CSP working paper series: 02B/01/2023

Identifying Data Quality Dimensions

By

Debarun Sengupta
Anjula Gurtoo
Manjunath B.S.

January 2023

Centre for Society and Policy

Indian Institute of Science

Bengaluru

Abstract:

Data quality is a multi-dimensional concept, indicating qualities that satisfy the needs of a data user. A well-defined data quality offers myriad opportunities and puts the data user in a competitive position as it reduces operational and process costs. However, supporting a uniform quality across the data exchange platform becomes tricky due to heterogeneous datasets. The primary objective of the report is to identify potential dimensions that can be used to assess the overall quality of the dataset. The first part of the report presents a general overview of data quality and exchange platforms, followed by a brief literature review on data quality assessments. In the second part, we have mentioned the basic quality dimensions and also try to indicate some domain-specific dimensions that can be used to improve the overall data quality.

Contents

Abstract: ii

1. Introduction: 4
2. Data quality and exchange platforms: An overview 5
3. Review of literature: 6
4. Data quality indicators: 9
5. Conclusion: 11

References: 13

1. Introduction:

In the twenty-first century, data have become an essential input for innovative products and services, addressing social and environmental challenges, and improving living standards. Data help to optimise the decision-making process, provide operational efficiency and achieve competitive advantages for public and private organisations (Ramos , 2022). For developing countries like India, data-driven strategies can help to deliver better governance in the area of social importance and enhance citizens' lives. However, rapid digitalisation and resulting data streams become the primary challenge for the organisation to conserve data quality – essential input for designing successful data-driven strategies (Ministry of Electronics and Information Technology, 2022).

To promote digital governance and avoid challenges related to data quality, the Government of India (GoI) is planning to introduce a National Governance Framework Policy (NDGFP) (Ministry of Electronics and Information Technology, 2022). The primary objectives of the NDGFP are as follows. First, accelerate digital governance; second, ensure privacy, safety and trust of public digital platforms; third, set quality standards of the non-personal database; and finally, ensure citizens' awareness and participation related to digital governance.

This report comprehensively captures the data quality standards and provides a general overview of data quality dimensions. These findings help to assess the overall quality and identify dimensions that may require further attention. The rest of the report is organised as follows. Section 2 provides a general overview of data quality and the data exchange platform, followed by a brief literature review on data quality in Section 3. Section 4 indicates the potential quality dimensions that can be used to evaluate the data quality. Finally, Section 5 concludes our report by summarising the findings and mentioning future research plans.

2. Data quality and exchange platforms: An overview

According to the Oxford learner's dictionary, data is "facts or information, especially when examined and used to find out things or to make decisions". In the modern digital society, data quality offers strategic advantages for business operations (Bovee et al., 2003). Data quality, in general, can be defined as the level of conformance with the objectivity associated with the data set. If the level of conformance is high, then the dataset quality is considered good; otherwise, it is considered poor. A good quality dataset offers myriad opportunities and puts the data user in a competitive position by reducing operational and process costs[1]. Furthermore, good data quality also helps build trust among users, leading to a higher acceptance rate of business decisions based on the data (Jesiļevska, 2017).

Data exchange platforms commercialise data for digital governance. In order to maximise the data partnership, an exchange platform might select its own data organisational structure. Batini et al. (2009) suggest an organisation structure based on the following data arrangements:

- *Structure data* represent the attributes of a particular domain in a consistent order and are stored in a well-defined fixed scheme, making them easy to analyse. Example: any statistical tables.
- *Semi-structure data* do not follow a well-defined fixed scheme like structure data. However, some organisational schemes are maintained, which makes them easy to process. Example: markup languages like XML and TCP/IP packets.
- *Un-structure data* are the sequence of attributes following random patterns which are difficult to understand. Example: questionnaires with free text, reports, and email bodies.

[1] Operational costs are the sum of lost and missed revenues due to poor data quality; while process costs indicate the cost associated with re-execution of the process due to errors in data (Batini et al. 2009).

Shankaranarayanan et al. (2000) propose an alternative organisational structure based on the following manufacturing process:

- *Raw data* do not undergo any process during or after procurement and generally store for long periods.
- *Semi-process (or component) data* are temporarily generated by processing the raw data and used to generate the final dataset.
- *Information data* are the final dataset generated after processing the raw or semi-process data.

Based on the above discussions, we can conclude that data quality and structure are strongly associated. Here, both structure and information represent the highest data quality level, while un-structure or raw data represents the lowest data quality levels.

3. Review of literature:

Literature on data quality reveals that researchers follow a different methodology to evaluate data quality. Here we chronologically discuss studies to provide the reader with a glimpse of empirical techniques to evaluate the data quality dimensions. We have restricted our review to only the studies published in the last three decades. Wang and Strong (1996) propose a user-centric framework for data quality. The study follows a two-stage survey method to understand individual preferences for quality dimensions. In stage one, they have listed the possible data quality attributes, and in the follow-up (i.e., second stage), they rank those dimensions based on individual preferences. During stage one, the study considers 112 MBA students and 25 individuals working in the industry. For MBA students, they use a self-administrated survey method, while for working individuals, the survey is followed after a meeting on data quality. From the first round survey, they have identified around 118 quality dimensions which can be further merged into 20 quality dimensions. In the follow-up survey round, authors have mailed

questionnaires to 1,500 individuals to identify the essential dimensions from these 20 dimensions. The findings indicate that accuracy and correctness are two important data quality dimensions. The study also groups the quality dimensions into the following four categories – intrinsic, contextual, representational, and accessibility to develop a hierarchical framework. An advantage of such a framework is that it helps professionals meet consumer requirements.

Lee et al. (2002) have developed a methodology for benchmarking the data quality. Like Wang and Strong (1996), they organise the key dimensions into four groups (i.e., sound information, dependable information, usable information, and useful information). The study considers 261 respondents from five organisations to collect data quality information. The findings suggest internal consistency of quality dimensions and support the multi-dimensional claim. Furthermore, the gap analysis technique indicates the deficiencies which can be used to determine the organisation-specific benchmarking. Pipino et al. (2002) have argued that data quality is a multi-dimensional concept which can be classified into subjective (i.e., the experience of the data user) and objective (i.e., general and contextual dimensions) dimensions. The overall quality can be determined through a combined approach of data quality assessment. However, the study mainly focuses on objective evaluation. It suggests the following three techniques – simple ratio (i.e., a ratio of the desired outcome to total outcome), min or max operation (i.e., computing the minimum or maximum value based on the normalised value of the individual quality attributes), and weighted average (i.e., an alternative to min or max operation). These functional forms may help to identify the existing discrepancies for each attribute, followed by the define the quality metric for data quality attributes.

Bovee et al. (2003) present a user-centric model to determine data quality and solve the pre-existing problem of interdependence and confusion within quality dimensions. Their proposed model follows a simple and intuitive framework incorporating four attributes (accessibility, interpretability, relevance, and integrity). To evaluate the nature of these data qualities, the

study employs the belief-function framework under heterogeneous scenarios (i.e., online users and assurance providers) that capture the impact of data quality. Moreover, the study also uses two different algorithms (logical and weighted average) to check the sensitivity of each scenario. The findings suggest that an individual's experience plays a vital role in determining his or her belief about data quality. De Amicis et al. (2006) use dependencies among data attribute to explain the overall data quality. They employ the Shannon entropy to develop an analytical framework to define the possible correlation between attributes, which can be further classified into three groups – independency, partial dependency, and perfect dependency. The proposed analytical framework has been implemented to find an association between two financial variables to evaluate the quality of a financial database. However, to detect the error in the data, the study compares the result with independent external data, which is assumed to be correct. The proposed data-driven technique analyses 4175 observations and detects a partial dependency between data quality attributes. Furthermore, the study identifies the lack of timeliness between two variables as the major limitation of data quality.

Vaziri et al. (2016) have developed a process-driven method or task-based data quality (TBDQ) technique to evaluate the quality of structure data. The TBDQ technique follows assessments and improvement phases to reach a desired level of quality. In order to identify the potential risks, the assessment stage can be further divided into two sub-stages – planning (i.e., defining the scope and objectivity of each quality dimension) and evaluation (i.e., computing the dimension-specific weights employing a hierarchical process). Similarly, to recommend future betterments, the improvement stage uses evolution (i.e., prioritise qualities to suggest further improvements) and execution (i.e., the process of improving the data quality) sub-stages. The authors employ the methodology to evaluate the data quality for an international seed trading company in Iran. The findings suggest that the methodology can effectively address the optimal data quality activity.

Jesilevska (2017) presents expert reviews on data quality issues to ensure optimal data quality. Her assessment framework contains 13 quality dimensions, tested under four different usage contexts (i.e., scientific research, decision-making, analysis progress of research, and object modelling and forecasting). The study follows a paper-based questionnaire and targets 11 (five academics and six national governments) experts to define the optimal dimensions. During the survey, respondents are asked to rank the proposed data quality dimensions from most essential to least essential based on data usage. The study findings indicate that the ranking of quality dimensions might vary across data usage, but data completeness remains essential for every context. The study also compares the data quality requirement between a statistician and a data user to combine the quality dimensions. The high and significant non-parametric rank correlation test statistic implies that both experts follow a similar identification strategy to improve overall data quality. González-Vidal et al. (2022) evaluate the data quality of an open data repository. The study considers five basic quality metrics (completeness, timeliness, plausibility, artificiality, and concordance) and employs principal component analysis (PCA) to visually represent data quality dimensions. For the case study, the authors select the Cambridge Open Data Repository. The findings suggest that the open database may follow the proposed methodology to improve quality and attract new users with minimum computational cost and human intervention.

4. Data quality indicators:

Data quality or fitness for use is a multi-dimensional concept, indicating different qualities that satisfy the data user's needs. Note that data quality is subjective and can differ across data users. Table 1 shows the list of possible data quality attributes. Each column stands for quality assessment methodology, while each row indicates the quality dimensions. Among these 54 dimensions, we have identified the top six dimensions covered by most assessment methodologies. These six dimensions can also be considered the six fundamental quality

dimensions for any database. However, interested readers may refer to Batini et al. 2009, who present detailed classifications of quality attributes.

<Table 1: Data quality attributes>

1. **Accessibility:** Data accessibility is the metric of difficulty in accessing the existing data. From a user perspective, accessibility is the primary dimension allowing individuals to use data for their respective analyses. A data exchange platform can follow two types of accessibility approaches. First is random access, which gives direct access to requested data, and second is sequential access, where requested data are obtained through a particular sequence. Generally, data users prefer random access as it requires fewer seek operations to obtain a particular data element.
2. **Accuracy:** The accuracy dimensions measure the degree to which the data values are correct, reliable and certified to present real-world situations. Data accuracy is vital in decision-making, as inaccurate data can increase the operational cost of business organisations.
3. **Consistency:** After accuracy, data consistency is another crucial dimension that defines data quality. A data value may be accurate but remain inconsistent due to its different representational forms. By data consistency, we mean that each value remains identical for all instances of data use. Note that if data values are inconsistent, at least one should be incorrect.
4. **Completeness:** The completeness dimension is the fundamental dimension of data quality, indicating whether any required elements are missing in the database. Here completeness may indicate the missing element in a particular schema or a particular column. One can also identify the missing observations within a dataset relative to the population.

5. **Interpretability:** The interpretability dimension defines the context in which data elements can be easily understood and analysed. It includes using standard concepts, terminology and availability of supplementary information, which make data more accessible.
6. **Timeliness:** Timeliness is a time-related data dimension of the data quality, indicating the time lag between the occurrence of an event and the availability of the corresponding data to support the decision-making based on data.

Note that one can make the following four clusters based on the underlying aspects of data quality (Wang and Strong, 1996; Lee et al., 2002). An advantage of such an exercise is that it develops a uniform framework for data quality dimensions that public and private organisations can follow. Below we discuss each cluster to deliver a glimpse of data quality dimensions.

- *Intrinsic quality* implies that data has its quality which includes data accuracy, objectivity, reputation, and believability. Among these qualities, data accuracy is considered the primary intrinsic quality dimension.
- *Contextual quality* measures the context or the underline dimensions of data required for a specific task. It includes dimensions like data completeness and timeliness.
- *Representational quality* indicates the data-quality dimensions associated with data consistency and data interpretations.
- *Accessibility quality* differs from other data qualities as it primarily focuses only on the accessibility dimension of the data.

5. Conclusion:

This present report tries to provide a general overview of data quality and identifies the potential data quality dimensions to define the quality of a database. Moreover, the study also indicates a hierarchical framework to recognise the aspects which can be used to improve data

quality further. However, these quality dimensions or aspects are subjective and may require expert opinion for more detailed and robust identification.

Note that these potential dimensions are not sufficient to define the overall data quality. For example, relevance and reputation may be necessary for datasets related to citizen grievances (i.e., waste management and civic animates). Similarly, we must consider precision, reputation, and believability for traffic and meteorological databases. Furthermore, these quality attributes are not equality importance for each database. For example, accuracy and timeliness are essential for traffic data relative to other dimensions. Similarly, completeness and interpretability may play a vital role in determining the quality of environmental or meteorological datasets. Therefore, we propose another study as future research to validate the data quality dimensions across domains.

References:

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for Data Quality Assessment and Improvement. *ACM Computing Surveys, 41*(3).

Bovee, M., Srivastava, R. P., & Mak, B. (2003). A Conceptual Framework and Belief-Function Approach to Assessing Overall Information Quality. *International Journal of Intelligent Systems, 18*, 51–74.

De Amicis, F., Barone, D., & Batini, C. (2006). An analytical framework to analyse dependencies among data quality dimensions. *11th International Conference on Information Quality (ICIQ)*, (pp. 369-383).

González-Vidal, A., Ramallo-González, A., & Skarmeta, A. (2022). Intrinsic and extrinsic quality of data for open data repositories. *ICT Express, 8*, 328–333.

Jesiļevska, S. (2017). Data Quality Dimensions to Ensure Optimal Data Quality. *The Romanian Economic Journal, 20*(63).

Lee, Y., Strong, D., Kahn, B., & Wang, R. (2002). AIMQ: a methodology for information quality assessment. *Information & Management, 40*, 133-146.

Ministry of Electronics and Information Technology. (2022). *National Data Governance Framework Policy (Draft)*. Delhi: Ministry of Electronics and Information Technology, Government of India. Retrieved from https://meity.gov.in/writereaddata/files/National-Data-Governance-Framework-Policy.pdf

Pipino, L., Lee, Y., & Wang, R. (2002). Data Quality Assessment. *Communications of the ACM, 45*(4).

Ramos , J. (2022, June 8). *Overcome these six data consumption challenges for a more data-driven enterprise.* Retrieved from Journey to AI Blog: https://www.ibm.com/blogs/journey-to-ai/2022/06/overcome-these-six-data-consumption-challenges-for-a-more-data-driven-enterprise/

Shankaranarayanan, G., Wang, R., & Ziad, M. (2000). IP-MAP: Representing the Manufacture of an Information Product. *Proceedings of the 2000 Conference on Information Quality*, (pp. 1-16).

Vaziri, R., Mohsenzadeh, M., & Habibi, J. (2016). TBDQ: A Pragmatic Task-Based Method to Data Quality Assessment and Improvement. *PLoS One, 11*(5).

Wang, R., & Strong, D. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems, 12*(4), 5-33.

Table 1: Data quality dimensions

Sl. No	Dimensions	TDQM	DWQ	TIQM	AIMQ	CIHI	DQA	IQM	ISTAT	AMEQ	COLDQ	DaQuinCIS	QAFD	CDQ	HDQM	TQPA
1	Accessibility	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES			YES		
2	Accuracy	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
3	Adaptability					YES										
4	Applicability							YES								
5	Appropriateness	YES			YES		YES			YES	YES					
6	Believability	YES			YES		YES			YES						
7	Business conformance			YES												
8	Clarity							YES			YES					
9	Comparability					YES										
10	Completeness	YES	YES	YES	YES		YES		YES	YES	YES	YES	YES	YES		YES
11	Comprehensiveness							YES			YES					
12	Consistent representation	YES			YES		YES	YES								
13	Concurrency of redundant data			YES												
14	Consistency		YES	YES			YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
15	Consistent representation									YES						
16	Contextual clarity			YES												
17	Convenience							YES								
18	Correctness		YES					YES		YES				YES		
19	Cost			YES						YES	YES			YES		
20	Credibility		YES													
21	Currency										YES	YES	YES	YES	YES	YES
22	Derivation integrity			YES												
23	Documentation					YES										
24	Ease of manipulation	YES			YES		YES									
25	Equivalence of redundant data			YES												
26	Flexibility										YES					

Note: (i) Column stands for assessment methodology; (ii) row indicates the quality dimensions
Source: Author's creation

Table 1: Data quality dimensions (Contd.)

Sl. No	Dimensions	TDQM	DWQ	TIQM	AIMQ	CIHI	DQA	IQM	ISTAT	AMEQ	COLDQ	DaQuinCIS	QAFD	CDQ	HDQM	TQPA
27	Interactivity							YES								
28	Interpret ability	YES	YES		YES	YES	YES			YES	YES					
29	Linkage ability					YES										
30	Maintainability							YES								
31	Meaningful Ness									YES						
32	Metadata		YES								YES					
33	Minimality		YES													
34	Non-duplication			YES												
35	Objectivity	YES			YES		YES			YES						
36	Precision			YES						YES	YES					
37	Presentation										YES					
38	Portability										YES					
39	Readable									YES						
40	Relevance	YES			YES	YES	YES			YES	YES					
41	Reliability									YES						
42	Reputation				YES	YES	YES			YES	YES	YES	YES	YES		
43	Responsiveness		YES													
44	Security	YES			YES	YES	YES	YES			YES					
45	Speed							YES								
46	Standardisation					YES										
47	Timeliness	YES	YES	YES	YES	YES	YES	YES			YES			YES		YES
48	Traceability		YES					YES		YES						
49	Unambiguity									YES						
50	Understandability	YES			YES		YES			YES						
51	Uniqueness												YES			YES
52	Usefulness		YES	YES		YES					YES					
53	Value added	YES			YES	YES	YES			YES						
54	Volatility													YES		

Note: (i) Column stands for assessment methodology; (ii) row indicates the quality dimensions
Source: Author's creation

Table 1A: List of quality assessment methodologies

Abbreviation	Full forms
TDQM	Total Data Quality Data Management
DWQ	Data Warehouse Quality
TIQM	Total Information Quality Management
AIMQ	A Methodology for Information Quality Assessment
CIHI	Canadian Institute for Health Information
DQA	Data quality Assessment
IQM	Information Quality Management
ISTAT	Italian National Bureau of Census
AMEQ	Activity-based Measuring and Evaluating of product information Quality
COLDQ	Cost-effect Of Low Data Quality
DaQuinCIS	DaQuinCIS Data Quality in Cooperative Information Systems
QAFD	Quality Assessment on Financial Data
CDQ	A comprehensive methodology for Data Quality
HDQM	Heterogeneous Data Quality Management