The Five Facets of Data Quality Assessment

Sedir Mohammed[1], Lisa Ehrlinger[1], Hazar Harmouch[2], Felix Naumann[1], Divesh Srivastava[3]
[1]Hasso Plattner Institute, University of Potsdam, Germany
[2]University of Amsterdam, The Netherlands
[3]AT&T Chief Data Office, USA
sedir.mohammed@hpi.de, lisa.ehrlinger@hpi.de
h.harmouch@uva.nl, felix.naumann@hpi.de, divesh@research.att.com

ABSTRACT

Data-oriented applications, their users, and even the law require data of high quality. Research has divided the rather vague notion of data quality into various dimensions, such as accuracy, consistency, and reputation. To achieve the goal of high data quality, many tools and techniques exist to clean and otherwise improve data. Yet, systematic research on actually assessing data quality in its dimensions is largely absent, and with it, the ability to gauge the success of any data cleaning effort.

We propose five facets as ingredients to assess data quality: *data*, *source*, *system*, *task*, and *human*. Tapping each facet for data quality assessment poses its own challenges. We show how overcoming these challenges helps data quality assessment for those data quality dimensions mentioned in Europe's AI Act. Our work concludes with a proposal for a comprehensive data quality assessment framework.

1 The Many Dimensions of Data Quality

Data quality (DQ) has been an important research topic for the last decades [10, 43, 62], reflecting its critical role in all fields where data are used to gain insights and make decisions. A manifold of DQ dimensions exists that regard data and their properties from various perspectives and contribute to understanding and characterizing the complex nature of data [10, 62].

The high demand for DQ. Especially in the fast-moving landscape of *artificial intelligence* (AI), where data plays a pivotal role, the significance of DQ is dramatically increasing, so much so that literature calls this trend a paradigm shift from a model-centric view to a data-centric one [64]. Data-centric AI emphasizes the data and their impact on the underlying model [44, 45, 63]. Literature showed that DQ, with its various dimensions, significantly influences prediction accuracy [24, 36, 40, 45]. Domain-specific particulars provide a context that imposes specific requirements on DQ assessment, such as

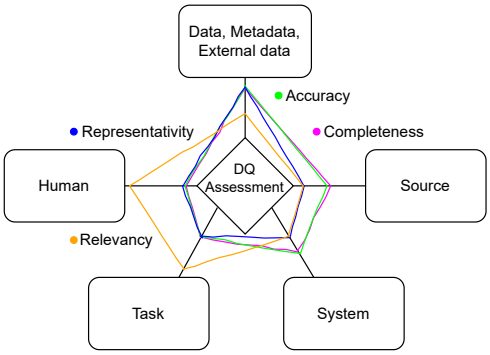

Figure 1: The five *facets* of DQ assessment and exemplary characteristics for DQ dimensions.

the *Health Insurance Portability and Accountability Act* (HIPAA), which focuses on privacy but promotes DQ dimensions, such as accuracy and completeness for ensuring trust [2].

Such requirements have also become part of regulation, as in the *General Data Protection Regulation* (GDPR) [25] and the EU *AI Act* [22]. For instance, the AI Act mentions in *Article 10* the DQ dimensions representativity, accuracy (free of errors), completeness and relevancy [22]. Similar initiatives to regulate DQ and AI are also being made by the United States [31] and China [52], which underlines the international interest in the topic of DQ.

Examining DQ is by no means just an academic problem [12]. Industry is also concerned about the impact of DQ on business [53]. Companies have shifted from internal "data gazing" [37] to hiring auditing firms for quality assurance. The literature shows that poor DQ has an enormous economic impact on organizations, either through loss of revenues or through additional internal costs [41, 50].

In addition to recognizing the relevance of DQ and understanding it in terms of the various dimensions, the goal is to improve DQ by cleaning the data. Yet, quality cannot be improved if it can-

not be measured [60]: we need concrete *assessment methods* to evaluate DQ in individual dimensions. Batini et al. [12] define DQ assessment as the measurement of DQ and the comparison with reference values for diagnosing it. As such, apart from the pure measurement of DQ, assessment includes classifying whether the measured quality is *sufficient* (or "fit") for the underlying task. *Measuring* vs. *judging* whether the measured DQ suffices for a task at hand are challenges of rather different natures.

Vision statement. Given a dataset, a use case (task specification), a set of DQ dimensions, and their formal definitions, our goal is to develop effective and efficient assessment procedures for each DQ dimension. These procedures should compute values that accurately align with the formal definitions.

Mission statement. To achieve the vision, we want to identify facets upon which assessment procedures across DQ dimensions depend. These facets enable individual dimensions to benefit from solutions to shared assessment challenges and streamlined implementation of assessment procedures.

Contribution. This paper proposes a new perspective on DQ research: through the lens of so-called *facets*. We discuss five *facets* of DQ assessment as potential sources for DQ information. Each *facet* presents its own set of challenges and opportunities. To overcome the challenges and capitalize on the opportunities, we identify a wide range of technologies that require cross-community expertise. We envision a thorough implementation of these technologies by different research communities. The ultimate goal is the integration of these technologies into a robust framework. We advocate for the development of a *DQ assessment framework* to accurately and efficiently measure all dimensions of the DQ. The framework enables (1) the integration of deeper data profiling methods [5], (2) compliance with given regulations, (3) enhancement of data cleaning, as well as (4) *judging* whether DQ meets user expectations. While this paper focuses on structured data, we believe it can also be extended to semi-structured or unstructured data.

2 Data Quality Assessment by Facets

Data quality assessment in its variety of dimensions [9, 43] poses many definitional, computational, and organizational challenges. We propose five *facets* (see Figure 1) that serve as foundation for DQ assessment: (i) the *data* itself, including metadata and external data; (ii) the *source* of the data; (iii) the *system* to store, handle, and access the data; (iv) the *task* to be performed on the data; and (v) the *humans* who interact with the data. These five facets are inspired by the stages of a typical data life cycle [59]: all relevant components of each stage can be mapped to one or more facets.

Each *facet* poses its own challenges and opportunities for future research. We hypothesize that addressing these challenges per *facet* adresses problems that arise from more than one DQ dimension. We propose *facets* as an additional layer to structure DQ research, allowing all dimensions involved in the assessment of a specific *facet* to benefit simultaneously from solving these challenges.

In the following, we define and discuss each of the five *facets* and their key challenges. We list exemplary DQ dimensions (see [39] for definitions) that specifically benefit from resolving these challenges.

2.1 The Data Facet

Raw data values are intended to represent real-world concepts and entities. The data facet includes the data semantics and their digital representation. It also includes metadata, such as schema information and other documentation, and any assessment-relevant external knowledge (as data), like a knowledge base (e.g., DBpedia [35]) to validate data. The data facet encompasses all challenges related to the data being assessed, its metadata, and external data.

As data occur in different *granularity* (e.g., values, records, columns), DQ assessment must identify the necessary level of detail and devise quality-metric aggregation methods to cross levels of granularity. Also, *metadata*, such as schema and data types, should be available and of high quality itself. When external knowledge is needed, challenges arise in discovering, matching, and assessing the quality of *reference data*. If data is encrypted, it cannot be assessed directly, so DQ assessment must handle *encrypted data* and, in case of distribution, also work in a *federated setup*.

In the following, we highlight two well-known DQ dimensions (mentioned in the AI Act) where the data facet is involved in the assessment.

Example DQ Dimensions

Accuracy: Typical metrics to assess accuracy require *reference data* to determine how closely the data matches the reality.

Completeness: Placeholders represent missing values, using either obvious placeholders like "NaN" or less obvious placeholders. The assessment needs *metadata* that contains information about the placeholder representation.

2.2 The Source Facet

The source of data represents a logical perspective. This *facet* encompasses evaluating the data generation and collection processes, as well as assessing the source's integrity and organizational compliance. The main aspect of the source facet is data *provenance*, which includes information on the origins, providers, and other organizations involved in creating and transforming the data [29].

One key challenge is ensuring *data lineage* traceability, including the data origin and its transformations [26]. Additionally, a *process-oriented view* is crucial, which includes evaluating the transformation process and the credibility of annotating agents in the DQ assessment. It is also important to consider the *time range* for assessing reliability over time; longer histories provide a more comprehensive view, while shorter intervals highlight recent changes.

Example DQ Dimensions

Reputation: The assessment requires evaluating a data source's credibility and reliability, and thus, considering historical reliability with *data lineage*.

Believability: The key challenge is to verify the data origin (*time range*), source transformations (*data lineage*), and involved entities (*process*).

2.3 The System Facet

The system facet pertains to a physical perspective, including the infrastructure and technology for storing, handling, and accessing the data. It also covers the system's technical compliance with legal and regulatory requirements, ensuring adherence to necessary data management standards.

The system facet raises challenges, such as *clarity* or *auditability*. The *clarity* includes documenting the system's architecture, data processing capabilities, interoperability with other systems, security features, and user interface aspects. *Auditability* is crucial to verify compliance with regulations, such as data deletion and security standards.

Example DQ Dimensions

Recoverability: Assessing the ability to restore a prior state of the data requires knowledge about the file system, backup procedures (*clarity*) and long-term storage regulations (*auditability*).

Portability: The key challenge is to understand the storage system, including file formats (*clarity*) and interoperability standards (*auditability*).

2.4 The Task Facet

The task facet pertains to the specific use case and the context in which the data is employed. Thus, it inherently aligns with the "fitness for use" definition of DQ [10, 62]. The task influences which parts of the data (e.g., columns, tuples) are considered and how well they represent the real world.

The task facet poses challenges regarding the *relevance* of the data, including the identification of relevant attributes and tuples. Also, the *risk* of the task, according to the AI Act, which defines minimal-, limited-, high- and unacceptable-risk AI systems, can determine the way DQ is assessed [1]. Higher risk categories require more stringent DQ assessment methods, including strict validation processes and documentation, to ensure compliance.

Example DQ Dimensions

Timeliness: The key challenge is defining an acceptable timeframe for tasks and to classify how long data are considered up-to-date or *relevant*.

Relevancy: The assessment involves balancing the need for complete information (*relevance*) against the risk of including unnecessary data that can violate legal requirements (*risk*).

2.5 The Human Facet

The human facet introduces a subjective view, while including the diverse groups that interact with the data, perform the task, and interpret the results. It aligns DQ with the specific needs and contexts in which users operate. Some DQ dimensions (e.g., relevancy, believability), require user surveys to assess experiences and challenges in handling the data. This subjective perspective makes it challenging to fully automate the assessment. The human facet presents challenges such as the need to *design surveys* that capture a range of expertise levels, or also the consideration of the *intent* of different user groups and their perspectives (e.g., developers, customers).

Example DQ Dimensions

Ease of manipulation: Since manipulability can impact accessibility positively and data integrity negatively, the assessment must consider the users *intent* of manipulation.

Relevancy: Determining relevant data varies by user perspective (*intent*). The evolving nature of relevancy with changing user needs, market trends, and legal standards complicates maintaining up-to-date assessments (*survey design*).

3 Facet Application

In the previous section, we listed example DQ dimensions per *facet*, for which the considered *facet* is involved in the assessment. Of course, the participation of the *facets* in assessing a DQ dimension occurs to varying degrees. We use a three-level system ("++", "+", "-") to indicate a *facets*' participation: "++" for strong involvement, "+" for medium, and "-" for low to no involvement. We determined the involvement of the *facets* through several discussion rounds among all authors until we reached a consensus. When determining the involvement of *facets*, we deliberately voted in favor of an objective and automatic assessment and thus tried to minimize the involvement of the human facet. Although DQ is often defined as "fitness for use" [62] the task facet is not necessarily included in the assessment.

In the following, we discuss the *facet* involvement and implications with respect to specific technologies for each DQ dimension from the AI Act: accuracy (free of errors), representativity, completeness, and relevancy [22] (see Figure 1). Additionally, we include a discussion on accuracy and relevancy as examples to illustrate why certain facets are not involved in the assessment.

3.1 DQ Dimension: Accuracy

Definition Accuracy describes the correspondence between a phenomenon in the world and its description as data [10].

Data	Source	System	Task	Human
++	+	+	-	-

The data facet is the primary contributor to the assessment of accuracy. Further aspects from the source facet (e.g., data provenance) and the system facet (e.g., storage technologies) are also relevant. Conversely, the task and human facets are less relevant: accuracy can be measured on a purely objective level, considering factual correctness and alignment with truth.

The literature established several metrics to assess accuracy [12, 28]. Most metrics require reference data, which corresponds to the data facet. To address this challenge, the reference data must be defined (e.g., its level of detail) and collected. Open data platforms, such as Kaggle [3] or general knowledge bases (e.g., Wikidata [4], DBpedia [35]), are well suited to collect a variety of data. To make use of such external data, they must be matched with the data using *schema matching* approaches [11, 19, 30, 49], which must handle different formats to process reference data from different sources [38]. This is particularly challenging with data that include *natural language*, demanding methods for semantic and syntactic processing, potentially using *large language models* [23].

In cases where access to such data platforms is too expensive or where no relevant data of sufficient quality could be found, *semantic web technologies* combined with *information retrieval approaches* would allow gathering data from the web, as external data for assessment [14, 27, 55].

In terms of the source facet, error detection and cleaning methods, such as NADEEF [18] or HoloClean [51], can be used to identify and correct data errors. The transformations applied must be clearly documented in the metadata (see Section 3.3).

The system in which the data is stored might be responsible for erroneous values caused by system failures, such as crashes or bugs. Thus, the system can lose information when saving new values, such as decimal points. Consequently, system robustness, data replication, and recovery processes must be included in the metadata. These aspects require a cataloging system to format the metadata in a machine-readable format (see also Section 3.3).

The system in which the data and metadata are located must ensure that access to them aligns with the relevant privacy provisions. If the data owner grants consent, where the consent information can also be part of the cataloging system, a partial decryption can be performed. Alternatively, encryption schemes such as *homomorphic encryption* can be used to assess and process the data/metadata while they are encrypted [6]. Compliance with privacy provisions is independent of the assessment of specific DQ dimensions.

3.2 DQ Dimension: Representativity

Definition Representativity aims to ensure that the characteristics of the reference data are present in the considered data [17, 33].

Data	Source	System	Task	Human
++	-	-	-	-

The data facet is the main contributor to the assessment of representativity.

Similar to accuracy, metrics to assess representativity require information on the reference data [15, 17]. Thus, the reference data must first be defined to establish a baseline for comparison in the assessment. In contrast to accuracy, assessing representativity does not require the complete reference data – summary statistics, respectively, data distributions of the attributes, are often sufficient. Depending on the data source, *metadata* may already contain

information about summary statistics and distributions. These metadata must be in a *structured format* (e.g., JSON or RDF) to enable automated access and further processing. Beyond uniform formatting, information must follow a *uniform schema* and *vocabulary* across data sources to ensure interoperability. The use of an *ontology* (e.g., Croissant [7] or DSD [21]) would ensure a standardized schema and vocabulary, improving interoperability.

Still, the data must be matched with the given data, even if it is in an aggregated format. But, *data matching* with less data is an easier task because there are fewer records and attributes to compare, reducing computational complexity and processing time. This simplifies schema matching, data cleaning, and handling diverse formats, leading to fewer errors and more straightforward and accurate matching criteria. Nevertheless, if the external data sources do not provide this information, the technologies the assessment requires to obtain and match the reference data overlap with the technologies mentioned in the context of accuracy.

3.3 DQ Dimension: Completeness

Definition Completeness refers to the extent to which data, including entities and attributes, are present according to the data schema [46].

Data	Source	System	Task	Human
++	+	+	-	-

When focusing on entry-level completeness, the data facet is primarily involved in the assessment; the source and system facets partially.

Since completeness represents the presence of the data, its assessment requires the measurement of missing values. While `null` or conventional placeholders like "NaN" for missing values are easily identified, more research is required to also identify so-called "hidden missing values" like "-99", "EMPTY", or default values [13, 48]. Identifying these hidden missing values can either be done through prior knowledge (in terms of metadata and sophisticated *Data Catalogs* [20] or, particularly suited for the ML context, with *Data Cards* [47]) or alternatively learned with ML models taking into account the context. Placeholders can differ for each data source or be domain-specific, which is why strict documentation is important. In addition, transformations on missing values, like deleted records or applied imputation strategies, must also be part of the metadata.

Similar to accuracy, the system in which the data is located might cause missing values, e.g., due to hardware failure. In the context of completeness, the system can lose data or fail to store new values, again necessitating metadata for recovery processes.

3.4 DQ Dimension: Relevancy

Definition Relevancy describes the extent to which the data are applicable and helpful for a given task [62].

Data	Source	System	Task	Human
+	-	-	++	++

While the task and the human facet mainly support the assessment of relevancy, the data facet is also involved. Conversely, the source and system facets are less relevant, as relevancy is solely determined by the data's usefulness for fulfilling a specific task, regardless of how or where it was created or stored.

To assess relevancy, stakeholders must *define* the given task, requiring domain experts to incorporate best practices and to understand the task's intricacies. Given the task, stakeholders and experts have to assess the relevancy of individual attributes and tuples. Alternatively, *statistical methods* can assess relevancy, e.g., Shapley or LIME calculate the feature importance to determine each feature's contribution to an ML model's prediction [56, 57, 61]. As feature importance is computationally complex, manual assessment might still be needed.

This manual assessment can be supported with *data profiling* [42] methods, comprising several tasks, such as, the automatic identification of distributions, functional dependencies, or data types. Based on the gathered information, experts can define domain- and task-specific criteria to assess the relevance of individual attributes and tuples using a *rating system* (e.g., Likert scale). Depending on the underlying task and its criticality, a larger-scale user study must be conducted to reflect various stakeholders and their perspectives. These surveys must follow the principles of good user *survey design* principles [34] and their creation should be independent from a given dataset to ensure an automated reuse for new or changed datasets.

4 Vision: A DQ Assessment Framework

In previous sections, we explored the challenges associated with different *facets* of DQ assessment and their applications to DQ dimensions. To promote this fresh look on DQ research, we envision a *DQ assessment framework* that implements the assessment methods along the *facets*. For instance, relevancy and timeliness intersect within the task facet: the specification of the downstream task (e.g., ML-based classification) determines whether the data

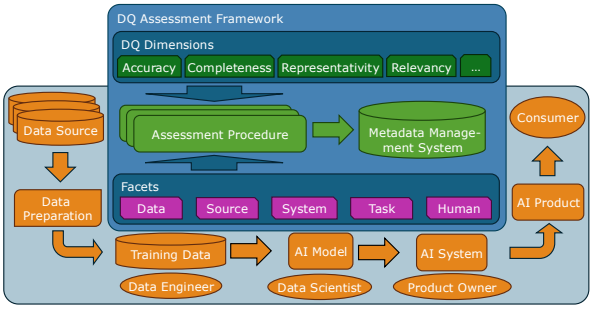

Figure 2: DQ assessment for an AI pipeline.

is relevant and also sufficiently up-to-date. The assessment of both dimensions benefits from that task specification.

Figure 2 shows the DQ assessment framework in the context of an AI pipeline. As part of this pipeline, data passes through various stages from its creation to the final product delivered to the customer. We can map the facets to these different stages of the pipeline. Thus, our proposed framework and the concept of facets are integrated into the AI pipeline: The data, in its digital representation (data facet) originate from various sources. A data engineer must prepare them using data preparation techniques, where all transformations must be traceable (source facet). The prepared data serve as training data, used by a data scientist to train an AI model, constituting a task (task facet). All these tasks can be deployed in an AI system (system facet), managed by a product owner, which in turn, can be part of an AI product that is delivered to customers. Finally, the various involved individuals should also be part of the DQ assessment (human facet). The assessment of each DQ dimension, together with the *facet's* participation, results in a dedicated assessment procedure.

We conducted an initial analysis of the participation of the *facets* per DQ dimension [39]. Apart from the facet-specific challenges to measure DQ in its various dimensions, building a framework that supports DQ measurement and management along the entire pipeline gives rise to further challenges:

Efficiency. The assessment effort and time should be low from a user perspective [8]. Data consumers might be unable or unwilling to wait for assessment results, and experts might not have much time to complete questionnaires or help in DQ assessment.

Explainability. Due to their ambiguity [32], assessment results must be explainable to consumers. In addition, the results should be traceable to their root cause, enabling measures to improve quality.

Metadata Management. Deploying the DQ assessment procedure requires an effective mechanism to store and query vast, diverse metadata (see *Metadata Management System* in Figure 2). An example solution and its challenges are discussed in [58].

5 Related Work

This section discusses representative works on DQ assessment and compares them to our fresh look through the lens of *facets*. Over the last decades, a number of DQ assessment frameworks have been proposed [12, 16]. For instance, Stvilia et al. [60] identified various sources for DQ assessment and distinguished intrinsic, relational, and reputational information quality. Batini et al. [12] divide the assessment into different phases and discuss metrics for DQ dimensions. Pipino et al. [46] present an approach combining subjective and objective DQ assessment results. In their vision paper, Sadiq et al. identify two dimensions to empirical DQ management [54]: the *metric* type (intrinsic vs. extrinsic) and the method *scope* (generic vs. tailored). They encourage the community to regard DQ beyond what we call the data facet – this paper follows that call. Other works [9, 10, 46] discuss challenges associated with specific DQ dimensions, e.g., the need for external data to assess accuracy [9].

In summary, many existing works implicitly mention individual facets (e.g., the human or the data facet) and the impact of their challenges on the assessment of DQ dimensions. However, so far, a unified view on how to address these different aspects was missing. We believe that addressing common DQ challenges per *facet* enables researchers the exploration of many DQ dimensions jointly.

6 Conclusion

We propose five assessment *facets* as foundational ingredients to assess *data quality* (DQ) and outline specific challenges and opportunities for each *facet*, highlighting the complexity of DQ assessment. We suggest how to overcome these challenges for the DQ dimensions mentioned in the AI Act as examples. Finally, we envision a DQ assessment framework that implements various methods to assess the DQ dimension through the lens of the *facets*.

Acknowledgements

This research was partially funded by the KITQAR project, supported by Denkfabrik Digitale Arbeitsgemeinschaft im Bundesministerium für Arbeit und Soziales (BMAS).

References

[1] EU AI Act: first regulation on artificial intelligence, 2023. URL https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence. (Last accessed: 2024-07-25).

[2] HIPAA privacy rule to support reproductive health care privacy, 2024. URL https://www.federalregister.gov/documents/2024/04/26/2024-08503/hipaa-privacy-rule-to-support-reproductive-health-care-privacy. (Last accessed: 2024-07-25).

[3] Kaggle: Your machine learning and data science community, 2024. URL https://www.kaggle.com/. (Last accessed: 2024-07-15).

[4] Wikipedia, the free encyclopedia, 2024. URL https://www.wikipedia.org/. (Last accessed: 2024-07-15).

[5] Ziawasch Abedjan, Lukasz Golab, and Felix Naumann. Profiling relational data: a survey. *VLDB Journal*, 24(4):557–581, 2015. doi: 10.1007/S00778-015-0389-Y.

[6] Abbas Acar, Hidayet Aksu, A. Selcuk Uluagac, and Mauro Conti. A survey on homomorphic encryption schemes: Theory and implementation. *ACM Computing Surveys*, 51(4): 79:1–79:35, 2018. doi: 10.1145/3214303. URL https://doi.org/10.1145/3214303.

[7] Mubashara Akhtar, Omar Benjelloun, Costanza Conforti, Pieter Gijsbers, Joan Giner-Miguelez, Nitisha Jain, Michael Kuchnik, Quentin Lhoest, Pierre Marcenac, Manil Maskey, Peter Mattson, Luis Oala, Pierre Ruyssen, Rajat Shinde, Elena Simperl, Goeffry Thomas, Slava Tykhonov, Joaquin Vanschoren, Jos van der Velde, Steffen Vogler, and Carole-Jean Wu. Croissant: A metadata format for ml-ready datasets. In *Proceedings of the International Conference on Management of Data (SIGMOD)*, pages 1–6. ACM, 2024. doi: 10.1145/3650203.3663326. URL https://doi.org/10.1145/3650203.3663326.

[8] Donald P Ballou, InduShobha N Chengalur-Smith, and Richard Y Wang. Sample-based quality estimation of query results in relational database environments. *IEEE Transactions on Knowledge and Data Engineering*, 18(5):639–650, 2006.

[9] Carlo Batini and Monica Scannapieco. *Data quality: concepts, methodologies and techniques*. Data-centric systems and applications. Springer, 2006. ISBN 978-3-540-33172-8 978-3-642-06970-3.

[10] Carlo Batini and Monica Scannapieco. *Data and Information Quality: Dimensions, Principles and Techniques*. Springer Berlin Heidelberg, 2016. ISBN 978-3-319-24104-3.

[11] Carlo Batini, Maurizio Lenzerini, and Shamkant B. Navathe. A comparative analysis of methodologies for database schema integration. *ACM Computing Surveys*, 18(4): 323–364, 1986. doi: 10.1145/27633.27634. URL https://doi.org/10.1145/27633.27634.

[12] Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, 41(3):16:1–16:52, 2009. doi: 10.1145/1541880.1541883. URL https://doi.org/10.1145/1541880.1541883.

[13] Michal Bechny, Florian Sobieczky, Jürgen Zeindl, and Lisa Ehrlinger. Missing data patterns: From theory to an application in the steel industry. In *Proceedings of the International Conference on Scientific and Statistical Database Management (SSDBM)*, page 214–219, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384131. doi: 10.1145/3468791.3468841. URL https://doi.org/10.1145/3468791.3468841.

[14] Bettina Berendt, Andreas Hotho, and Gerd Stumme. Towards semantic web mining. In Ian Horrocks and James A. Hendler, editors, *Proceedings of the International Semantic Web Conference (ISWC)*, volume 2342 of *Lecture Notes in Computer Science*, pages 264–278. Springer, 2002. doi: 10.1007/3-540-48005-6_21. URL https://doi.org/10.1007/3-540-48005-6_21.

[15] Marcin Budka, Bogdan Gabrys, and Katarzyna Musial. On accuracy of PDF divergence estimators and their applicability to representative data sampling. *Entropy*, 13(7):1229–1266, 2011. doi: 10.3390/E13071229. URL https://doi.org/10.3390/e13071229.

[16] Corinna Cichy and Stefan Rass. An overview of data quality frameworks. *IEEE Access*, 7: 24634–24648, 2019.

[17] Line H. Clemmensen and Rune D. Kjærsgaard. Data representativity for machine learning and AI systems. *CoRR*, abs/2203.04706, 2022. doi: 10.48550/ARXIV.2203.04706. URL https://doi.org/10.48550/arXiv.2203.04706.

[18] Michele Dallachiesa, Amr Ebaid, Ahmed Eldawy, Ahmed K. Elmagarmid, Ihab F. Ilyas, Mourad Ouzzani, and Nan Tang. NADEEF: a commodity data cleaning system. In *Proceedings of the International Conference on Management of Data (SIGMOD)*, pages 541–552. ACM, 2013. doi: 10.1145/2463676.2465327. URL https://doi.org/10.1145/2463676.2465327.

[19] AnHai Doan, Alon Halevy, and Zachary Ives. *Principles of Data Integration*. Morgan Kaufmann, 2012. ISBN 978-0-12-416044-6. doi: 10.1016/C2011-0-06130-6. URL https://doi.org/10.1016/C2011-0-06130-6.

[20] Lisa Ehrlinger, Johannes Schrott, Martin Melichar, Nicolas Kirchmayr, and Wolfram Wöß. Data catalogs: A systematic literature review and guidelines to implementation. In *DEXA Workshops Proceedings*, volume 1479 of *Communications in Computer and Information Science*, pages 148–158. Springer, 2021. doi: 10.1007/978-3-030-87101-7_15. URL https://doi.org/10.1007/978-3-030-87101-7_15.

[21] Lisa Ehrlinger, Johannes Schrott, and Wolfram Wöß. Dsd: the data source description vocabulary. In *International Conference on Database and Expert Systems Applications (DEXA)*, pages 3–10. Springer, 2023.

[22] European Parliament. Artifical inteligence act. 2024. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689. Version from 2024-06-13.

[23] Raul Castro Fernandez, Aaron J. Elmore, Michael J. Franklin, Sanjay Krishnan, and Chenhao Tan. How large language models will disrupt data management. *PVLDB*, 16(11):3302–3309, 2023. doi: 10.14778/3611479.3611527. URL https://www.vldb.org/pvldb/vol16/p3302-fernandez.pdf.

[24] Daniele Foroni, Matteo Lissandrini, and Yannis Velegrakis. Estimating the extent of the effects of data quality through observations. In *Proceedings of the International Conference on Data Engineering (ICDE)*, pages 1913–1918. IEEE, 2021. doi: 10.1109/ICDE51399.2021.00176. URL https://doi.org/10.1109/ICDE51399.2021.00176.

[25] GDPR. General data protection regulation (last accessed: 2024-02-13), 2016. URL https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02016R0679-20160504.

[26] Boris Glavic and Klaus R. Dittrich. Data provenance: A categorization of existing approaches. In *Proceedings of the Conference Datenbanksysteme in Business, Technologie und Web Technik (BTW)*, volume P-103 of *LNI*, pages 227–241. GI, 2007. URL https://dl.gi.de/handle/20.500.12116/31801.

[27] David A. Grossman and Ophir Frieder. *Information retrieval: algorithms and heuristics*. Number 15. Springer, 2nd ed edition, 2004. ISBN 978-1-4020-3004-8 978-1-4020-3003-1.

[28] Tom Haegemans, Monique Snoeck, and Wilfried Lemahieu. Towards a precise definition of data accuracy and a justification for its measure. In *Proceedings of the International Conference on Information Quality*, pages 16–16. MIT Information Quality (MITIQ) Program, 2016.

[29] Melanie Herschel, Ralf Diestelkämper, and Houssem Ben Lahmar. A survey on provenance: What for? what form? what from? *VLDB Journal*, 26(6):881–906, 2017. doi: 10.1007/S00778-017-0486-1. URL https://doi.org/10.1007/s00778-017-0486-1.

[30] Thomas N. Herzog, Fritz Scheuren, and William E. Winkler. *Data quality and record linkage techniques*. Springer, 2007. ISBN 978-0-387-69502-0. OCLC: ocn137313060.

[31] The White House. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence, 2023. URL https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

[32] Vimukthi Jayawardene, Shazia W. Sadiq, and Marta Indulska. The curse of dimensionality in data quality. In *Australasian Conference on Information Systems (ACIS)*, page 165, 2013. URL https://aisel.aisnet.org/acis2013/165.

[33] William Kruskal and Frederick Mosteller. Representative sampling, III: The current statistical literature. *International Statistical Review / Revue Internationale de Statistique*, 47(3): 245–265, 1979. doi: 10.2307/1402647.

[34] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. *Research Methods in Human Computer Interaction*. Elsevier, second edition, 2017. ISBN 978-0-12-805390-4.

[35] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. DBpedia - A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195, 2015. doi: 10.3233/SW-140134.

[36] Peng Li, Xi Rao, Jennifer Blase, Yue Zhang, Xu Chu, and Ce Zhang. Cleanml: A study for evaluating the impact of data cleaning on ML classification tasks. In *Proceedings of the International Conference on Data Engineering (ICDE)*, pages 13–24. IEEE, 2021. doi: 10.1109/ICDE51399.2021.00009. URL https://doi.org/10.1109/ICDE51399.2021.00009.

[37] Arkady Maydanchik. *Data quality assessment*. Data quality for practitioners series. Technics Publications, 2007. ISBN 978-0-9771400-2-2.

[38] Tova Milo and Sagit Zohar. Using schema matching to simplify heterogeneous data translation. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, pages 122–133, 1998. URL http://www.vldb.org/conf/1998/p122.pdf.

[39] Sedir Mohammed, Hazar Harmouch, Felix Naumann, and Divesh Srivastava. Data quality assessment: Challenges and opportunities. *CoRR*, abs/2403.00526, 2024. doi: 10.48550/ARXIV.2403.00526. URL https://doi.org/10.48550/arXiv.2403.00526.

[40] Sedir Mohammed, Lukas Budach, Moritz Feuerpfeil, Nina Ihde, Andrea Nathansen, Nele Sina Noack, Hendrik Patzlaff, Felix Naumann, and Hazar Harmouch. The effects of data quality on machine learning performance on tabular data. *Information Systems (IS)*, 132, 2025. doi: 10.1016/J.IS.2025.102549. URL https://doi.org/10.1016/j.is.2025.102549.

[41] Tadhg Nagle, Tom Redman, and David Sammon. Assessing data quality: A managerial call to action. *Business Horizons*, 63(3):325–337, 2020. ISSN 00076813. doi: 10.1016/j.bushor.2020.01.006. URL https://linkinghub.elsevier.com/retrieve/pii/S0007681320300069.

[42] Felix Naumann. Data profiling revisited. *SIGMOD Rec.*, 42(4):40–49, 2013. doi: 10.1145/2590989.2590995. URL https://doi.org/10.1145/2590989.2590995.

[43] Felix Naumann and Claudia Rolker. Assessment methods for information quality criteria. In *Fifth Conference on Information Quality (IQ 2000)*, pages 148–162. MIT, 2000.

[44] Felix Neutatz, Binger Chen, Ziawasch Abedjan, and Eugene Wu. From cleaning before ML to cleaning for ML. *IEEE Data Engineering Bulletin*, 44(1):24–41, 2021. URL http://sites.computer.org/debull/A21mar/p24.pdf.

[45] Felix Neutatz, Binger Chen, Yazan Alkhatib, Jingwen Ye, and Ziawasch Abedjan. Data cleaning and automl: Would an optimizer choose to clean? *Datenbank-Spektrum*, 22 (2):121–130, 2022. doi: 10.1007/s13222-022-00413-2. URL https://doi.org/10.1007/s13222-022-00413-2.

[46] Leo L. Pipino, Yang W. Lee, and Richard Y. Wang. Data quality assessment. *Communications of the ACM*, 45(4):211–218, 2002. ISSN 0001-0782. doi: 10.1145/505248.506010. URL https://doi.org/10.1145/505248.506010.

[47] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible AI. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FaCCT)*, page 1776–1826, New York, NY, USA, 2022. Association for Computing Machinery. doi: 10.1145/3531146.3533231. URL https://doi.org/10.1145/3531146.3533231.

[48] Abdulhakim A. Qahtan, Ahmed Elmagarmid, Raul Castro Fernandez, Mourad Ouzzani, and Nan Tang. Fahes: A robust disguised missing values detector. In *Proceedings of the International Conference on Knowledge discovery and data mining (SIGKDD)*, page 2100–2109, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355520. doi: 10.1145/3219819.3220109. URL https://doi.org/10.1145/3219819.3220109.

[49] Erhard Rahm and Philip A. Bernstein. A survey of approaches to automatic schema matching. *VLDB Journal*, 10(4):334–350, 2001. doi: 10.1007/S007780100057. URL https://doi.org/10.1007/s007780100057.

[50] Thomas C Redman. *Data quality: the field guide*. Digital press, 2001.

[51] Theodoros Rekatsinas, Xu Chu, Ihab F. Ilyas, and Christopher Ré. HoloClean: Holistic data repairs with probabilistic inference. *PVLDB*, 10(11):1190–1201, 2017. doi: 10.14778/3137628.3137631. URL http://www.vldb.org/pvldb/vol10/p1190-rekatsinas.pdf.

[52] Huw Roberts, Josh Cowls, Jessica Morley, Mariarosaria Taddeo, Vincent Wang, and Luciano Floridi. The chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation. *AI & SOCIETY*, 36(1):59–77, 2021. ISSN 0951-5666, 1435-5655. doi: 10.1007/s00146-020-00992-2. URL https://link.springer.com/10.1007/s00146-020-00992-2.

[53] Shazia Sadiq, editor. *Handbook of data quality: research and practice*. Springer, 2013. ISBN 978-3-642-36256-9. doi: 10.1007/978-3-642-36257-6.

[54] Shazia Sadiq, Tamraparni Dasu, Xin Luna Dong, Juliana Freire, Ihab F. Ilyas, Sebastian Link, Miller J. Miller, Felix Naumann, Xiaofang Zhou, and Divesh Srivastava. Data quality: The role of empiricism. *SIGMOD Record*, 46(4):35–43, 2018. URL https://doi.org/10.1145/3186549.3186559.

[55] Urvi Shah, Timothy W. Finin, and Anupam Joshi. Information retrieval on the semantic web. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*, pages 461–468, 2002. doi: 10.1145/584792.584868. URL https://doi.org/10.1145/584792.584868.

[56] Lloyd S Shapley. A value for n-person games. In *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press, Princeton, 1953.

[57] Dylan Slack, Anna Hilgard, Sameer Singh, and Himabindu Lakkaraju. Reliable post hoc explanations: Modeling uncertainty in explainability. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 9391–9404, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/4e246a381baf2ce038b3b0f82c7d6fb4-Abstract.html.

[58] Divesh Srivastava and Yannis Velegrakis. Intensional associations between data and metadata. In *Proceedings of the International Conference on Management of Data (SIGMOD)*, pages 401–412. ACM, 2007. doi: 10.1145/1247480.1247526.

[59] Victoria Stodden. The data science life cycle: a disciplined approach to advancing data science as a science. *Communications of the ACM*, 63 (7):58–66, 2020. doi: 10.1145/3360646. URL https://doi.org/10.1145/3360646.

[60] Besiki Stvilia, Les Gasser, Michael B. Twidale, and Linda C. Smith. A framework for information quality assessment. *J. Assoc. Inf. Sci. Technol.*, 58(12):1720–1733, 2007. doi: 10.1002/ASI.20652. URL https://doi.org/10.1002/asi.20652.

[61] Mukund Sundararajan and Amir Najmi. The many Shapley values for model explanation. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 119, pages 9269–9278. PMLR, 2020. URL http://proceedings.mlr.press/v119/sundararajan20b.html.

[62] Richard Y. Wang and Diane M. Strong. Beyond accuracy: What data quality means to data consumers. *J. Manag. Inf. Syst.*, 12(4):5–33, 1996. doi: 10.1080/07421222.1996.11518099. URL https://doi.org/10.1080/07421222.1996.11518099.

[63] Steven Euijong Whang, Yuji Roh, Hwanjun Song, and Jae-Gil Lee. Data collection and quality challenges in deep learning: a data-centric AI perspective. *VLDB Journal*, 32 (4):791–813, 2023. doi: 10.1007/S00778-022-00775-9. URL https://doi.org/10.1007/s00778-022-00775-9.

[64] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. Data-centric artificial intelligence: A survey. *CoRR*, abs/2303.10158, 2023. doi: 10.48550/ARXIV.2303.10158. URL https://doi.org/10.48550/arXiv.2303.10158.