“FAIR” is not enough – A Metrics Framework to ensure Data Quality through Data Preparation

Valerie Restat[1] Meike Klettke[2] Uta Störl[3]

Abstract: Data-driven systems and machine learning-based decisions are becoming increasingly important and are having an impact on our everyday lives. The prerequisite for good results and decisions is good data quality, which must be ensured by preprocessing the data. For domain experts, however, the following difficulties arise: On the one hand, they have to choose from a multitude of different tools and algorithms. On the other hand, there is no uniform evaluation method for data quality. For this reason, we present the design of a framework of metrics that allows for a flexible evaluation of data quality and data preparation results.

Keywords: data quality; metrics; evaluation; data preparation

1 Introduction

Real data is rarely error-free. To use it for analysis and to ensure the quality of the derived decisions, data preparation is necessary [Ab16]. A variety of different tools exist for this purpose, which are often combined in a preprocessing pipeline. The difficulty, however, is to choose from this range of tools and possible combinations depending on the data and the use case. More support is required for domain experts without in-depth IT knowledge. As a first step in this direction, we see the need for a framework of metrics to assess data quality. Such a framework would generate a number of advantages:

- The results of different data preparation tools become comparable.
- New solutions can be evaluated.
- Data quality is measurable and thus the quality of analyses and automated decisions can be ensured.

To the best of our knowledge, such a systematic framework does not yet exist. Hence, in this paper we propose a framework of metrics that takes into account many different aspects of data quality. Depending on the use case, this allows metrics to be selected that are suitable for the scenario in question.

After the summarization of related work in Section 2, we present this framework in Section 3. It consists of two dimensions, which are described in Section 3.1 and Section 3.2

[1] University of Hagen, Universitätsstr. 1, 58097 Hagen, Germany valerie.restat@fernuni-hagen.de
[2] University of Regensburg, Bajuwarenstr. 4, 93053 Regensburg, Germany meike.klettke@ur.de
[3] University of Hagen, Universitätsstr. 1, 58097 Hagen, Germany uta.stoerl@fernuni-hagen.de

doi:10.18420/BTW2023-61

subsequently. The measurement of the proposed metrics is defined in Section 3.3. We conclude the paper in Section 4 and present future work.

2 Related Work

Well-known principles are the FAIR principles (*Findable*, *Accessible*, *Interoperable* and *Re-usable*) for automatically finding, using, and re-using data [Wi16]. However, these do not take into account the quality of the data, which must be ensured by preprocessing.

In software engineering, there is a standard (ISO/IEC 9126) that defines software quality criteria, and various metrics have been proposed to prove each criterion. For data, there is no such standardization yet. There is some initial research work here.

A variety of different definitions exist for data quality, including timeliness, currency, accuracy and completeness, credibility, and presentation quality [CZ15; Si12]. In addition, aspects such as relevance and fairness must be considered [CZ15; Pi20; SS20]. We have provided a more detailed description of the different dimensions of data quality in [RKS22]. Since the focus of this paper is on the metrics framework, only the most important aspects are mentioned here.

The multitude of different definitions also leads to the fact that there is no standardized evaluation. Different metrics exist (e.g. [HH16] and [BM11]), but these only cover a few aspects of data quality. To the best of our knowledge, a systematic framework that takes into consideration many different data quality dimensions does not yet exist.

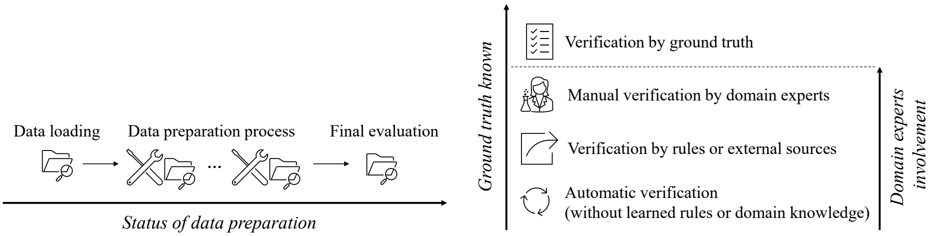
Data quality validation is also addressed in Schelter et al. [Sc18]. The authors describe a system for automating the verification of data quality. The following aspects are considered: Completeness, consistency, and accuracy. The system scales well for large data sets and provides users with a declarative API. By combining common quality constraints and custom validation code, it offers many possibilities for evaluation. However, our classification is even more comprehensive. It provides a more detailed classification of the metrics and distinguishes more precisely to what extent domain knowledge is required.

In Polyzotis et al. [Po17], challenges related to machine learning are addressed, with particular reference to the deviation between serving and training data. Our framework can be used to compare the quality of serving and training data. In future work, we also want to explore how we can further extend the framework to address the challenges that specifically arise in context of machine learning.

3 Metrics Framework

Our proposed framework consists of two dimensions, which are explained in more detail in the following subsections:

- Horizontal dimension – The status of data preparation (Section 3.1)
- Vertical dimension – The extent to which ground truth is known (Section 3.2)


Data loading
Data preparation process
Final evaluation
...
Status of data preparation
Ground truth known
Verification by ground truth
Manual verification by domain experts
Verification by rules or external sources
Automatic verification
(without learned rules or domain knowledge)
Domain experts involvement

(a) Horizontal dimension

(b) Vertical dimension

Fig. 1: Metrics framework

The horizontal dimension, shown in Figure 1a, describes the status of data preparation. Some aspects have to be considered already at the time of data collection or loading. Other aspects are continuously improved by preprocessing the data. At the end of preprocessing, a final evaluation should be performed again.

On the vertical dimension, shown in Figure 1b, a distinction is made between the extent to which ground truth is known. It is only rarely the case that ground truth is completely available. In other cases, the evaluation depends on the degree to which domain experts are involved. If these are available, a manual evaluation can be performed. Otherwise, the data quality must be checked using rules or external sources. If these are also not available, an automated check can be performed in individual cases.

In the following, the two dimensions and the corresponding metrics are explained in more detail. It is described which aspects of data quality must be taken into account. Subsequently, Section 3.3 outlines how the metrics can be measured.

3.1 Horizontal Dimension

First, the horizontal dimension is considered. It defines the point in time when the evaluation of data quality takes place.

Data loading At the beginning of data preparation, the data must be loaded. In the process, some aspects of data quality must first be checked. This includes the following aspects, which have been mentioned in Section 2:

- Timeliness
- Credibility
- Relevance

These aspects must be checked at the beginning and are prerequisites for data quality. If they are not met, an improvement by e.g. data cleaning is not possible. For example, if the data is not suitable for the application purpose, data preparation will not improve it either. New data must then be collected instead.

Data preparation process Other aspects of data quality are only achieved through the preparation process itself. Continuous evaluation is possible here. After each step, it can be checked whether the data quality has improved. At the end of data preparation, a final evaluation should be performed. In [Re22], we have already declared an extensive error classification that covers a variety of different data quality aspects. Therefore, we base these kinds of metrics on the error classification presented in [Re22], illustrated in Figure 2. The different types of errors are shown in Table 1. In the first step, we focus on single relation levels to assess the quality of individual data sets. In future work this classification should be extended.

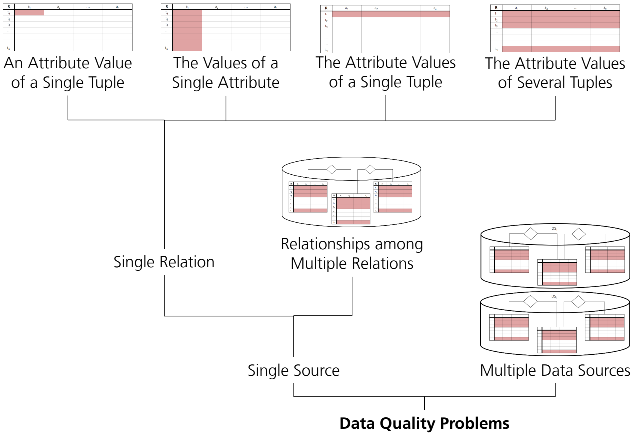

An Attribute Value of a Single Tuple
The Values of a Single Attribute
The Attribute Values of a Single Tuple
The Attribute Values of Several Tuples
Relationships among Multiple Relations
Single Relation
Single Source
Multiple Data Sources
Data Quality Problems

Fig. 2: Error classification specified in [Re22]

Final evaluation At the end of the data preparation process, all aspects from Table 1 should be checked again. As an additional point, depending on the use case, the evaluation of the presentation quality, mentioned in [RKS22], is also conceivable here.

In the following section, the metrics are described and subdivided in the vertical dimension.

3.2 Vertical Dimension

This dimension describes the extent to which ground truth is incorporated into the metrics. It also indicates how much domain experts are involved in the verification.

Tab. 1: Error types specified in [Re22]

Level	**Error Type**
An Attribute Value of a Single Tuple	Missing value
	Syntax violation
	Interval violation
	Set violation
	Misspelled error
	Inadequate value to the attribute context
	Value items beyond the attribute context
	Meaningless Value
	Erroneous entry
The Values of a Single Attribute	Uniqueness value violation
	Synonyms existence
	Outlier
	Missing Attribute
The Attribute Values of a Single Tuple	Semi-empty tuple
	Inconsistency among attribute values
	Irrelevant observation
The Attribute Values of Several Tuples	Redundancy about an entity
	Inconsistency about an entity
	Bias
	Noise

Verification by ground truth When ground truth is present, automated matching can take place. However, it is only rarely the case that ground truth is fully available. The creation may also require a high level of involvement of domain experts. In addition, it must be noted that there is no ground truth for some quality aspects. Along the horizontal dimension, this concerns mainly the aspects of data loading, as will be shown in Section 3.2.1. In the final evaluation there is also one aspect for which no ground truth can be determined, shown in Section 3.2.3.

Manual verification by domain experts If ground truth is not available, human involvement is necessary. Normally, domain experts know the data best and are therefore in the best position to judge the quality of the data. However, this process is very time consuming and therefore mostly not feasible.

Verification by rules or external sources For this reason, the next step is to permit domain experts to specify rules that can be used to check errors in the data. In some cases, it is also possible to use external sources. It should be noted that there are a number of tools that use pattern enforcement or machine learning models to detect errors. HoloDetect [He19] and Raha [Ma19] are examples for such tools. Theoretically, those models or rules learned by models can also be used for evaluation. The result of these could then be compared with the results present. However, it is difficult to state about which results are more correct. For this reason, it was not considered in the present framework.

Automatic verification (without learned rules or domain knowledge) If no rules or external sources are available, there are a number of aspects that can be checked automatically without any rules. Examples are missing values or duplicates.

In the following, the categories of the horizontal dimension, presented in Section 3.1, are listed. The individual aspects of these categories are classified on the basis of the categories of the vertical dimension just presented.

3.2.1 Data loading

On the horizontal axis, the first category is data loading. Table 2 shows the aspects mentioned and describes how an evaluation can be done.

Tab. 2: Evaluation of data quality at the time of data loading

	Evaluation method			
Data quality aspect	Automatic	Rules	Experts	Ground Truth
Timeliness	-	(✓)	✓	-
Credibility	-	(✓)	✓	-
Relevance	-	(✓)	✓	-

None of the aspects can be tested by automated evaluation without rules or domain knowledge. With rules, the following evaluations are conceivable: For timeliness, checks can be made according to the arrival time and intervals of the data. For credibility, a verification by domain experts is necessary. In addition, checks according to the range of data or accepted values are conceivable. In terms of relevance, the assessment of whether data is suitable for a specific use case, primarily depends on the goal of the analysis or prediction. It must be evaluated by a domain expert. Information about the distribution of the data, warnings of protected features and fairness metrics can be supportive in the decision-making process. However, all these rules can only support the evaluation. The final assessment must be made by a domain expert. Ground truth cannot be detected for these aspects. For example, there is no ground truth that tells whether a data set is credible. Such aspects can only be assessed by domain experts. This is why ground truth cannot be used for an evaluation of these aspects.

3.2.2 Data preparation process

The goal of data preprocessing is to achieve the best possible data quality. As described, many different approaches and tools exist for this purpose. For better comparability, it must be possible to evaluate the results of such tools. The following metrics are suitable for this purpose, but also to subject the data to continuous evaluation. Thus, the quality of the data can be tracked at any time. Table 3 shows how the quality aspects can be evaluated in the context of the data preparation process.

All these aspects can be checked with ground truth, if available. If this is not available, the next step could be a manual verification by domain experts. If domain experts can not support, all these aspects could be checked using rules or external sources. The better the rules, the more accurate the assessment of the data quality. If no rules or external sources are available, only certain data quality aspects can be checked automatically. This includes the following aspects:

Missing values can be easily detected automatically. Rules are only needed here if missing values are encoded by specific values, such as −9999.

Duplicates can – to a certain degree – be checked automatically. At column level (*Uniqueness value violation*), duplicate detection is easy, but it is usually allowed for most columns that several rows have the same value. Further information is therefore needed to decide for which columns duplicates must not exist. On row level (*Redundancy about an entity*) it can be checked automatically if identical rows exist. For rows that are merely similar, however, further information is again required for evaluation.

Outliers can – to a certain extent – be detected automatically as well. A common approach is to apply the *three sigma rule*, which states that if the data is normally distributed, 99.7% of the values will be within three standard deviations of the mean [CBK09]. However, if the values of a column do not follow the normal distribution, this rule is not applicable. Furthermore, problems arise in this context when data changes and therefore the statistical

Tab. 3: Evaluation of data quality at the time of the data preparation process

	Evaluation method			
Data quality aspect	Automatic	Rules	Experts	Ground Truth
Missing Value	✓	✓	✓	✓
Syntax violation	-	✓	✓	✓
Interval violation	-	✓	✓	✓
Set violation	-	✓	✓	✓
Misspelled error	-	✓	✓	✓
Inadequate value to the attribute context	-	✓	✓	✓
Value items beyond the attribute context	-	✓	✓	✓
Meaningless Value	-	✓	✓	✓
Erroneous entry	-	✓	✓	✓
Uniqueness value violation	(✓)	✓	✓	✓
Synonyms existence	-	✓	✓	✓
Outlier	(✓)	✓	✓	✓
Missing Attribute	-	✓	✓	✓
Semi-empty tuple	✓	✓	✓	✓
Inconsistency among attribute values	-	✓	✓	✓
Irrelevant observation	-	✓	✓	✓
Redundancy about an entity	✓	✓	✓	✓
Inconsistency about an entity	-	✓	✓	✓
Bias	(✓)	✓	✓	✓
Noise	-	✓	✓	✓

parameters may need to be re-examined. For more extensive outlier detection, additional rules or sources are needed.

Semi-empty tuples can be checked automatically. As with missing values, rules are needed only in case empty fields are encoded by values such as −9999.

Bias: Already at the time of data loading, as described in Section 3.2.1, it should be checked whether the data is relevant. This also includes the analysis of whether the data is representative for the use case. However, as described in [SS20], a bias can also be introduced into the data by preprocessing. Therefore, as a first indicator, an automated check can be made to see if the distribution of the data changes significantly during preprocessing.

All other aspects cannot be checked automatically. Further information, rules or external sources are required for an evaluation.

The framework does not yet consider error hierarchies. For example, spelling errors may also lead to a set violation. After the spelling error has been corrected, there would possibly no longer be a set violation. In future work, this should be included in the framework.

3.2.3 Final evaluation

As already mentioned, all aspects specified in Section 3.2.2 should be checked again in the final evaluation. In addition, the presentation quality can also be evaluated. This is not possible without any rules or domain knowledge, as shown in Table 4.

Tab. 4: Evaluation of data quality for the final evaluation

	Evaluation method			
Data quality aspect	Automatic	Rules	Experts	Ground Truth
Presentation Quality	-	(✓)	✓	-

For an evaluation against rules, checks can be made based on given standards or specifications. A detailed evaluation can only be done by a domain expert. Ground truth does not exist in this case.

3.3 Measurement

In the following, we will look at how the metrics presented can be applied and how a measurement of quality can be made. Along the vertical dimension, evaluation becomes more accurate once domain knowledge becomes available. Along the horizontal axis, a distinction is made depending on the time of data preparation.

3.3.1 Data loading

For all these criteria, it is only possible to check whether they are met or not. No percentage is given. For example, a data set may or may not be suitable for the use case. It is not checked whether it is 50% suitable. In future work, a more precise distinction could be made here.

3.3.2 Data preparation process

For each error listed in Table 1, the corresponding metric indicates whether that error occurs. Depending on the level of error classification, a distinction is made: In some cases, the percentage of errors per column is considered. In other cases, it is only checked whether the error occurs in the data (yes/no). The mapping depending on the level of error classification is shown in Table 5.

Tab. 5: Evaluation type per error classification level

Level	**Evaluation Type**
An Attribute Value of a Single Tuple	Percentage per column
The Values of a Single Attribute	Percentage per column (except for *Missing Attribute*)
The Attribute Values of a Single Tuple	Percentage per data set
The Attribute Values of Several Tuples	Yes/no per data set

For the error types of the first level, *An Attribute Value of a Single Tuple*, the percentage per column can be measured in each case. The same applies to the second level, *The Values of a Single Attribute*. However, the error type *Missing Attribute* is an exception. Here, the percentage per data set must be applied. This is also considered for the third level, *The Attribute Values of a Single Tuple*. For the fourth level, a percentage can no longer be calculated. It is only checked whether the corresponding error type occurs in the data set or not.

An exception is when ground truth is available. In this case, metrics such as *precision* and *recall* can be used for evaluation.

3.3.3 Final evaluation

The same aspects apply to the final evaluation. However, for the examination of the presentation quality (as with the aspects at the time of data loading) it can only be stated whether it is provided or not.

Example Each quality criterion corresponding to the rows of tables 2, 3, and 4 is assigned a metric. Thus, the framework is easily extendable. The metric for *Missing Value* is described here as an example: The error type belongs to the first level *An Attribute Value of a Single Tuple* (see Table 1). Therefore, the metric is measured by the percentage per column (see Table 5). As can be seen in Table 3, an automatic verification is possible. Thus, the percentage of missing values per column in the data set can be calculated automatically. If there is a special encoding of missing values, further rules are necessary to be able to determine the percentage correctly. If these rules are not available, a domain expert can give the correct percentage. In case that ground truth is present, precision and recall can be applied for evaluation as described.

As a result of the evaluation, a corresponding report with all described metrics would be generated. Via color coding, the reliability of the results could be marked according to the vertical dimension. Thus, a detailed evaluation of data quality would be possible. In addition to validating individual data sets, different data sets can also be compared, for example to contrast the quality of serving and training data in context of machine learning.

4 Conclusion and Future Work

To measure data quality and evaluate data preparation tools, we created a framework of metrics. This can be used flexibly, depending on the extent to which ground truth is available or domain experts are available for verification. It also considers the time of the evaluation. Different aspects need to be checked at the beginning of data preprocessing rather than during the process itself. It was shown that domain knowledge is needed especially at the point of data loading and the initial investigation of various quality aspects, such as the usability of the data. As data preparation continues, more automation in evaluation is possible.

In the future, the presented framework will be made available to domain experts for an empirical validation. They should review the framework to ensure that all relevant data quality aspects for their domain and data engineering application are included in the classification. Beyond that, we would like to further develop this framework in future work. This includes further levels of the error classification presented as well as the consideration of error hierarchies. Moreover, only the metric for *Missing Value* was described as an example. The elaboration of the other metrics will be done accordingly to emphasize practical relevance. In addition to the further development of the theoretical foundations, implementation also has to be carried out. In our vision, a framework like this must be an integral part of every data engineering pipeline to ensure data quality through data preparation.

References

[Ab16] Abedjan, Z. et al.: Detecting Data Errors: Where are we and what needs to be done? Proc. VLDB Endow. 9/12, pp. 993–1004, 2016, URL: http://www.vldb.org/pvldb/vol9/p993-abedjan.pdf.

[BM11] Blake, R. H.; Mangiameli, P.: The Effects and Interactions of Data Quality and Problem Complexity on Classification. ACM J. Data Inf. Qual. 2/2, 8:1–8:28, 2011, URL: https://doi.org/10.1145/1891879.1891881.

[CBK09] Chandola, V.; Banerjee, A.; Kumar, V.: Anomaly detection: A survey. ACM Comput. Surv. 41/3, 15:1–15:58, 2009, URL: https://doi.org/10.1145/1541880.1541882.

[CZ15] Cai, L.; Zhu, Y.: The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Sci. J. 14/, p. 2, 2015, URL: https://doi.org/10.5334/dsj-2015-002.

[He19] Heidari, A. et al.: HoloDetect: Few-Shot Learning for Error Detection. In: Proceedings SIGMOD 2019. ACM, pp. 829–846, 2019, URL: https://doi.org/10.1145/3299869.3319888.

[HH16] Heinrich, B.; Hristova, D.: A quantitative approach for modelling the influence of currency of information on decision-making under uncertainty. J. Decis. Syst. 25/1, pp. 16–41, 2016, URL: https://doi.org/10.1080/12460125.2015.1080494.

[Ma19] Mahdavi, M. et al.: Raha: A Configuration-Free Error Detection System. In: Proceedings SIGMOD 2019. ACM, pp. 865–882, 2019, URL: https://doi.org/10.1145/3299869.3324956.

[Pi20] Pitoura, E.: Social-minded Measures of Data Quality: Fairness, Diversity, and Lack of Bias. ACM J. Data Inf. Qual. 12/3, 12:1–12:8, 2020, URL: https://dl.acm.org/doi/10.1145/3404193.

[Po17] Polyzotis, N.; Roy, S.; Whang, S. E.; Zinkevich, M.: Data Management Challenges in Production Machine Learning. In: Proceedings SIGMOD 2017. ACM, pp. 1723–1726, 2017, URL: https://doi.org/10.1145/3035918.3054782.

[Re22] Restat, V. et al.: GouDa - Generation of universal Data Sets: Improving Analysis and Evaluation of Data Preparation Pipelines. In: Proceedings DEEM ’22. ACM, 2:1–2:6, 2022, URL: https://doi.org/10.1145/3533028.3533311.

[RKS22] Restat, V.; Klettke, M.; Störl, U.: Towards a Holistic Data Preparation Tool. In: Proceedings DATAPLAT ’22. Vol. 3135, CEUR-WS.org, 2022, URL: https://ceur-ws.org/Vol-3135/dataplat_short1.pdf.

[Sc18] Schelter, S.; Lange, D.; Schmidt, P.; Celikel, M.; Bießmann, F.; Grafberger, A.: Automating Large-Scale Data Quality Verification. Proc. VLDB Endow./, pp. 1781–1794, 2018, URL: http://www.vldb.org/pvldb/vol11/p1781-schelter.pdf.

[Si12] Sidi, F. et al.: Data quality: A survey of data quality dimensions. In: 2012 International Conference on Information Retrieval & Knowledge Management. IEEE, pp. 300–304, 2012, URL: https://doi.org/10.1109/InfRKM.2012.6204995.

[SS20] Schelter, S.; Stoyanovich, J.: Taming technical bias in machine learning pipelines. IEEE Data Engineering Bulletin (Special Issue on Interdisciplinary Perspectives on Fairness and Artificial Intelligence Systems) 43/4, pp. 39–50, 2020.

[Wi16] Wilkinson, M. D. et al.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data/, p. 160018, Mar. 2016, ISSN: 2052-4463, URL: https://doi.org/10.1038/sdata.2016.18.