Appendix 1 - Summary of Well-known Data Quality Issues and the Data Quality Dimension Violations Associated

Table 1 - Data Quality Issues and Data Quality Dimension Violations [Adapted from Visengeriyeva and Abedjan, 2020]

|  |  |  |  |
| --- | --- | --- | --- |
| # | Data Quality Issue | Description | Data Quality Dimensions |
| 1 | Missing data | Comprises missing tuples and missing values. Tuple completeness requires that all tuples are present in the table. Missing value issue consists of either null values or disguised values. Value completeness requires that all values are present in the table, while null values indicate that the value is unknown or non-existent. | Accuracy, Completeness |
| 2 | Incorrect data | Data that differ from the values of the real-world entity (e.g., wrong date of birth). | Accuracy |
| 3 | Misspellings | Syntactic deviation of the data value from its ground truth (e.g., “Smiht” instead of “Smith”). | Accuracy |
| 4 | Ambiguous data | Data values which might be interpreted in several ways (e.g., abbreviations or cryptic values). | Accuracy, Consistency |
| 5 | Extraneous data | Presence of additional data in the attribute value (e.g., the address column contains a person’s name in addition to the address). | Consistency, Uniqueness |
| 6 | Outdated temporal data | Values that are obsolete or outdated. | Timeliness |
| 7 | Misfielded values | Values that are placed inside the wrong attribute. | Accuracy, Consistency, Completeness |
| 8 | Incorrect references | Entities that contain wrong information concerning the reference relation (e.g., the employee is associated with a wrong department). | Accuracy |
| 9 | Duplicates | Tuples/values that represent the same real-world entity. | Uniqueness |
| 10 | Structural conflicts | Conflicting duplicates in different sources. | Consistency, Uniqueness |
| 11 | Different word orderings | Values that violate the expected word order (e.g., first name precedes last name). | Consistency, Uniqueness |
| 12 | Different aggregation levels | Entities produced by applying different aggregation methods (e.g., entries per quartal vs. entries per year). | Accuracy, Consistency |
| 13 | Temporal mismatch | Refers to erroneous data that arise due to non-enforcement of integrity constraints for temporal data. | Accuracy, Timeliness |
| 14 | Different units/representations | Occurrence of multiple representations for the same concept (e.g., Price in different currencies). | Consistency |
| 15 | Domain violation | Values that violate semantic rules defined on the specific attribute. | Accuracy |
| 16 | FD violation | Values that violate previously specified functional dependencies. | Accuracy, Consistency |
| 17 | Wrong data type | Values that violate the data type specification of the corresponding attribute, i.e., data type constraint violation. | Consistency |
| 18 | Referential integrity violation | Tuples that violate the referential integrity constraints defined on multiple relations (e.g., missing foreign key). | Accuracy, Consistency, Completeness |
| 19 | Uniqueness violation | Duplication of values under the uniqueness constraint. | Uniqueness |
| 20 | Use of synonyms | Occurrence of synonymous representations for the same concept inside the same column (e.g., “lecturer” and “professor”). | Uniqueness |
| 21 | Use of special characters (space, no space, dash, parentheses) | Refers to different representations of compound data, such as Social Security Number or phone number. | Consistency |
| 22 | Different encoding formats | Inconsistent usage of encodings for values within a dataset (e.g., ASCII or EBCDIC). | Consistency |