Data Quality—Concepts and Problems

Subjects: Architecture And Design
Contributor: Max Hassenstein

Data Quality is, in essence, understood as the degree to which the data of interest satisfies the requirements, is free of flaws, and is suited for the intended purpose. Data Quality is usually measured utilizing several criteria, which may differ in terms of assigned importance, depending on, e.g., the data at hand, stakeholders, or the intended use.

Keywords: data quality ; information quality ; data quality dimensions ; data life cycle

---

The word data is the plural form of the Latin noun datum (verbatim “something given”) [1]. In general, data is “information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer” [2]. The sense of data may vary depending on the context; for example, researchers often work with data sets, which are data in an accumulated and structured, often in tabularized form.

Old papyrus fragments from ancient Egypt (specifically, from the 26th century BC) indicate the early existence of logbooks and documentation, thus proving data collection as a phenomenon as old as early civilizations [3]. The Roman Empire, for instance, also recorded and collected data as evidenced by its historical censuses to create population registers containing asset estimations and medical examinations for military service [4].

Today, and due to the digital age, data have become omnipresent in private, commercial, political and scientific environments. Computing underwent drastic transformation within the past 40 years: until the 1980s, centralized data centers gathered data and were business-orientated, and by 2000, data centers expanded their data management capabilities, and individual users increasingly had access to a private computer and the World Wide Web (WWW) [5]. Since 2000 and with the increasing spread of the internet, data centers have expanded their capacities to cloud computing, resulting in considerably increased amounts of data collected and available [5].

Shannon [6], a pioneer of information theory, defined information as a simple unit of message (e.g., a binary digit, known as bit), either stand-alone or as a sequence, sent by a sender to a receiver. However, we see a certain degree of distinction between the terms data and information; from our point of view, data are relatively solitary and of a technical nature, and require interpretation or placement to become information [7][8].

The word quality has multiple origins, among others, from the Latin noun qualitas (verbatim characteristic, nature). According to ISO 9000:2015 [9], quality is the “degree to which a set of inherent characteristics of an object fulfills requirements.” Nevertheless, the requirements remain undefined at this point. Therefore, in our context, quality broadly refers to the extent of the goodness of a thing (for instance, our data).

Based on the presented terms for quality and data, a definition for data quality can already be deduced: the degree to which the data of interest fulfills given requirements, as similarly defined by Olson [10]. However, the literature offers additional interpretations of the data quality concept. These are, in essence: Whether the data are fit for (the intended) use and free of flaws [11] or meet the needs and requirements of their users [12][13]. In this regard, data quality requirements may be imposed by standards, legislation, regulations, policies, stakeholders, or their intended use [14].

For instance, the wide availability of modern information technology, such as smart devices (phones, tablets, and wearables), has made people eager to track their physical activity, sleep and other health data, or dietary habits as a hobby [15][16]. Likewise, companies have turned data into a business model (for instance, Google or Meta, previously known as Facebook) or accumulate data for knowledge management. Furthermore, specific scientific disciplines, such as epidemiology, acquire data to research health conditions and their causes [17]. These are just a few examples of how much data has become part of everyday life. However, the ubiquity of data goes hand in hand with the ubiquity of data quality issues. Simple examples from everyday life are outdated phone numbers or unregistered changes of residence in a contact directory, which may lead to an inability to contact a particular person or bias statistical analyses that consider geographical variables, challenging the usefulness of the directory.

The quality and applicability of data should not and cannot be assumed by default, as they may directly impact data processing as well as the results and conclusions derived from the data [18][19].

High-quality research and analyses require reliable data, frequently referenced inversely as “garbage in, garbage out” [20][21]. Even if, from our point of view, quality considerations concerning the data collected might be as old as the collection procedure itself, we only find the rather modern literature to discuss this matter [22][23]. Nevertheless, data quality was already labeled “a key issue of our time” [24], at a much lower digitization level in 1986.

The primary motivation for work in the field of data quality is generally to ensure data integrity and, thus, in principle, to ensure the usability and usefulness of the data. Thereby, the stakeholders of data quality are data producers, data users, analysts, and people who derive conclusions from interpreted data, such as the readership of a study or the recipients of information provided via the WWW. Regardless, data quality considerations should primarily concern the people either involved in data collection, data generation, or those analyzing or providing data, as well as people with direct data access, as they have the means to address data quality issues. Ideally, data quality considerations precede and accompany the data collection phase and may imply, for example, measures to assure the data structure or value range controls. However, as discussed in Section 2.2.3, quality assurance may be a continuous process.

Our contribution is structured as follows. The following section presents data quality concepts and discusses data quality assessment within these frameworks. Section 3 illustrates data quality issues in real-life examples, focusing on the health sciences, to give the readers a better grasp of the theoretical concepts presented earlier. Finally, Section 3.3 describes the challenges associated with the practical application of the data quality frameworks before we close the paper with a conclusion in Section 4.

References

1. Lexico English Dictionary (Online). Data. Oxford University Press, 2021. Available online: https://www.lexico.com/definition/data (accessed on 20 January 2022).
2. Cambridge Dictionary. Data. Available online: https://dictionary.cambridge.org/dictionary/english/data (accessed on 20 January 2022).
3. Tallet, P. Les Papyrus de la Mer Rouge I: Le Journal de Merer (PAPYRUS JARF A ET B); Institut Français D’archéologie Orientale: Kairo, Egypt, 2017.
4. Unruh, F. “... Dass alle Welt geschätzt würde”: Volkszählung im Römischen Reich; Gesellschaft für Vor- und Frühgeschichte in Württemberg und Hohenzollern e.V.; Thiess: Stuttgart, Germany, 2001; Volume 54.
5. Reinsel, D.; Gantz, J.; Rydning, J. Data Age 2025: The Evolution of Data to Life-Critical. An IDC White Paper; International Data Corporation (IDC): Framingham, MA, USA, 2017.
6. Shannon, C.E. A Mathematical Theory of Communication. Repr. Correct. Bell Syst. Tech. J. 1948, 27, 379–423.
7. Logan, R.K. What Is Information?: Why Is It Relativistic and What Is Its Relationship to Materiality, Meaning and Organization. Information 2012, 3, 68–91.
8. Hewitt, S.M. Data, Information, and Knowledge. J. Histochem. Cytochem. 2019, 67, 227–228.
9. International Organization for Standardization. ISO 9000:2015, Quality Management Systems—Fundamentals and Vocabulary, 5th ed.; International Organization for Standardization: Geneva, Switzerland, 2015.
10. Olson, J.E. Data Quality: The Accuracy Dimension; Morgan Kaufmann: San Francisco, CA, USA, 2003.
11. Redman, T.C. Data Quality: The Field Guide; Digital Press: Boston, MA, USA, 2001.
12. Wang, R.Y.; Strong, D.M. Beyond Accuracy: What Data Quality Means to Data Consumers. J. Manag. Inf. Syst. 1996, 12, 5–33.
13. Kahn, B.K.; Strong, D.M.; Wang, R.Y. Information quality benchmarks: Product and service performance. Commun. ACM 2002, 45, 184–192.
14. Fürber, C. Data Quality Management with Semantic Technologies, 1st ed.; Springer Gabler: Wiesbaden, Germany, 2015.
15. Piwek, L.; Ellis, D.A.; Andrews, S.; Joinson, A. The Rise of Consumer Health Wearables: Promises and Barriers. PLoS Med. 2016, 13, e1001953.

16. Jones, S. Health & Fitness Wearables: Market Size, Trends & Vendor Strategies 2020–2025; Juniper Research Ltd.: Basingstoke, Hampshire, UK, 2020.
17. Rothman, K.J. Epidemiology: An Introduction, 2nd ed.; Oxford University Press: New York, NY, USA, 2012.
18. Loh, W.-Y.; Zhang, Q.; Zhang, W.; Zhou, P. Missing data, imputation and regression trees. Stat. Sin. 2020, 30, 1697–1722.
19. McCausland, T. The Bad Data Problem. Res.-Technol. Manag. 2021, 64, 68–71.
20. Arias, V.B.; Garrido, L.E.; Jenaro, C.; Martinez-Molina, A.; Arias, B. A little garbage in, lots of garbage out: Assessing the impact of careless responding in personality survey data. Behav. Res. Methods 2020, 52, 2489–2505.
21. Kilkenny, M.F.; Robinson, K.M. Data quality: “Garbage in-garbage out”. Health Inf. Manag. J. 2018, 47, 103–105.
22. Naroll, F.; Naroll, R.; Howard, F.H. Position of women in childbirth. A study in data quality control. Am. J. Obstet. Gynecol. 1961, 82, 943–954.
23. Vidich, A.J.; Shapiro, G. A Comparison of Participant Observation and Survey Data. Am. Sociol. Rev. 1955, 20, 28–33.
24. Jensen, D.L.; Wilson, T.F.; United States Bureau of Justice Statistics; Search Group. Data Quality Policies and Procedures: Proceedings of a BJS/SEARCH Conference: Papers; U.S. Department. of Justice, Bureau of Justice Statistics: Washington, DC, USA, 1986.