

Data Quality: Defining, Measuring and Improving

Hans – J. Lenz,

Freie Universität Berlin, Germany
hjlentz@wiwiss.fu-berlin.de

In industry the term „Quality“ used in the context of “Quality Control or Assurance” of products - and later services - has a history of about one hundred years. It is used in an ISO norm as “Suitability for use relative to a given objective of usage”. Looking at “Products” and “Processes” one distinguishes between “Quality of Design” and “Quality of Performance”.

“Data Quality” is a term which is used at Statistical Offices and supranational Organizations (OECD, UN NA-Group etc.) for about the same time. It became popular in computer science twenty years ago, when data quality problems related to data warehousing, ETL, data cleansing, data mining and data integration were detected.

Data Quality is mostly defined as above, i.e. fitness for use given an objective of data processing on a specific domain. For example, the objective may be web-mining where semi-structured data is to be integrated. Evidently, the term “data quality” has many various facets. Stepwise refining the granularity starting from several data sources to a single value of an attribute (variable) one can differ between multi-sources or data bases, single databases (on the schema or data level), records and values. For instance, on the data level errors, outliers, null-values (missing values), inconsistent (incoherent) values or simply semantic misuse of data are of concern while on the schema level integrity constraints may be violated. All these factors may lead to low data quality.

Wang and Strong (1996) widened the view on data quality when they unfolded the term and advocated four dimensions like intrinsic data quality (accuracy, reputation etc.), contextual data quality (relevancy, completeness etc.), representational data quality (consistency, conciseness etc.) and access quality (accessibility, security etc.).

An open question is how to define and measure the involved indicators (“metrics”) of data quality as well as how to use them for the design, operating, monitoring and controlling of the data entering, transformation, production, analysis and dissemination process.

Therefore the tutorial focuses on the following views on data quality:

1. Definitions and measurements of data quality indicators.
2. The role of an extended repository (meta data information system) is discussed.
3. The “Big Three Trouble Makers”: Missing Data (Null Values), Outliers and Errors (of measurement, sampling, simulation or prediction).
4. A Benchmark Database and the methodology useful for resolving some typical data conflicts in ETL.
5. A probabilistic and possibilistic approach to generalize the concept of edits applied to dirty or missing metric data (“statistical or fuzzy edits”) when balance equations (i.e. a model) are at hand.
6. We present a framework and a sound methodology of object identification (“approximate joins”)
7. Statistical and Semantic Consistency when mixing transformations, aggregations and data-cube operators
8. We give hints to real world applications (Census (ARC), eNAQ, EcoOil etc.).

Lit.:

C. Batini and M. Scannapieco (2006)
M.A. Hernandez and S. J. Stolfo (1998)
I.P. Fellegi and Holt, D. (1976)
I.P. Fellegi and A.B. Sunter (1969)
H.-J. Lenz and B. Thalheim (2005)
F. Naumann (2002)
M. Neiling et al. (2003)
R.Y. Wang and D.M. Strong (1996)
W.E. Winkler (1995)