The Missing Pieces in Classic and Modern Data Error Classification

Dimas Cassimiro Nascimento, Daliton da Silva

[1]Universidade Federal do Agreste de Pernambuco (UFAPE)
Garanhuns - PE – Brazil

{dimas.cassimiro,daliton.silva}@ufape.edu.br

Abstract. *Data quality research has produced extensive catalogs of data error types and data quality problems. However, gaps remain when contrasting classical taxonomies with recent consolidated surveys. This paper identifies three categories of data errors that are insufficiently captured in both foundational and contemporary works: ambiguous abbreviation values, intra-tuple dependency violations based on reference semantics, and cross-source unit inconsistencies. We provide formal definitions for each error type, illustrate them with representative examples, and discuss practical considerations for their detection in real-world scenarios. Our contributions help bridge classical data quality literature with modern data integration challenges, supporting a more comprehensive assessment of data quality in heterogeneous environments.*

1. Introduction

The systematic identification of data errors has long been recognized as a fundamental prerequisite for effective data cleaning and data quality assessment. Over the past two decades, several taxonomies have been proposed to characterize and organize data quality problems. Early contributions presented in [Rahm and Do 2000, Kim et al. 2003, Müller and Freytag 2005] established foundational classifications of dirty data, distinguishing between schema-level and instance-level issues, as well as syntactic, semantic, and coverage-related anomalies. Subsequently, the authors of [Oliveira et al. 2005a, Oliveira et al. 2005b] advanced the field by introducing formally defined data quality problems organized across multiple granularity levels. More recently, the work [Bhadauria et al. 2026] consolidated a broad set of error types and indicators into a unified catalog, reflecting contemporary data quality concerns.

Despite these advances, important gaps remain when contrasting classical taxonomies with recent consolidated surveys. Classical works tend to emphasize issues such as lexical ambiguity, semantic consistency within tuples, and heterogeneities arising from data integration scenarios. In contrast, modern catalogs focus on phenomena such as disguised missing values, bias, and distributional anomalies, which are particularly relevant in large-scale and machine learning-driven applications. As a result, no single framework currently provides comprehensive coverage of both classical and contemporary data error categories. This fragmentation poses practical challenges for practitioners, who often need to combine multiple taxonomies to achieve adequate coverage in real-world data quality assessments.

In this paper, we revisit this gap and identify three categories of data errors that remain insufficiently captured across both foundational and recent works: ambiguous

abbreviation values, intra-tuple dependency violations grounded in reference semantics, and cross-source unit inconsistencies. For each category, we provide formal definitions, illustrative examples, and discuss practical considerations for their detection. Finally, we analyze how these error types can be integrated into existing data profiling and quality assessment workflows, particularly in heterogeneous and data integration settings.

2. Related Work

Data quality research has developed along multiple complementary directions, including error taxonomies, formal definitions, quality dimensions, and detection techniques. Early work by Rahm and Do [Rahm and Do 2000] provided one of the first systematic overviews of data cleaning problems, distinguishing between single-source and multi-source scenarios, as well as schema-level and instance-level errors. Kim et al. [Kim et al. 2003] further refined this perspective by organizing dirty data into hierarchical categories such as missing [Nascimento et al. 2025], wrong, and unusable values. Similarly, the authors of [Müller and Freytag 2005] classified data anomalies into syntactic, semantic, and coverage-related issues, highlighting the diversity of data quality problems encountered in practice.

A significant step toward formalization was made by Oliveira et al. [Oliveira et al. 2005a, Oliveira et al. 2005b], who introduced mathematically grounded definitions for data quality problems organized across different granularity levels. Their work distinguishes itself from prior taxonomies by providing precise criteria for identifying errors such as imprecise values, meaningless values, and structural inconsistencies, enabling more rigorous analysis and potential automation. In parallel, research has also focused on higher-level data quality dimensions and assessment methodologies. Batini et al. [Batini et al. 2009] and the extended treatment in [Batini and Scannapieco 2024] synthesize core dimensions such as completeness, accuracy, consistency, and timeliness, and describe methodologies for measuring and improving data quality. These works emphasize how low-level data errors impact aggregate quality indicators, but they do not aim to provide exhaustive catalogs of fine-grained error types.

More recent efforts have sought to consolidate and systematize existing knowledge. The survey presented in [Ehrlinger and Wöß 2022] reviews tools for data quality measurement and monitoring, highlighting the growing importance of automated detection techniques. Complementing this perspective, Bhadauria et al. [Bhadauria et al. 2026] propose a comprehensive catalog of data errors and indicators, covering a wide range of issues including disguised missing values, out-of-vocabulary words, and bias-related anomalies.

Despite these advances, the literature remains fragmented across different perspectives. Classical taxonomies emphasize semantic ambiguity, intra-tuple consistency, and heterogeneities arising from data integration, while recent catalogs tend to focus on indicators relevant to large-scale and machine learning-driven settings. Furthermore, detection-oriented research, including recent approaches based on large language models [Zhang et al. 2025], has expanded the range of detectable issues without necessarily revisiting the completeness of underlying error taxonomies. As a result, certain categories of data errors remain insufficiently captured across both classical and modern frameworks. In the next section, we identify and formalize three such categories.

3. Problem Definition

We identify three categories of data errors that are only partially captured in both classical taxonomies and recent consolidated catalogs. Each category reflects a distinct aspect of data quality that remains insufficiently formalized in existing error classification frameworks, particularly in data integration and heterogeneous environments.

The first category concerns ambiguous abbreviation values. Classical work by Oliveira et al. [Oliveira et al. 2005a] characterizes imprecise values as those in which abbreviations or acronyms admit multiple valid interpretations. For example, the value *Ant.* in a customer-related attribute may correspond to *Anthony*, *Antonia*, or *Antonio*. This type of ambiguity differs from general semantic ambiguity [Bhadauria et al. 2026] in that it arises from lexical compression rather than from uncertainty in entity reference or context. While interpretability is recognized as a data quality dimension [Batini et al. 2009], current catalogs do not explicitly isolate abbreviation-induced ambiguity as a distinct and formally characterized error type.

The second category involves dependency violations observable within a single tuple when evaluated against reference semantics. Classical functional dependency theory [Date 2005] defines dependencies at the relation level, requiring comparisons across multiple tuples. However, in many practical settings, domain knowledge is externalized in reference data (e.g., code tables or master data), allowing inconsistencies to be identified even in isolated tuples. For example, consider a tuple where *CountryCode* = PT" and *CountryName* = Spain". Given a reference mapping in which PT" corresponds uniquely to Portugal", this tuple violates the intended dependency between code and name. Such inconsistencies are not detectable through standard FD violation checks [Bhadauria et al. 2026], which rely solely on co-occurrence patterns within the observed dataset.

The third category concerns unit inconsistencies across data sources. Classical taxonomies [Rahm and Do 2000, Oliveira et al. 2005b] recognize heterogeneity in measurement units as a common issue in data integration scenarios. However, recent catalogs [Bhadauria et al. 2026] primarily address incorrect or inconsistent units within a single dataset. In practice, different data sources may represent semantically equivalent attributes using distinct units, such as salaries expressed in euros in one source and in dollars in another. When such data are integrated without proper normalization, the resulting dataset becomes semantically inconsistent, even if each source is internally coherent. Although data integration challenges are acknowledged in the literature [Batini and Scannapieco 2024], the explicit characterization of cross-source unit inconsistencies remains limited.

4. Proposed Error Definitions

We now present formal definitions for three data error types. Each definition follows the notation established in [Oliveira et al. 2005a] and extended in [Bhadauria et al. 2026]. Let r denote a relation instance, $t \in r$ a tuple, A the set of attributes, $a \in A$ an attribute, and $v(t, a)$ the value of attribute a in tuple t. For a set of attributes $X \subseteq A$, we denote by $v(t, X)$ the tuple of values corresponding to attributes in X. Let $Dom(a)$ denote the domain of attribute a.

4.1. Ambiguous Abbreviation Value

Let w be a token occurring in a string value $v(t, a)$, where $type(a) = string$. Let $M(a)$ be the set of permissible full-form values for attribute a, and let $expand(w) \subseteq M(a)$ be a function that returns the set of possible full-form expansions of w according to a domain-specific abbreviation dictionary.

Definition 1. Given a tuple $t \in r$, an attribute a with $type(a) = string$, and a token $w \in tokens(v(t, a))$, we say that w is a *potentially ambiguous abbreviation* if:

$$|expand(w)| > 1$$

Let $C(t, w) \subseteq expand(w)$ denote the subset of expansions that are consistent with the remaining attribute values in t. The ambiguity status of w in t is defined as follows: it is considered resolved ambiguity when $|C(t, w)| = 1$; unresolved ambiguity when $|C(t, w)| > 1$; and inconsistency when $|C(t, w)| = 0$.

Example 1. Consider a Customer relation where the attribute ContactMethod contains the value "Call Ant.". The abbreviation "Ant." may expand to *Anthony* or *Antonia*. Since $|expand(\text{"Ant."})| = 2$, the abbreviation is potentially ambiguous. If no additional attribute (e.g., Gender or Title) restricts the set of consistent interpretations, then $C(t, \text{"Ant."}) = \{\textit{Anthony}, \textit{Antonia}\}$, and the ambiguity remains unresolved. Each expansion is individually valid, and the issue arises from the multiplicity of admissible interpretations rather than from semantic invalidity.

4.2. Reference-Based Dependency Violation

Let $X \subseteq A$ and $Y \subseteq A$ be sets of attributes such that a dependency $X \rightarrow Y$ is expected to hold based on domain knowledge. Let r_{ref} be a reference relation encoding valid mappings between values of X and Y.

Definition 2. A tuple $t \in r$ violates a reference-based dependency $X \rightarrow Y$ if:

$$v(t, Y) \notin \pi_Y(\sigma_{X=v(t,X)}(r_{ref}))$$

where $v(t, X)$ denotes the value of attributes X in tuple t, and σ and π denote selection and projection, respectively.

This definition captures inconsistencies detectable through external domain knowledge, even when no conflicting tuples exist in r. Note that r_{ref} may encode one-to-many mappings between X and Y. When $r_{ref} = r$ and $|r| \geq 2$, the definition becomes equivalent to a set-based formulation of functional dependency violation.

Example 2. Consider a relation where *CountryCode* = "PT" and *CountryName* = "Spain". A reference relation specifies that "PT" corresponds to "Portugal". The tuple violates the dependency between code and name, even if no other tuple in the dataset shares the same code value.

4.3. Cross-Source Unit Inconsistency

Let r_1 and r_2 be relations from distinct data sources representing the same entity type, and let a be a semantically equivalent numeric attribute in both relations. Let $Unit_{r_i}(a)$ denote the measurement unit associated with attribute a in relation r_i, when such information is available.

Definition 3. A cross-source unit inconsistency exists for attribute a between r_1 and r_2 if:

$$Unit_{r_1}(a) \neq Unit_{r_2}(a)$$

or, in the absence of explicit unit metadata, if there does not exist a consistent transformation function f such that for corresponding tuples $t_1 \in r_1$ and $t_2 \in r_2$ representing the same entity (according to a given matching process):

$$v(t_2, a) = f(v(t_1, a))$$

In most practical scenarios, f is expected to be a linear or affine transformation. When f is linear, it may be expressed as $f(v) = k \cdot v$. If such a factor k exists but is not explicitly represented in metadata, the inconsistency is considered implicit. If no consistent transformation can be identified based on observed corresponding tuples, the inconsistency is considered irreconcilable.

Example 3. One data source stores employee salaries in USD, while another stores salaries in EUR. If these sources are integrated without applying a conversion function, salary values become incomparable across tuples representing similar entities. This inconsistency arises at the integration level, even when each source is internally consistent.

5. Practical Considerations

The detection of ambiguous abbreviation values relies on the availability of a domain-specific abbreviation dictionary that maps tokens to their possible expansions. Such dictionaries can be constructed from schema metadata, data profiling statistics, or external knowledge sources [Batini and Scannapieco 2024]. A key challenge is that abbreviations are inherently context-dependent and may evolve over time or vary across organizational settings. To address this issue, we recommend a feedback-driven approach in which ambiguous cases are reviewed by data stewards, and validated expansions are incrementally incorporated into the dictionary.

From a computational perspective, detection involves scanning string attributes and applying the $expand(\cdot)$ function to candidate tokens. Assuming an average of w tokens per value and efficient dictionary lookup, the overall complexity is $O(n \cdot w)$ for a relation with n tuples. Additional context-based filtering, when applied, may increase this cost depending on the complexity of cross-attribute consistency checks.

In turn, the detection of reference-based dependency violations depends on the availability and quality of a reference relation encoding valid mappings between determinant and dependent attributes. In practice, such reference data may be incomplete or partially incorrect. When no explicit reference is available, approximate alternatives can be derived by mining association rules or frequent value mappings from curated datasets [Batini et al. 2009]. However, these approaches introduce uncertainty and may lead to false positives or false negatives. A conservative strategy is to report violations only when the determinant value is sufficiently represented in the reference and the observed dependent value is not among the admissible mappings. Confidence thresholds based on frequency or support can further reduce spurious detections.

Lastly, the detection of cross-source unit inconsistencies requires aligning semantically equivalent attributes across data sources. In many real-world scenarios, unit meta-

data is missing or inconsistently documented [Rahm and Do 2000]. In such cases, heuristic approaches can be employed. For example, comparisons of value distributions or ratios between matched entities may reveal systematic scaling differences indicative of unit mismatches. If corresponding records exhibit consistent multiplicative discrepancies, a transformation factor may be inferred. However, distinguishing unit inconsistencies from legitimate value variation remains challenging, particularly in noisy or dynamic datasets. For this reason, automated methods should prioritize the identification and ranking of suspicious cases, leaving final validation to human experts.

The proposed error types complement existing data quality detection techniques and can be incorporated into data profiling and monitoring pipelines. As highlighted by Abedjan et al. [Abedjan et al. 2016], modern data cleaning systems benefit from combining multiple detection strategies, including rule-based, statistical, and knowledge-based approaches. Integrating the proposed checks into such frameworks can improve coverage, particularly in data integration scenarios where heterogeneous sources introduce semantic inconsistencies that are not captured by traditional single-source validation methods.

6. Conclusion and Future Work

This paper revisits the gap between classical data quality taxonomies and recent consolidated error catalogs, identifying three categories of data errors that remain insufficiently formalized in existing frameworks: ambiguous abbreviation values, reference-based dependency violations, and cross-source unit inconsistencies. For each category, we provide formal definitions, illustrative examples, and discuss practical considerations for detection. By explicitly characterizing these error types, our work contributes to a more comprehensive understanding of data quality issues, particularly in heterogeneous and data integration scenarios where semantic inconsistencies are prevalent. Our findings indicate that, despite substantial progress in data quality research, the coverage of error types remains fragmented across taxonomies, quality dimensions, and detection-oriented approaches. Addressing this fragmentation is essential for building more complete and reliable data quality assessment pipelines.

Several directions for future work emerge from this study. First, empirical evaluation on real-world datasets is needed to assess the prevalence of the proposed error types and their impact on downstream analytical and machine learning tasks. Second, integrating these definitions into existing data profiling and monitoring tools, such as those surveyed by [Ehrlinger and Wöß 2022], would support practical validation and adoption. Third, extending the proposed concepts to semi-structured and unstructured data formats, including JSON documents and graph data, would broaden their applicability in modern data ecosystems. Fourth, recent advances in data cleaning using large language models [Zhang et al. 2025] open opportunities to automate tasks such as abbreviation disambiguation and semantic validation, potentially reducing manual effort. Finally, an important open challenge is the development of a unified framework that systematically integrates classical taxonomies with modern error indicators and detection techniques, providing a coherent foundation for more robust and comprehensive data quality management.

Referências

Abedjan, Z., Chu, X., Deng, D., Fernandez, R. C., Ilyas, I. F., Ouzzani, M., Papotti, P., Stonebraker, M., and Tang, N. (2016). Detecting data errors: Where are we and what

needs to be done? *Proceedings of the VLDB Endowment*, 9(12):993–1004.

Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, 41(3):Article 16.

Batini, C. and Scannapieco, M. (2024). *Data Quality: Concepts, Methodologies and Techniques*. Springer, second edition.

Bhadauria, D., Harmouch, H., Naumann, F., Srivastava, D., and Ehrlinger, L. (2026). A catalog of data errors. *arXiv preprint arXiv:2604.09277*.

Date, C. J. (2005). *Database in depth: relational theory for practitioners*. "O'Reilly Media, Inc.".

Ehrlinger, L. and Wöß, W. (2022). A survey of data quality measurement and monitoring tools. *Frontiers in Big Data*, 5:850611.

Kim, W., Choi, B.-J., Hong, E.-K., Kim, S.-K., and Lee, D. (2003). A taxonomy of dirty data. *Data Mining and Knowledge Discovery*, 7(1):81–99.

Müller, H. and Freytag, J. C. (2005). Problems, methods, and challenges in comprehensive data cleansing. Technical Report HUB-IB-164, Humboldt University, Berlin.

Nascimento, D. C., da Silva, D., and Pereira, L. F. A. (2025). Análise teórica do impacto de dados faltantes em atributos sensíveis sobre a métrica de fairness p%-rule. In *Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 767–773. SBC.

Oliveira, P., Rodrigues, F., and Henriques, P. R. (2005a). A formal definition of data quality problems. In *Proceedings of the International Conference on Information Quality (ICIQ)*.

Oliveira, P., Rodrigues, F., Henriques, P. R., and Galhardas, H. (2005b). A taxonomy of data quality problems. In *Proceedings of the 2nd International Workshop on Data and Information Quality*, pages 219–233.

Rahm, E. and Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Bulletin of the Technical Committee on Data Engineering*, 23(4):3–13.

Zhang, S., Huang, Z., and Wu, E. (2025). Data cleaning using large language models. In *Proceedings of the International Conference on Data Engineering (ICDE)*, pages 28–32.