluation Model of AI-based Knowledge Graph System	(Xu et al., 2021)
Quality factory and quality notification service in data warehouse	(Li and Osei-Bryson, 2010)
Quality model and metrics of ontology for semantic descriptions of web services	(Zhu et al., 2017)
Rating quality in metadata harvesting	(Kapidakis, 2015)
Towards a Critical Data Quality Analysis of Open Arrest Record Datasets	(Wickett and Newman, 2024)
Towards a Data Quality Framework for Heterogeneous Data	(Micic et al., 2017)
Towards a meta-model for data ecosystems	(Iury et al., 2018)
Towards a Metadata Management System for Provenance, Reproducibility and Accountability in Federated Machine Learning	(Peregrina et al., 2022)

Table 1: Presentation of the 32 articles identified as a result of the systematic literature search

Search Strategy. Following Kitchenham et al. (2004), we first defined a search string based on keywords from known base literature (Stvilia et al., 2007; Batini and Scannapieco, 2006; Pernici and Scannapieco, 2003; Madnick et al., 2009).

The title, abstract, and full text of the results were then filtered based on pre-defined inclusion and exclusion criteria. Finally, we conducted a backward search to identify further relevant studies. An overview of our search methodology, including the number of articles found at each step, is presented in Figure 2 and described in detail below. The resulting search string is as follows:

> (”meta data” OR ”meta-data” OR ”metadata” OR ”knowledge graph” OR ”knowledgegraph” OR ”knowledge-graph”) AND (”quality model” OR ”quality framework” OR ”quality concept”)

The broader field of ”data quality” encompasses ”data asset quality” because ensuring the accuracy, consistency, and reliability of data assets is essential for their effective use in business monitoring and decision-making. A total of 395 papers were identified through the search string described (Step 1). The initial filtering removed duplicates and only included studies that were accepted at a peer-reviewed conference, accessible to the authors (open access), written in English or German, and belonging to the research field of computer science or any related subfields. A full overview of the inclusion and exclusion criteria is provided in Table 2.

Subsequent steps focus on filtering and refining these aspects to enhance the overall utility of specific data assets. The filtering and evaluation steps were conducted separately to mitigate researcher bias and ensure comprehensive coverage of all relevant articles (Step 3).

Filtering Search Results. Both reviewers independently evaluated the titles of the 127 articles identified based on the inclusion and exclusion criteria. Following this, they reviewed the abstracts to identify potentially suitable studies. During the filtering process, the two researchers independently excluded non-relevant studies that either focused on data quality management or did not focus on metadata and knowledge graphs. Additionally, work that did not propose a methodology or framework for assessing the quality of metadata or knowledge graphs was excluded. In cases of discrepancies during the merging of the lists, issues were resolved either by mutual consensus or by compiling a list of articles for a more detailed review. The reviewers compared their selections and, by mutual agreement, produced a final list of 48 articles for the first reviewer and 43 articles for the second reviewer to be included in the study.

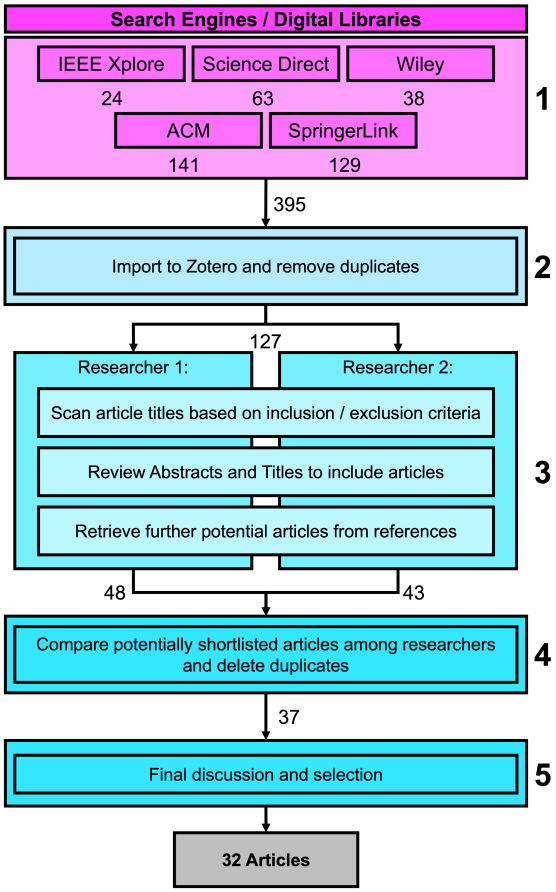

Figure 2: Process of the systematic literature search

To ensure comprehensive coverage of relevant articles, a snowballing approach was employed by:

- Checking the references in the selected articles.
- Using Google Scholar to look up the article titles and retrieving the ”Cited by” articles to compare them with the eligibility criteria.
- Performing a search for related articles for each individual data quality dimension.

This process identified 10 additional articles that met the eligibility criteria (RS1: 7; RS2: 3). Lastly, a total of 37 articles were identified as relevant to our survey. Both reviewers compared their notes and reached a consensus on the selection. Ultimately, 31 relevant articles were chosen. The resulting research is listed in Table 1, which shows the titles of the papers and their authors.

Inclusion Criteria	Exclusion Criteria
Be accepted at a peer-reviewed conference	Have not been peer-reviewed or published, with the exception of dissertations
Be written in English	Use evaluation methods published only as poster abstracts
Belong to the research field of computer science or any related subfields	Focus on data quality management
Be freely available (open access)	Do not focus on metadata or knowledge graphs
Be written in English.	Do not propose a methodology or framework for assessing the quality of metadata or knowledge graphs

Table 2: Inclusion and exclusion criteria for literature

4 Scalability Assurance Forms (SAF)

In this chapter, we present the results of our holistic quality assessment framework. The sets of DAQ dimensions were derived from the extensive literature review described in Section 3 and are shown in Figure 3. Throughout this process, we systematically extracted and analyzed the quality dimensions mentioned in various papers. Through iterative clustering and synthesis, we consolidated these dimensions into the aforementioned groups, ensuring comprehensive coverage and alignment with the dimensions defined in ISO 25012.

- **Accessibility:** The degree to which a DA is available and obtainable for use by authorized entities, ensuring that users can access the DA when needed.
- **Accuracy:** The closeness of DA values to the true values or accepted standard, reflecting the correctness and precision of the data.
- **Connectivity:** The capability of a DA to be connected and interlinked with other data sources, enhancing its usability and integration across systems.
- **Integrity:** The extent to which a DA is complete, consistent, and free from unauthorized modification, ensuring its reliability and trustworthiness.
- **Presentation:** The clarity and interpretability of a DA, including its format and structure, make it comprehensible and usable by intended users.
- **Relevance:** The pertinence and applicability of a DA to the context in which it is used, ensuring that it meets the needs and requirements of users.
- **Security:** The protection of a DA against unauthorized access and breaches, ensuring confidentiality, integrity, and availability of the data.
- **Operational Efficiency:** The degree to which a DA supports effective and efficient business operations, including performance and process optimization.
- **Regulatory Compliance:** The extent to which a DA adheres to laws, regulations, and policies relevant to its use and management, ensuring legal and regulatory conformance.
- **System Flexibility:** The adaptability and maintainability of DA systems to accommodate changes and evolving requirements, ensuring long-term usability and scalability.

Each group contains different dimensions, which are shown in different colours and shapes. The color distinguishes the dependency of the dimensions between inherent, inherent and system-dependent.

- **Inherent Quality** refers to the inherent potential of a DA to satisfy both explicit and implicit requirements under certain conditions, including domain values, constraints, data-asset-value relationships, and metadata.
- **System Dependent Quality** depends on the technological capability of computer systems, including hardware and software, to access a DA, maintain its accuracy, recover it, and facilitate its portability.
- **Inherent and System Dependent Quality** is a hybrid dimension that recognizes the complexity of DQ that arises both inherently and through system interaction and requires a holistic approach to assessment.

The form of the dimension describes the quality dimension and distinguishes between DI, SI and hybrid indicators (HI). The first two dimensions have already been introduced in Chapter 1. The hybrid indicators combine data and semantic indicators in order to thoroughly assess the suitability of data for cross-system use. These dimensions, therefore, apply to both data and semantic descriptions.

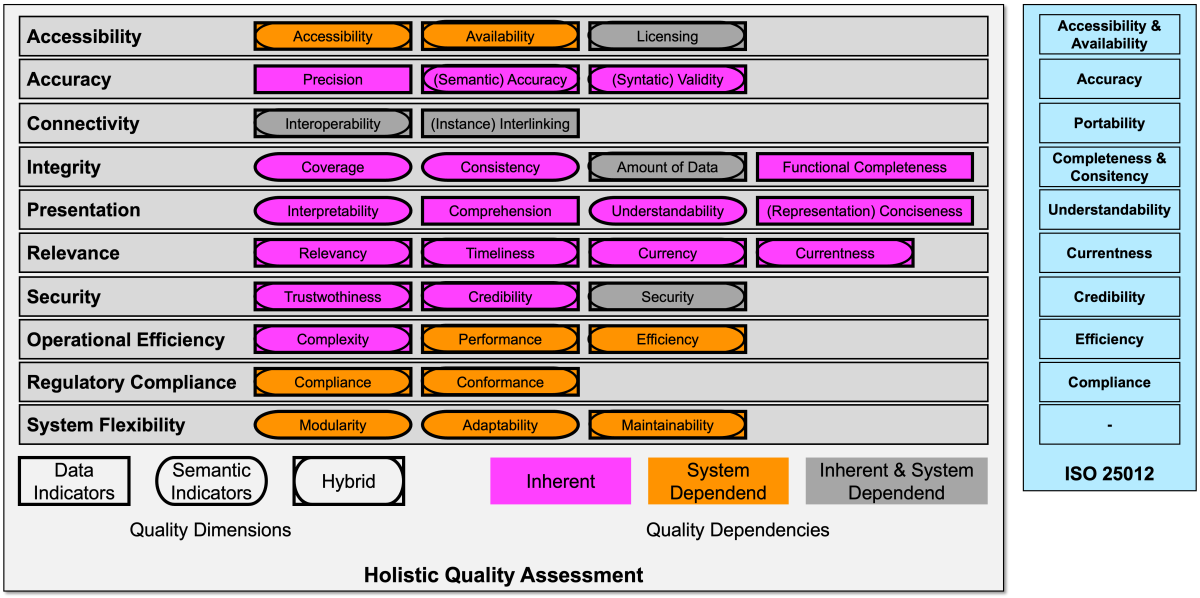

Figure 3: Holistic Quality Assessment overview: The dimensions are classified under overarching groups, reflecting their inherent and system-dependent qualities, and are further mapped onto the ISO 25012 standard.

Figure 3 shows that there are both groups containing only inherent quality dimensions (presentation) and groups containing only system-dependent dimensions (system flexibility). Overall, however, the derived groups correspond very well to the groups defined in ISO 25012. ”System flexibility” is the only group that cannot be mapped to a corresponding complementary group in the ISO standard. A total of 31 dimensions and 10 superordinate groups were extracted from the literature. The dimensions are divided into 5 DI, 6 SI, and 20 HI, as well as 17 inherent dimensions, 9 system-dependent dimensions, and 5 inherent & system-dependent dimensions. The objective of this selection is to provide users with the option to select the dimensions that are pertinent to the specific application from a range of dimensions and groups. The clustering of the quality dimensions and the quality dependencies enables the user to be even more specific in their selection.

4.1 SAF Scores

A deterministic calculation of scores is essential to ensure comparability between different DAs and to meet the individual needs of users and departments. These scores should allow users to decide on an individual and application basis whether the SI or the DI is more important and should, therefore, be weighted more heavily.

The SAF scores are calculated on the basis of a systematic and mathematically sound method and the dimensional assignments from Figure 1. The proposed framework is based on the assumption that for each dimension, there exists an appropriate metric that can be collected and calculated for the corresponding DA. The objective is to combine the metrics for DI and SI in such a way that SAF = DI + SI. First, the mean score for each parent group is calculated by averaging the scores of the underlying features. Let c_i be the score for the i^{th} dimension metric within a group and n be the total number of dimension metrics in that group. This means that a dimension can have multiple metrics. The mean $\overline{C}$ for the group is given by

$$\overline{C} = \frac{1}{n}\sum_{i=1}^{n} c_i$$

We then calculate the DI and SI values. For a dimension classified as a DI, labelled DI_k and belonging to a group with a mean value $\overline{C}$ its calculated value V_{DI_k} is

$$V_{DI_k} = DI_k \cdot \overline{C}$$

Similarly, for a dimension SI_k identified as a semantic indicator, the value V_{SI_k} is calculated using the same formula. Each feature within the DI and SI groups is subjected to this calculation and the results are aggregated to give the overall DI or SI score:

$$\text{Total DI} = \sum V_{DI_k}$$

$$\text{Total SI} = \sum V_{SI_k}$$

The SAF score is then the sum of the total DI and the total SI. To allow for the weighting of DI and SI values, enabling users to prioritize dimensions according

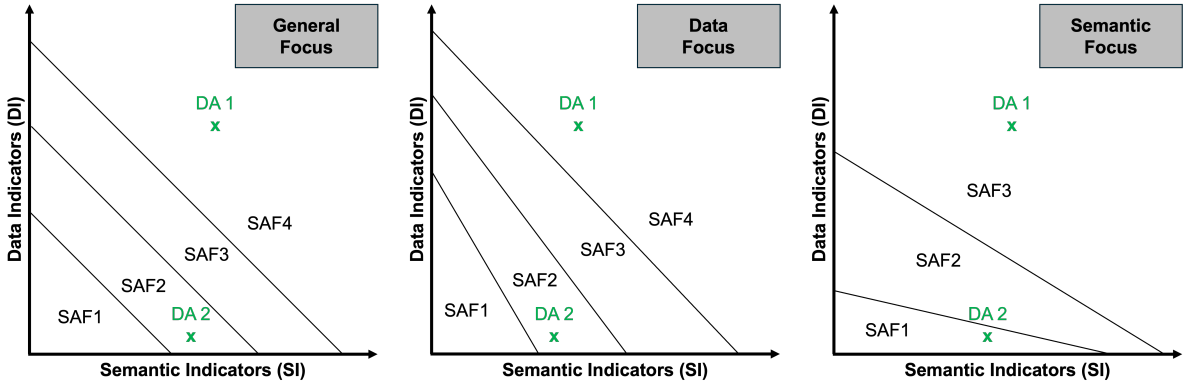

Figure 4: The SAF assessment framework is a comprehensive approach for evaluating heterogeneous DA in three distinct forms. The three diagrams illustrate the distribution of SAF levels based on the assessment focus: general (left), data-oriented (center) and semantic-oriented (right). Furthermore, the framework is adaptable to the assessment priorities defined by the user and the granularity of the SAF grading. It is at the discretion of the user to determine the number of SAF levels and the respective thresholds for these levels for DI and SI. Different DA are shown in green as examples; the X and Y values of the DA are identical in all three diagrams.

to their importance, we introduce weight factors w_{DI_k} and w_{SI_k} for each dimension. The weighted values W_{DI_k} and W_{SI_k} are calculated as follows:

$$W_{DI_k} = w_{DI_k} \cdot V_{DI_k}$$

$$W_{SI_k} = w_{SI_k} \cdot V_{SI_k}$$

The total weighted DI and SI scores are then:

$$\text{Total Weighted DI} = \sum W_{DI_k}$$

$$\text{Total Weighted SI} = \sum W_{SI_k}$$

Finally, the SAF score, incorporating the weights, is calculated as the sum of the total weighted DI and the total weighted SI:

$$\text{SAF} = \text{Total Weighted DI} + \text{Total Weighted SI}$$

Initial, the weight factors w_{DI_k} and w_{SI_k} are set to 1, ensuring that the weighting between SI and DI is balanced when no user-defined weights are applied.

Finally, these values are presented in Figure 4, where SI is on the X-axis and DI on the Y-axis. The evaluation framework illustrates the methodological rigor of the SAF. In the event that a data ecosystem is comprised solely of disparate data assets, a corresponding SAF value would be calculated for each asset, with a weighting defined by the user. These values would then be transferred to the corresponding diagram.

In the left-hand diagram, DI and SI are weighted equally, enabling a balanced assessment. In contrast, the middle diagram represents a semantics-centred evaluation, while the right-hand diagram represents a data-centred approach. Depending on the use case and the focus of the intended application, the user can decide whether to give equal weighting to DA and SI or place a special emphasis on one of the two focal points. For example, in the center diagram in Figure 4, it is important for a data asset to achieve a higher overall rating for DI to advance from level 1 to level 2. A good semantic description and a higher overall SI score would not be as effective for this goal as having high-quality raw data. The opposite is true in the right diagram in Figure 4, where high semantic descriptions (high SI value) are more important than high-quality raw data (high DI value).

The exemplary positions of two different DAs in each of the three graphs in Figure 4 demonstrate that different levels are achieved in the SAF framework depending on the weighting. Consequently, the same DA is assigned a different rating in different contexts. The framework's flexibility is also reflected in the variable SAF levels, which allow the user to define different levels of granularity, represented by lines in Figure 4.

In practice, these levels can be used as thresholds for determining the suitability or usability of different DAs. The delineation of scores in the diagrams illustrates the nuanced interplay of quality attributes within the two main dimensions. This visual tool facilitates a comprehensive assessment of data integrity and highlights the complex dynamics at play in data ecosystems. This adaptability is central to the SAF's usefulness as it allows users to calibrate the assessment based on specific application requirements. It also ensures that the SAF can be tailored to different contexts and needs, thereby demonstrating its potential as a robust tool for a holistic evaluation of data quality. The SAF thus provides a solid frame-

work for a comprehensive, multi-dimensional assessment of data quality, which is crucial for the integrity and utility of data in today's technologically diverse ecosystem.

5 Discussion and Conclusion

In this paper, we developed the Scalability Assurance Forms (SAF) framework, a comprehensive method for assessing data asset quality in data ecosystems. Grounded in ISO 25012, the SAF framework systematically integrates DI and SI to offer a holistic evaluation of data assets. This dual approach ensures that both intrinsic DQ and contextual semantic richness are thoroughly addressed, which is essential for the reliability and scalability of AI applications. The SAF framework presents several advantages. It allows users to prioritize dimensions according to their importance through weight factors, offering a customizable approach to DAQ assessment. This adaptability is crucial for addressing the diverse needs of different data-driven environments and ensures that the quality assessments are both relevant and actionable. Furthermore, by providing a structured method for assessing data assets, the SAF framework supports better decision-making and enhances the trustworthiness of data used in various applications. The holistic view offered by the SAF framework is crucial for users, enabling them to make well-informed decisions and select the most appropriate data assets from complex data ecosystems.

However, there are limitations to the current framework. One significant challenge is the absence of predefined metrics for the various dimensions, which often need to be individually defined and tailored to specific contexts. This process can be complex and time-consuming, requiring extensive domain expertise. Additionally, the field of automated quality assessment in data ecosystems is still in its early stages, and further research is needed to develop robust methodologies and tools. Despite these limitations, future research will focus on defining specific metrics for each dimension and developing a prototype for automated quality assessment. This will enhance the framework's applicability and effectiveness, providing users with more precise and actionable quality assessments.

REFERENCES

Arias, V. B., Garrido, L. E., Jenaro, C., Martínez-Molina, A., and Arias, B. (2020). A little garbage in, lots of garbage out: Assessing the impact of careless responding in personality survey data. *Behavior Research Methods*, 52(6):2489–2505.

Ban, T., Wang, X., Chen, L., Wu, X., Chen, Q., and Chen, H. (2024). Quality Evaluation of Triples in Knowledge Graph by Incorporating Internal With External Consistency. *IEEE Transactions on Neural Networks and Learning Systems*, 35(2):1980–1992.

Batini, C. and Rula, A. (2021). From Data Quality to Big Data Quality: A Data Integration Scenario.

Batini, C. and Scannapieco, M. (2006). *Data Quality: Concepts, Methodologies and Techniques*. Data-Centric Systems and Applications. Springer, Berlin Heidelberg.

Brous, P., Overtoom, I., Herder, P., Versluis, A., and Janssen, M. (2014). Data Infrastructures for Asset Management Viewed as Complex Adaptive Systems. *Procedia Computer Science*, 36:124–130.

Cappiello, C., Daniel, F., Koschmider, A., Matera, M., and Picozzi, M. (2011). A Quality Model for Mashups. In Auer, S., Díaz, O., and Papadopoulos, G. A., editors, *Web Engineering*, volume 6757, pages 137–151. Springer Berlin Heidelberg, Berlin, Heidelberg.

Cappiello, C., Di Noia, T., Marcu, B. A., and Matera, M. (2016). A Quality Model for Linked Data Exploration. In Bozzon, A., Cudre-Maroux, P., and Pautasso, C., editors, *Web Engineering*, volume 9671, pages 397–404. Springer International Publishing, Cham.

Grillo, A. (2018). Developing a Data Quality Scorecard that Measures Data Quality in a Data Warehouse.

Günther, L. C., Colangelo, E., Wiendahl, H.-H., and Bauer, C. (2019). Data quality assessment for improved decision-making: A methodology for small and medium-sized enterprises. *Procedia Manufacturing*, 29:583–591.

Hassenstein, M. J. and Vanella, P. (2022). Data Quality—Concepts and Problems. *Encyclopedia*, 2(1):498–510.

Huang, H., Stvilia, B., Jörgensen, C., and Bass, H. W. (2012). Prioritization of data quality dimensions and skills requirements in genome annotation work. *Journal of the American Society for Information Science and Technology*, 63(1):195–207.

Hussein, H., Oelen, A., Karras, O., and Auer, S. (2022). KGMM – A Maturity Model for Scholarly Knowledge Graphs based on Intertwined Human-Machine Collaboration.

ISO25012 (2008). ISO/IEC 25012:2008.

Issa, S., Adekunle, O., Hamdi, F., Cherfi, S. S.-S., Dumontier, M., and Zaveri, A. (2021). Knowledge Graph Completeness: A Systematic Literature Review. *IEEE Access*, 9:31322–31339.

Iury, M., Oliveira, L., Ribeiro, M., and Lóscio, B. (2018). *Towards a Meta-Model for Data Ecosystems*.

Jarke, M., Jeusfeld, M. A., Quix, C., and Vassiliadis, P. (1999). Architecture and quality in data warehouses:

An extended repository approach. *Information Systems*, 24(3):229–253.

Jensen, D., Wilson, T., Statistics, U. S. B. o. J., and Group, S. (1986). *Data Quality Policies and Procedures: Proceedings of a BJS/SEARCH Conference : Papers*. U.S. Department of Justice, Bureau of Justice Statistics.

Ji, S., Pan, S., Cambria, E., Marttinen, P., and Yu, P. S. (2022). A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2):494–514.

Juran, J. M., Gryna, F. M., and Bingham, R. S., editors (1974). *Quality Control Handbook*. McGraw-Hill, New York, 3d ed edition.

Kapidakis, S. (2015). *Rating Quality in Metadata Harvesting*.

Kejriwal, M. (2022). Knowledge Graphs: A Practical Review of the Research Landscape. *Information*, 13(4):161.

Kilkenny, M. F. and Robinson, K. M. (2018). Data quality: "Garbage in – garbage out". *Health Information Management Journal*, 47(3):103–105.

Kitchenham, B. (2004). Procedures for Performing Systematic Reviews.

Krogstie, J. and Gao, S. (2015). A semiotic approach to investigate quality issues of open big data ecosystems. In Liu, K., Nakata, K., Li, W., and Galarreta, D., editors, *Information and Knowledge Management in Complex Systems*, pages 41–50, Cham. Springer International Publishing.

Kwaliteit, W. D. (2023). Dimensions of Data Quality | Stichting DAMA NL.

Li, X., Lyu, M., Wang, Z., Chen, C.-H., and Zheng, P. (2021). Exploiting knowledge graphs in industrial products and services: A survey of key aspects, challenges, and future perspectives. *Computers in Industry*, 129:103449.

Li, Y., Nadal, S., and Romero, O. (2022). A data quality framework for graph-based virtual data integration systems. In Chiusano, S., Cerquitelli, T., and Wrembel, R., editors, *Advances in Databases and Information Systems*, pages 104–117, Cham. Springer International Publishing.

Li, Y. and Osei-Bryson, K.-M. (2010). Quality factory and quality notification service in data warehouse. In *Proceedings of the 3rd Workshop on Ph.D. Students in Information and Knowledge Management*, PIKM '10, pages 25–32, New York, NY, USA. Association for Computing Machinery.

Loh, W.-Y., Zhang, Q., Zhang, W., and Zhou, P. (2020). Missing data, imputation and regression trees. *Statistica Sinica*, 30(4):1697–1722.

Madnick, S. E., Wang, R. Y., Lee, Y. W., and Zhu, H. (2009). Overview and Framework for Data and Information Quality Research. *Journal of Data and Information Quality*, 1(1):1–22.

Martins, L. A., Afonso Júnior, P., Freire, A. P., and Costa, H. (2020). Evolution of quality assessment in SPL: A systematic mapping. *IET Software*.

McCausland, T. (2021). The Bad Data Problem. *Research-Technology Management*, 64(1):68–71.

Micic, N., Neagu, D., Campean, F., and Habib Zadeh, E. (2017). *Towards a Data Quality Framework for Heterogeneous Data*.

Moher, D., Liberati, A., Tetzlaff, J., and Altman, D. G. (2010). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *International Journal of Surgery*, 8(5):336–341.

Montero, O., Crespo, Y., and Piatini, M. (2021). Big Data Quality Models: A Systematic Mapping Study. In Paiva, A. C. R., Cavalli, A. R., Ventura Martins, P., and Pérez-Castillo, R., editors, *Quality of Information and Communications Technology*, volume 1439, pages 416–430. Springer International Publishing, Cham.

Naroll, F., Naroll, R., and Howard, F. H. (1961). Position of women in childbirth. *American Journal of Obstetrics and Gynecology*, 82(4):943–954.

NIST, C. C. (2020). Data asset - Glossary | CSRC. https://csrc.nist.gov/glossary/term/data_asset.

Otto, B., Ten Hompel, M., and Wrobel, S., editors (2022). *Designing Data Spaces: The Ecosystem Approach to Competitive Advantage*. Springer International Publishing, Cham.

Pan, J. Z., Vetere, G., Gomez-Perez, J. M., and Wu, H., editors (2017). *Exploiting Linked Data and Knowledge Graphs in Large Organisations*. Springer International Publishing, Cham.

Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., and Wu, X. (2024). Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–20.

Peng, C., Xia, F., Naseriparsa, M., and Osborne, F. (2023). Knowledge Graphs: Opportunities and Challenges. *Artificial Intelligence Review*, 56(11):13071–13102.

Peregrina, J. A., Ortiz, G., and Zirpins, C. (2022). Towards a Metadata Management System for Provenance, Reproducibility and Accountability in Federated Machine Learning. In Zirpins, C., Ortiz, G., Nochta, Z., Waldhorst, O., Soldani, J., Villari, M., and Tamburri, D., editors, *Advances in Service-Oriented and Cloud Computing*, pages 5–18, Cham. Springer Nature Switzerland.

Pernici, B. and Scannapieco, M. (2003). Data Quality in Web Information Systems. In Goos, G., Hartmanis, J., Van Leeuwen, J., Spaccapietra, S., March, S., and Aberer, K., editors, *Journal on Data Semantics I*, volume 2800, pages 48–68. Springer Berlin Heidelberg, Berlin, Heidelberg.

Price, R. and Shanks, G. (2010). DQ tags and decision-making. In *2010 43rd Hawaii International Conference on System Sciences*, pages 1–10.

Quillian, M. R. (1967). Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral Science*, 12(5):410–430.

Radulovic, F., Mihindukulasooriya, N., García-Castro, R., and Gómez-Pérez, A. (2017). A comprehensive quality model for Linked Data. *Semantic Web*, 9(1):3–24.

Ramasamy, A. and Chowdhury, S. (2020). Big Data Quality Dimensions: A Systematic Literature Review. *Journal of Information Systems and Technology Management*, page e202017003.

Redman, T. (2001). Data quality: The field guide.

Schaal, M., Smyth, B., Mueller, R. M., and MacLean, R. (2012). Information quality dimensions for the social web. In *Proceedings of the International Conference on Management of Emergent Digital EcoSystems*, Medes '12, pages 53–58, New York, NY, USA. Association for Computing Machinery.

Stvilia, B., Gasser, L., Twidale, M. B., and Smith, L. C. (2007). A framework for information quality assessment. *Journal of the American Society for Information Science and Technology*, 58(12):1720–1733.

Tarver, H. and Phillips, M. E. (2021). EPIC: A proposed model for approaching metadata improvement. In Garoufallou, E. and Ovalle-Perandones, M.-A., editors, *Metadata and Semantic Research*, pages 228–233, Cham. Springer International Publishing.

Theissen-Lipp, J., Kocher, M., Lange, C., Decker, S., Paulus, A., Pomp, A., and Curry, E. (2023). Semantics in Dataspaces: Origin and Future Directions. In *Companion Proceedings of the ACM Web Conference 2023*, pages 1504–1507, Austin TX USA. ACM.

Unterkalmsteiner, M. and Abdeen, W. (2024). A compendium and evaluation of taxonomy quality attributes.

Wang, J. (2012). A Quality Framework for Data Integration. In MacKinnon, L. M., editor, *Data Security and Security Data*, volume 6121, pages 131–134. Springer Berlin Heidelberg, Berlin, Heidelberg.

Wang, X., Chen, L., Ban, T., Usman, M., Guan, Y., Liu, S., Wu, T., and Chen, H. (2021). Knowledge graph quality control: A survey. *Fundamental Research*, 1(5):607–626.

Wickett, K. M. and Newman, J. (2024). Towards a Critical Data Quality Analysis of Open Arrest Record Datasets. In Sserwanga, I., Joho, H., Ma, J., Hansen, P., Wu, D., Koizumi, M., and Gilliland, A. J., editors, *Wisdom, Well-Being, Win-Win*, pages 311–318, Cham. Springer Nature Switzerland.

Xu, Z., Gao, Y., and Yu, F. (2021). Quality Evaluation Model of AI-based Knowledge Graph System. In *2021 3rd International Conference on Natural Language Processing (ICNLP)*, pages 73–78, Beijing, China. IEEE.

Xue, B. and Zou, L. (2022). Knowledge Graph Quality Management: A Comprehensive Survey. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1.

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., and Auer, S. (2015). Quality assessment for Linked Data: A Survey: A systematic literature review and conceptual framework. *Semantic Web*, 7(1):63–93.

Zhang, L., Jeong, D., and Lee, S. (2021). Data Quality Management in the Internet of Things. *Sensors*, 21(17):5834.

Zhu, H., Liu, D., Bayley, I., Aldea, A., Yang, Y., and Chen, Y. (2017). Quality model and metrics of ontology for semantic descriptions of web services. *Tsinghua Science and Technology*, 22(3):254–272.