Data Quality Assessment

LEO L. PIPINO, YANG W. LEE, AND RICHARD Y. WANG

How good is a company's data quality? Answering this question requires usable data quality metrics. Currently, most data quality measures are developed on an ad hoc basis to solve specific problems [6, 8], and fundamental principles necessary for developing usable metrics in practice are lacking. In this article, we describe principles that can help organizations develop usable data quality metrics.

Studies have confirmed data quality is a multi-dimensional concept [1, 2, 6, 9, 10, 12]. Companies must deal with both the subjective perceptions of the individuals involved with the data, and the objective measurements based on the data set in question. Subjective data quality assessments reflect the needs and experiences of stakeholders: the collectors, custodians, and consumers of data products [2, 11]. If stakeholders assess the quality of data as poor, their behavior will be influenced by this assessment. One can use a questionnaire to measure stakeholder perceptions of data quality dimensions. Many healthcare, finance, and consumer product companies have used one such questionnaire, developed to assess data quality dimensions listed in Table 1 [7]. A major U.S. bank that administered the questionnaire found custodians (mostly MIS professionals) view their data as highly timely, but consumers disagree; and data consumers view data as difficult to manipulate for their business purposes, but custodians disagree [4, 6]. A follow-up investigation into the root causes of differing assessments provided valuable insight on areas needing improvement.

Objective assessments can be task-independent or task-dependent. Task-independent metrics reflect states of the data without the contextual knowledge of the application, and can be applied to any data set, regardless of the tasks at hand. Task-dependent metrics, which include the organization's business rules, company and government regulations, and constraints provided by the database administrator, are developed in specific application contexts.

LEO L. PIPINO (Leo_Pipino@uml.edu) is professor of MIS in the College of Management at the University of Massachusetts Lowell.
YANG W. LEE (y.wlee@neu.edu) is an assistant professor in the College of Business Administration at Northeastern University in Boston, MA.
RICHARD Y. YANG (rwang@bu.edu) is an associate professor at Boston University and Co-director of the Total Data Quality Management (TDQM) program at MIT Sloan School of Management in Cambridge, MA.

Table 1. Data quality dimensions.

Dimensions	Definitions
Accessibility	the extent to which data is available, or easily and quickly retrievable
Appropriate Amount of Data	the extent to which the volume of data is appropriate for the task at hand
Believability	the extent to which data is regarded as true and credible
Completeness	the extent to which data is not missing and is of sufficient breadth and depth for the task at hand
Concise Representation	the extent to which data is compactly represented
Consistent Representation	the extent to which data is presented in the same format
Ease of Manipulation	the extent to which data is easy to manipulate and apply to different tasks
Free-of-Error	the extent to which data is correct and reliable
Interpretability	the extent to which data is in appropriate languages, symbols, and units, and the definitions are clear
Objectivity	the extent to which data is unbiased, unprejudiced, and impartial
Relevancy	the extent to which data is applicable and helpful for the task at hand
Reputation	the extent to which data is highly regarded in terms of its source or content
Security	the extent to which access to data is restricted appropriately to maintain its security
Timeliness	the extent to which the data is sufficiently up-to-date for the task at hand
Understandability	the extent to which data is easily comprehended
Value-Added	the extent to which data is beneficial and provides advantages from its use

In this article, we describe the subjective and objective assessments of data quality, and present three functional forms for developing objective data quality metrics. We present an approach that combines the subjective and objective assessments of data quality, and illustrate how it has been used in practice. Data and information are often used synonymously. In practice, managers differentiate information from data intuitively, and describe information as data that has been processed. Unless specified otherwise, this paper will use *data* interchangeably with *information.*

Functional Forms

When performing objective assessments, companies should follow a set of principles to develop metrics specific to their needs. Three pervasive functional forms are simple ratio, min or max operation, and weighted average. Refinements of these functional forms, such as addition of sensitivity parameters, can be easily incorporated. Often, the most difficult task is precisely defining a dimension, or the aspect of a dimension that relates to the company's specific application. Formulating the metric is straightforward once this task is complete.

Simple Ratio. The simple ratio measures the ratio of desired outcomes to total outcomes. Since most people measure exceptions, however, a preferred form is the number of undesirable outcomes divided by total outcomes subtracted from 1. This simple ratio adheres to the convention that 1 represents the most desirable and 0 the least desirable score [1, 2, 6, 9]. Although a ratio illustrating undesirable outcomes gives the same information as one illustrating desirable outcomes, our experience suggests managers prefer the ratio showing positive outcomes, since this form is useful for longitudinal comparisons illustrating trends of continuous improvement. Many traditional data quality metrics, such as *free-of-error*, *completeness*, and *consistency* take this form. Other dimensions that can be evaluated using this form include *concise representation*, *relevancy*, and *ease of manipulation*.

The *free-of-error* dimension represents data correctness. If one is counting the data units in error, the metric is defined as the number of data units in error divided by the total number of data units subtracted from 1. In practice, determining what constitutes a data unit and what is an error requires a set of clearly defined criteria. For example, the degree of precision must be specified. It is possible for an incorrect character in a text string to be tolerable in one circumstance but not in another.

The *completeness* dimension can be viewed from many perspectives, leading to different metrics. At the most abstract level, one can define the concept of schema completeness, which is the degree to which entities and attributes are not missing from the schema. At the data level, one can define column completeness as a function of the missing values in a column of a table. This measurement corresponds to Codd's column integrity [3], which assesses missing values. A third type is called population completeness. If a column should contain at least one occurrence of all 50 states, for example, but it only contains 43 states, then we have population incompleteness. Each of the three types (schema completeness, column completeness, and population completeness) can be measured by taking the ratio of the number of incomplete items to the total number of items and subtracting from 1.

The *consistency* dimension can also be viewed from a number of perspectives, one being consistency of the same (redundant) data values across tables. Codd's Referential Integrity constraint is an instantiation of this type of consistency. As with the previously discussed dimensions, a metric measuring consistency is the ratio of violations of a specific consistency type to the total number of consistency checks subtracted from one.

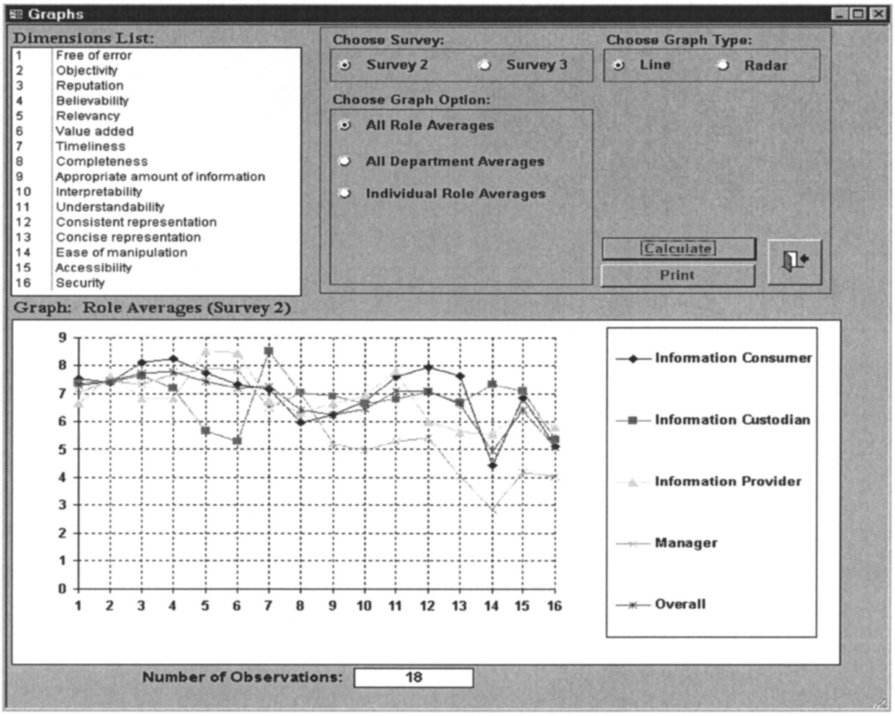
Min or Max Operation. To handle dimensions that require the aggregation of multiple data quality indicators (variables), the minimum or maximum operation can be applied. One computes the minimum (or maximum) value from among the normalized values of the individual data quality indicators. The min operator is conservative in that it assigns to the dimension an aggregate value no higher than the value of its weakest data quality indicator (evaluated and normalized to between 0 and 1).

The maximum operation is used if a liberal interpretation is warranted. The individual variables may be measured using a simple ratio. Two interesting examples of dimensions that can make use of the min operator are *believability* and *appropriate amount of data*. The max operator proves useful in more complex metrics applicable to the dimensions of *timeliness* and *accessibility*.

Believability is the extent to which data is regarded as true and credible. Among other factors, it may reflect an individual's assessment of the credibility of the data source, comparison to a commonly accepted standard, and previous experience. Each of these variables is rated on a scale from 0 to 1, and overall believability is then assigned as the minimum value of the three. Assume the believability of the data source is rated as 0.6; believability against a common standard is 0.8; and believability based on experience is 0.7. The overall believability rating is then 0.6 (the lowest number). As indicated earlier, this is a conservative assessment. An alternative is to compute the believability as a weighted average of the individual components.

A working definition of the *appropriate amount of data* should reflect the data quantity being neither too little nor too much. A general metric that embeds this tradeoff is the minimum of two simple ratios: the ratio of the number of data units provided to the number of data units needed, and the ratio of the number of data units needed to the number of data units provided.

Timeliness reflects how up-to-date the data is with respect to the task it's used for. A general metric to measure timeliness has been proposed by Ballou et al., who sug-



Dimensions List:

1	Free of error
2	Objectivity
3	Reputation
4	Believability
5	Relevancy
6	Value added
7	Timeliness
8	Completeness
9	Appropriate amount of information
10	Interpretability
11	Understandability
12	Consistent representation
13	Concise representation
14	Ease of manipulation
15	Accessibility
16	Security

Choose Survey: ● Survey 2 ○ Survey 3

Choose Graph Type: ● Line ○ Radar

Choose Graph Option:
● All Role Averages
○ All Department Averages
○ Individual Role Averages

Graph: Role Averages (Survey 2)

Number of Observations: 18

Figure 1. Dimensional data quality assessment across roles.

gest timeliness be measured as the maximum of one of two terms: 0 and one minus the ratio of currency to volatility [2]. Here, currency is defined as the age plus the delivery time minus the input time. Volatility refers to the length of time data remains valid; delivery time refers to when data is delivered to the user; input time refers to when data is received by the system; and age refers to the age of the data when first received by the system.

An exponent can be used as a sensitivity factor, with the max value raised to this exponent. The value of the exponent is task-dependent and reflects the analyst's judgment. For example, suppose the timeliness rating without using the sensitivity factor (equivalent to a sensitivity factor of 1) is 0.81. Using a sensitivity factor of 2 would then yield a timeliness rating of 0.64 (higher sensitivity factor reflects fact that the data becomes less timely faster) and 0.9 when sensitivity factor is 0.5 (lower sensitivity factor reflects fact that the data loses timeliness at a lower rate).

A similarly constructed metric can be used to measure *accessibility*, a dimension reflecting ease of data attainability. The metric emphasizes the time aspect of accessibility and is defined as the maximum value of two terms: 0 or one minus the time interval from request by user to delivery to user divided by the time interval from request by user to the point at which data is no longer useful. Again, a sensitivity factor in the form of an exponent can be included.

If data is delivered just prior to when it is no longer useful, the data may be of some use, but will not be as useful as if it were delivered much earlier than the cutoff. This metric trades off the time interval over which the user needs data against the time it takes to deliver data. Here, the time to obtain data increases until the ratio goes negative, at which time the accessibility is rated as zero (maximum of the two terms).

In other applications, one can also define accessibility based on the structure and relationship of the data paths and path lengths. As always, if time, structure, and path lengths all are considered important, then individual metrics for each can be developed and an overall measure using the min operator can be defined.

Weighted Average. For the multivariate case, an alternative to the min operator is a weighted average of variables. If a company has a good understanding of the importance of each variable to the overall evaluation of a dimension, for example, then a weighted average of the variables is appropriate. To insure the rating is normalized, each weighting factor should be between zero and one, and the weighting factors should add to one. Regarding the believability example mentioned earlier, if the company can specify the degree of importance of each of the variables to the overall believability measure, the weighted average may be an appropriate form to use.

Assessments in Practice

To use the subjective and objective metrics to improve organizational data quality requires three steps (see Figure 2):

- Performing subjective and objective data quality assessments;
- Comparing the results of the assessments, identifying discrepancies, and determining root causes of discrepancies; and
- Determining and taking necessary actions for improvement.

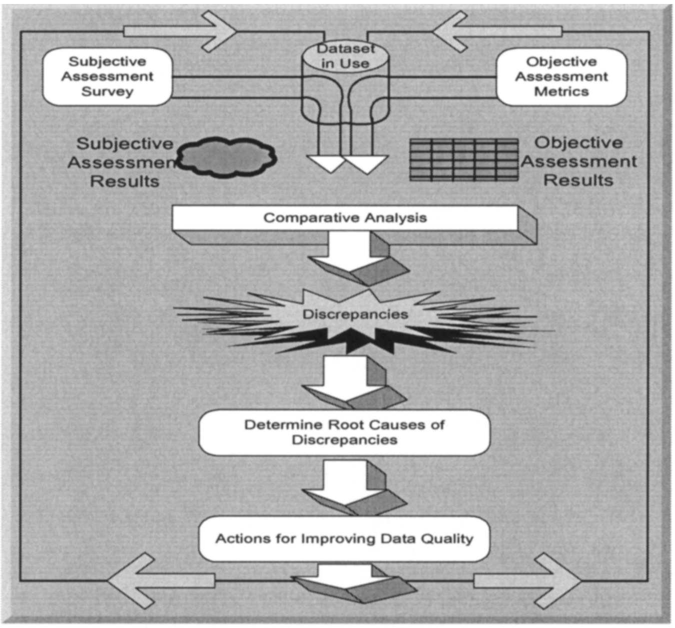

Dataset in Use
Subjective Assessment Survey
Objective Assessment Metrics
Subjective Assessment Results
Objective Assessment Results
Comparative Analysis
Discrepancies
Determine Root Causes of Discrepancies
Actions for Improving Data Quality

Figure 2. Data quality assessments in practice.

To begin the analysis, the subjective and objective assessments of a specific dimension are compared. The outcome of the analysis will fall into one of four quadrants (see Figure 3). The goal is to achieve a data quality state that falls into Quadrant IV. If the analysis indicates Quadrants I, II, or III, the company must investigate the root causes and take corrective actions. The corrective action will be different for each case, as we illustrate using the experiences of two companies.

Global Consumer Goods, Inc., (GCG), a leading global consumer goods company, has made extensive use of the assessments [4]. At GCG, results of subjective assessments across different groups indicated that consistency and completeness were two major concerns. When these assessments were compared to objective assessments of data being migrated to GCG's global data warehouse, the objective measures corroborated the subjective assessment (Quadrant I). This agreement led to a corporate-wide initiative to improve data consistency and completeness. Among the measurements used was a metric measuring column integrity of the transaction tables. Prior to populating their global data warehouse, GCG performed systematic null checks on all the columns of its detailed transaction files. GCG conducted column integrity analysis using a software tool called Integrity Analyzer [5] to detect missing values, which indicated the database state did not reflect the real-world state and any statistical analysis would be useless. Although GCG could simply have measured consistency and completeness on an ad hoc basis, performing the measurements based on the approach presented here enabled GCG to continually monitor both objective measures and user assessments, thereby institutionalizing its data quality improvement program.

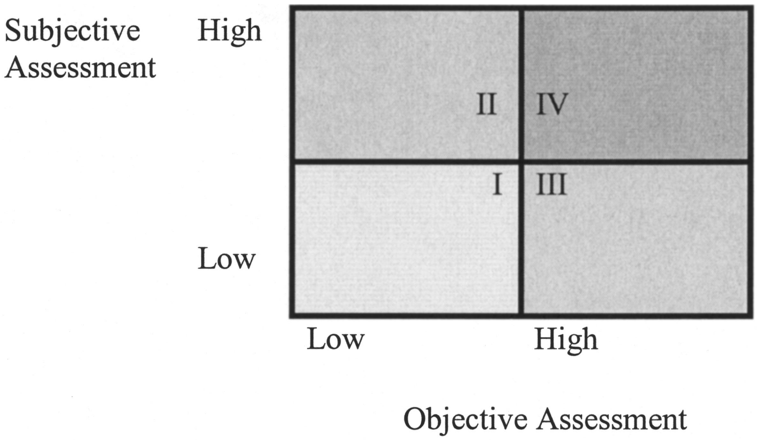


Figure 3. Subjective and objective assessments.

A leading data product manufacturing company, Data Product Manufacturing, Inc., (DPM), which provides data products to clients in the financial and consumer goods industries, among others, illustrates the issue of conflicting assessments. Unlike GCG, DPM found discrepancies between the subjective and objective assessments in its data quality initiative. DPM's objective assessment indicated its data products were of high quality, but its client's assessments (subjective assessments) indicated a lack of confidence in the data products in terms of believability, timeliness, and free of error (Quadrant III). Further analysis revealed the clients' subjective assessments were based on the historical reputation of the data quality. DPM proceeded to implement a data quality assurance program that included training programs for effective use of data. They also incorporated the results of the objective assessments in an overall report that outlined the complexities of client deliverables.

Companies like GCG and DPM that assess subjective and objective data quality go a long way toward answering the question posed at the beginning of this article: How good is my company's data quality? Such assessments also help answer other questions posed by practitioners: How does my data quality compare with others in my industry? Is there a single aggregate data quality measure? If dimensional data quality metrics are developed and assessment data is collected and analyzed over time across an industry, that industry can eventually adopt a set of data quality metrics as a de facto standard, or benchmark performance measure. In the long term, different benchmarks and aggregate performance measures can be established across industries.

In practice, companies wish to develop a single aggregate measure of their data quality—an index of data quality. A single-valued, aggregate data quality measure would be subject to all the deficiencies associated with widely used indexes like the Dow Jones Industrial Average and the Consumer Price Index. Many of the variables and the weights would be subjective. Issues that arise when combining values associated with different scale types (ordinal, interval, and ratio) further complicate matters. But if the assumptions and limitations are understood and the index is interpreted accordingly, such a measure could help companies assess data quality status. From the practitioner's viewpoint, such an index could help to succinctly communicate the state of data quality to senior management and provide comparative assessments over time.

Conclusion

Experience suggests a "one size fits all" set of metrics is not a solution. Rather, assessing data quality is an on-going effort that requires awareness of the fundamental principles underlying the development of subjective and objective data quality metrics. In this article, we have presented subjective and objective assessments of data quality, as well as simple ratio, min or max operators, and weighted average—three functional forms that can help in developing data quality metrics in practice. Based on these functional forms, we have developed illustrative metrics for important data quality dimensions. Finally, we have presented an approach that combines the subjective and objective assessments of data quality, and demonstrated how the approach can be used effectively in practice.

References

1. Ballou, D.P. and Pazer, H.L. Modeling data and process quality in multi-input, multi-output information systems. *Management Science 31*, 2, (1985), 150–162.

2. Ballou, D.P., Wang, R.Y., Pazer, H. and Tayi, G.K. Modeling information manufacturing systems to determine information product quality. *Management Science 44*, 4 (1998), 462–484.

3. Codd, E.F., Relational database: a practical foundation for productivity, the 1981 ACM Turing Award Lecture. *Commun. ACM 25*, 2 (1982), 109–117.

4. CRG, Information Quality Assessment (IQA) Software Tool. Cambridge Research Group, Cambridge, MA, 1997.

5. CRG, Integrity Analyzer: A Software Tool for Total Data Quality Management. Cambridge Research Group, Cambridge, MA, 1997.

6. Huang, K.,Lee, Y., and Wang, R. *Quality Information and Knowledge.* Prentice Hall, Upper Saddle River: N.J. 1999.

7. Kahn, B.K., Strong, D.M., and Wang, R.Y. Information Quality Benchmarks: Product and Service Performance. *Commun. ACM*, (2002).

8. Laudon, K.C. Data quality and due process in large interorganizational record systems. *Commun. ACM 29*,1 (1986), 4–11.

9. Redman, T.C., ed. Data *Quality for the Information Age.* Artech House: Boston, MA., 1996.

10. Wand, Y. and Wang, R.Y. Anchoring data quality dimensions in ontological foundations. *Commun. ACM 39*,11 (1996), 86–95.

11. Wang, R.Y. A product perspective on total data quality management. *Commun.ACM 41*, 2 (1998), 58–65.

12. Wang, R.Y. and Strong, D.M. Beyond accuracy: what data quality means to data consumers. *Journal of Management Information Systems 12*, 4 (1996), 5–34.