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

Maximilian Stäbler[1][a], Tobias Müller[2][b], Frank Köster[1] and Chris Langdon[3]

[1]German Aerospace Center (DLR) - Institute for AI Safety and Security, Ulm, Germany

[2]SAP SE, Walldorf, Germany

[3]Drucker School of Business, Claremont Graduate University, Claremont, U.S.A.

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 organization 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. Different studies (Günther et al., 2019; Loh et al., 2020; McCausland, 2021) highlight that due to diverse data processing approaches, data quality and applicability cannot be assumed to be uniform across different organizations 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 discussions on *Data Quality* (DQ) appear relatively recent in the literature, the concern with DQ is as longstanding as the practice of data collection itself (Naroll et al., 1961; Jensen et al., 1986).

To address these challenges, it is crucial to consider a holistic assessment of data assets, encompassing both the structure provided by *Knowledge Graphs* (KGs) and the quality of the raw data itself. A *Data Asset* (DA) refers to any organized collection of data used for business monitoring and decision-making, distinguishing it from unorganized raw data without immediate use (NIST, 2020). KGs, based on *Linked Data* (LD) principles, promote the publication and linking of data in a machine-readable format using web standards, enabling interoperability and reuse across organizational silos (Radulovic et al., 2017). However, focusing solely on the structure without considering the intrinsic quality of the raw data can lead to misleading conclusions and suboptimal decision-making. Likewise, high-quality raw data without a coherent structure lacks the context necessary for comprehensive analysis.

Therefore, a balanced and integrated approach to assessing data assets is necessary. This paper introduces ”Scalability Assurance Forms” (SAF), a novel framework designed to evaluate the quality of data as-

[a] https://orcid.org/0000-0003-1311-3568
[b] https://orcid.org/0000-0002-9088-5054

Stäbler, M., Müller, T., Köster, F. and Langdon, C.
Trust the Data You Use: Scalability Assurance Forms (SAF) for a Holistic Quality Assessment of Data Assets in Data Ecosystems.
DOI: 10.5220/0012915900003825
In *Proceedings of the 20th International Conference on Web Information Systems and Technologies (WEBIST 2024)*, pages 199-208
ISBN: 978-989-758-718-4; ISSN: 2184-3252

sets, including raw data and their semantic descriptions, with essential contextual information for cross-domain AI systems. The SAF framework standardizes data asset quality assessments through 31 dimensions and 10 overarching groups derived from the literature, enabling a holistic assessment that aligns with individual user requirements. This approach ensures the maintenance of data asset quality across interconnected data sources, supporting reliable data-driven services and robust AI application development. By addressing the need for trust in systems with diverse participants, the SAF framework promotes the quality and reliability of data assets in varied ecosystems.

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

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* (DI) and *Semantic Indicators* (SI). These categories are devised to provide a comprehensive framework for evaluating the robustness of datasets within KGs and across AI systems. The DI and SI were derived based on the literature review detailed in Section 3. Current research (Zaveri et al., 2015; Wang et al., 2021; Radulovic et al., 2017) emphasizes the importance of this distinction, highlighting that separate evaluation of intrinsic data quality and semantic richness is critical for effective data utilization in AI and linked data applications.

- *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 complete patient records. This ensures that foundational data used in AI algorithms is reliable, mitigating risks associated with poor DQ.
- *Semantic Indicators (SI)* pertain to the semantic descriptions of datasets, encompassing structured interlinking and contextual relevance. These indicators evaluate how effectively data is described and linked, similar to metadata or Linked Data (LD) standards, enhancing discoverability and usability. For instance, in a scholarly database, SI assesses the clarity and correctness of metadata, influencing data integration and retrieval across platforms.

DI would verify the accuracy, completeness, and synchronicity of bus departure times in the public transport domain, ensuring timestamps are precise and consistently formatted. This reliability is critical for AI systems in route optimization and predictive modeling.

SI would examine 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 comprehensive information that can integrate seamlessly with smart city infrastructures and deliver insightful, actionable information to end users.

Together, these indicators form the backbone of our methodology, addressing the dual aspects of DQ

(Kilkenny and Robinson, 2018) 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 aligns with the strategic goals of semantic interoperability and ensures that data and its contextual framework are optimized for cross-domain applications.

Data Quality Standards and ISO Standard 25012. Within the *ISO Standard 25012*[1], 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 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. Rather than diving deeply into individual metrics, this strategic focus on dimensions positions our research as a foundational reference point, facilitating subsequent detailed studies to refine 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 qualified through various metrics. Our work focuses 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.

Existing DQ dimensions and standards, such as *ISO 25012* and *ISO 8000-2*, play a crucial role in evaluating and assuring DQ in various contexts. *ISO 25012*, titled ”Data Quality Model,” provides a framework for assessing data quality 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 accessible 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 explicitly tailored to the complexities of KGs or 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 shared 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 fail to address the specific needs of emerging data architectures (Zhang et al., 2021).

Data Ecosystems. An example of such distributed environments are data ecosystems, a concept 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

[1] https://www.iso.org/standard/35736.html

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 realizing 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 accompanying 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 employs best practices using URIs and RDF for machine-readable, interoperable data distribution on the Web (Zaveri et al., 2015; Ji et al., 2022). KGs enhance LD by forming a graph-based knowledge base with interconnected entities, enabling advanced analytics and AI applications (Ban et al., 2024; Pan et al., 2017). KGs improve metadata quality, crucial for accurate data descriptions and interoperable AI systems, thus enhancing reliability and trustworthiness (Pan et al., 2024).

While quality models for KGs and LD exist, they lack standardization and consensus on dimensions and metrics (Zaveri et al., 2015; Radulovic et al., 2017). The dynamic nature of LD requires innovative assessment methods for scalable, high-quality data exchange across systems (Zaveri et al., 2015).

3 METHODOLOGY

Our methodology is based on a Structured Literature Review (SLR) and subsequent analysis of existing frameworks. We derived new high-level DAQ dimensions for DI and SI by anchoring our clustering process to the ISO 25012 framework (ISO25012, 2008), ensuring alignment with recognized standards.

Two scientists (RS1 and RS2) from different institutions independently conducted the SLR following (Moher et al., 2010; Kitchenham, 2004) to mitigate bias. This approach identifies open issues and contributes to a common conceptualization. We summarize established ISO 25012 dimensions and various methods for evaluating KGs and LD, proposing a framework for assessing data assets within a data ecosystem.

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

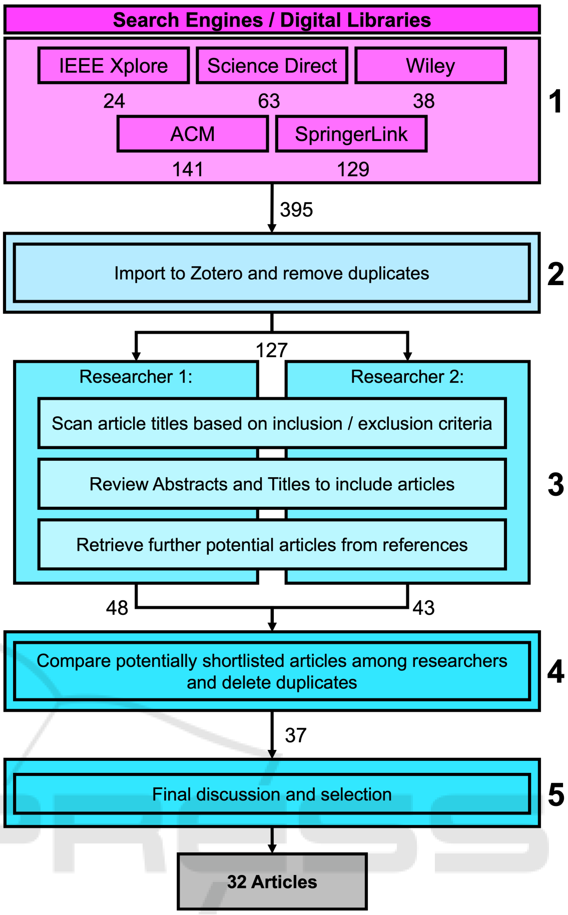
Search Engines / Digital Libraries
IEEE Xplore
Science Direct
Wiley
24
63
38
ACM
SpringerLink
141
129
1
395
Import to Zotero and remove duplicates
2
127
Researcher 1:
Researcher 2:
Scan article titles based on inclusion / exclusion criteria
Review Abstracts and Titles to include articles
Retrieve further potential articles from references
3
48
43
Compare potentially shortlisted articles among researchers and delete duplicates
4
37
Final discussion and selection
5
32 Articles

Figure 1: Process of the systematic literature search.

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

Titles, abstracts, and full texts were filtered using predefined inclusion and exclusion criteria, followed by a backward search to identify further relevant studies. This process, including the number of articles found at each step, is detailed in Figure 1. We identified 395 papers, removing duplicates and including only peer-reviewed, accessible, English or German studies in computer science.

Subsequent steps focused on filtering and refining data assets. Both reviewers independently evaluated the titles of 127 articles and reviewed abstracts to identify suitable studies, excluding those not focused on metadata or knowledge graphs or lacking a

Table 1: Presentation of the 31 articles identified as a result of the systematic literature search.

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)

methodology for quality assessment. Discrepancies were resolved by consensus or detailed review, resulting in a final list of 48 articles for RS1 and 43 for RS2. A snowballing approach ensured comprehensive coverage by checking references, using Google Scholar's ”Cited by” feature, and searching for related articles. This process identified 10 additional relevant articles. Finally, 31 articles were selected, listed in Table 1.

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 2. Throughout this process, we systematically extracted and analyzed the quality dimensions mentioned in various papers. We consolidated these dimensions into the aforementioned groups through iterative clustering and synthesis, 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 distinguishes between Data Indicators (DI), Semantic Indicators (SI), and Hybrid Indicators (HI). The first two dimensions were introduced in Chapter 1. Hybrid indicators combine DI and SI to assess the suitability of data for cross-system use, applying to both data and semantic descriptions.

Figure 2 shows groups with inherent quality dimensions (e.g., presentation) and system-dependent dimensions (e.g., system flexibility). Overall, these groups align well with ISO 25012, except for ”system flexibility,” which lacks a corresponding group in the ISO standard. We extracted 31 dimensions and 10 superordinate groups from the literature, comprising 5 DI, 6 SI, and 20 HI, as well as 17 inherent, 9 system-dependent, and 5 inherent and system-dependent dimensions. This selection allows users to choose dimensions relevant to their specific application. Clustering quality dimensions and dependencies enables more precise selection.

4.1 SAF Scores

The SAF scores are calculated to ensure comparability between different DAs and to meet the individual needs of users and departments. These scores

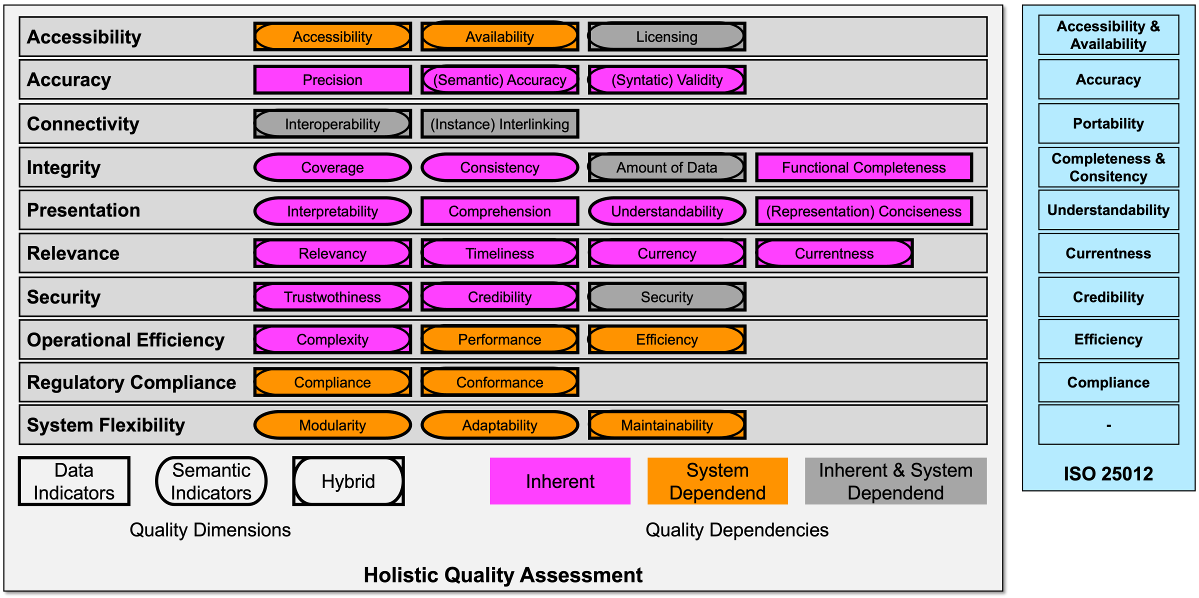

Figure 2: Holistic Quality Assessment overview: The dimensions are classified under overarching groups, reflecting their inherent and system-dependent qualities, and are further mapped onto the ISO 25012 standard.

allow users to prioritize whether SI or DI is more important for their specific application, enabling customized weighting.

The SAF scores are based on a systematic and mathematically sound method, utilizing the dimensional assignments from Figure 3. For each dimension, an appropriate metric is collected and calculated for the corresponding DA. The objective is to combine the metrics for DI and SI such that $\text{SAF} = \text{DI} + \text{SI}$. 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. 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, labeled 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 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:

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

Initially, the weight factors w_{DI_k} and w_{SI_k} are set to 1, ensuring balanced weighting when no user-defined weights are applied.

Figure 3 illustrates the methodological rigor of the SAF framework with three scenarios: balanced, semantics-centered, and data-centered evaluations. Users can define the weighting to reflect the focus of their application.

4.1.1 Example: Complex Data Asset Evaluation

Consider a complex DA distributed across multiple sites, such as a healthcare data system integrating patient records from various hospitals. Each site collects

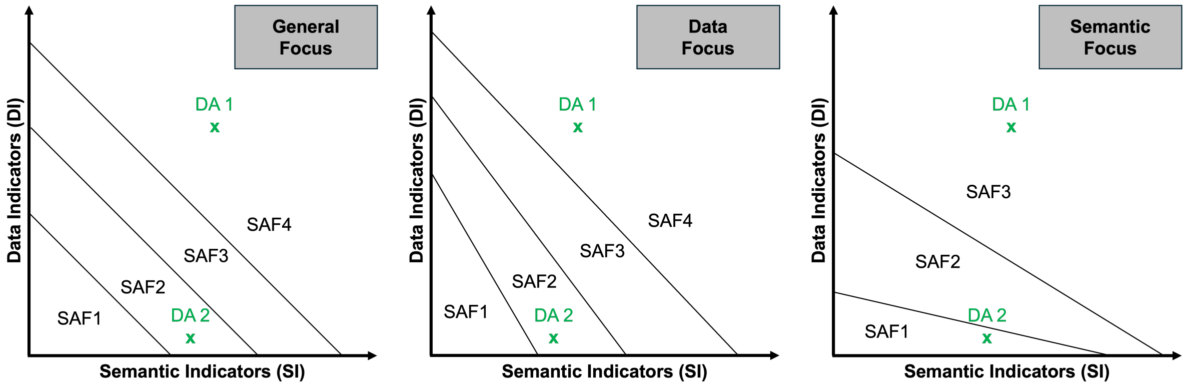

Figure 3: The SAF assessment framework is a comprehensive approach for evaluating heterogeneous DA in three distinct forms. The three diagrams illustrate the distribution of SAF levels based on the assessment focus: general (left), data-oriented (center) and semantic-oriented (right). Furthermore, the 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.

data including patient diagnostics, treatment records, and outcomes. The evaluation involves:

- Data Indicators (DI): Accuracy of diagnostic codes, completeness of treatment records, consistency of patient outcomes across sites.
- Semantic Indicators (SI): Clarity of metadata descriptions, interlinking of patient records, contextual relevance of treatment data.

For DI, metrics such as precision of diagnostic codes (c_1), completeness of records (c_2), and consistency (c_3) are collected. Suppose $\overline{C}_{DI}$ for these metrics is calculated as:

$$\overline{C}_{DI} = \frac{1}{3}(c_1 + c_2 + c_3)$$

Assuming DI_1, DI_2, and DI_3 are the weights for these metrics, the total DI score is:

$$\text{Total DI} = DI_1 \cdot c_1 + DI_2 \cdot c_2 + DI_3 \cdot c_3$$

For SI, metrics such as metadata clarity (c_4), interlinking (c_5), and contextual relevance (c_6) are collected. Suppose $\overline{C}_{SI}$ for these metrics is:

$$\overline{C}_{SI} = \frac{1}{3}(c_4 + c_5 + c_6)$$

Assuming SI_1, SI_2, and SI_3 are the weights for these metrics, the total SI score is:

$$\text{Total SI} = SI_1 \cdot c_4 + SI_2 \cdot c_5 + SI_3 \cdot c_6$$

Weight factors can be adjusted based on user priorities. For instance, in a scenario prioritizing data accuracy over metadata clarity, w_{DI_k} could be set higher than w_{SI_k}. Finally, the SAF score, incorporating these weights, is:

$$\text{SAF} = \sum w_{DI_k} \cdot V_{DI_k} + \sum w_{SI_k} \cdot V_{SI_k}$$

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 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.

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.

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.

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.

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.

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.

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.