Data Quality Dimensions to Ensure Optimal Data Quality

Svetlana Jesiļevska[1]

Quality is more difficult to define for data, moreover the meaning of 'quality' depends on the context in which it is applied. Paper gives a short overview of data quality dimensions which have been collected from literature research.
This paper presents some results of expert survey on data quality issues carried out by the author. The examples illustrate the fact that it is not necessary to use all the various dimensions of data quality provided by researchers, but the most essential data quality dimensions can be combined for a specific application. To support further applications of this approach, this paper contains comparison of data quality requirements to be met from statisticians and data users point of view. The empiric method (analysis of texts and documents) and the method of theoretical research (analysis of the expert survey data) are applied.

Keywords: data quality, data quality dimensions, data users

JEL classification: C00

Defining the problem

Some research results has shown that providing information about the quality of data can impact decision-making (Chengalur-Smith, Ballou, & Pazer, 1999; Ballou & Tayi, 1999) as it gives an opportunity for decision makers to use data more efficiently and effectively (Even, Shankaranarayanan & Watts, 2006). For example, decision makers need sufficient information on data quality in order to assess the reliability of the data (Shankaranarayan, et al., 2003).

There is no agreement on a standard definition of data quality that can be applied across all data domains. The quality of data should be defined in the context of being fit for a particular use. Data 'fitness for use' depends on the application of the data, the characteristics of quality that are necessary for that specific purpose and on the user's expectations of what they define to be useful information. Data quality is a multidimensional concept that is why data quality standards must include a

[1] University of Latvia, Aspazijas blvd. 5, Rīga, Latvia, LV-1050, mozir@inbox.lv

range of quality characteristics that incorporate the concept of 'fitness for use'. These elements of data quality standards should be considered and balanced in the design, implementation, and validation of data management processes and procedures.

The aim of the paper is to present the results of expert survey on data quality dimensions, as well as to identify the most significant data quality dimensions depending on the purpose of data usage. Moreover, based on the expert survey, this paper provides findings of the most essential data quality dimensions from data users and statisticians point of view. The empiric method (analysis of texts and documents) and the method of theoretical research (analysis of the statistical data) are applied.

The problem of low data quality

Nowadays, activities and decisions making in an organization and at the country level is based on statistical data and information obtained from data analysis of this data. Data analysis provides various possibilities for constructing reliable and accurate process for desion-making.

Low quality data can imply a plenty of negative consequences, for example, poor data quality increases operational costs since time and other resources are spent to detect and correct errors. To start with, data mistakes that are not identified and corrected can have extremely negative economic and social impacts on an organization (Ballou et al., 2004; Wang & Strong, 1996). For business users low quality data can have the following negative effects: less customer satisfaction, increased running costs, inefficient decision-making processes, lower performance and lowered employee job satisfaction (Kahn et al., 2003; Leo et al., 2002; Redman, 1998) etc. as data are critical inputs to almost all decisions in an enterprise. Data constitute a significant contributor to organizational culture as a result poor data quality can have negative effects on the organizational culture (Levitin & Redman, 1998; Ryu et al., 2006). Data quality is a survival issue for the Nation Statistical Office (NSO) as poor data quality also means that it becomes difficult to build trust in the NSO's data, which may imply a lack of user acceptance of any initiatives based on such data. Poor data quality has far-reaching effects and consequences both for data users and statisticians.

There is a general agreement in literature that poor quality data is a problem for decision-making both in the country and companies level. For example, much academic literature argue that statistical data of poor quality represent a significant cost factor for many companies, which is supported by findings from several surveys from industrial experts (Marsh, 2005). At the same time, Eppler and Helfert (2004) claim that only very few studies demonstrate how to identify, categorize and assess such costs (i.e. how to determine the causal links between poor data quality and monetary impact). This is supported by Kim and Choi (2003) who state: "There have been limited efforts to systematically understand the effects of low quality data. The efforts have been directed to investigating the effects of data errors on computer-based models such as neural networks, linear regression models, rule-based systems, etc." and "In practice, low quality data can bring monetary damages to an organization in a variety of ways". According to Kim (2002), the types of damage that poor quality data can cause depend on the nature of data, the purpose of the use of data, the types of responses to the damages, etc.

To improve data quality as well as to evaluate the current data quality level, the effect of data quality initiatives have to be measured. Several authors point out that: "Only what can be measured can be improved" (Wand & Wang 1996, Wang & Strong 1996, English 1999). What is needed is a measurement approach to determine the level of data quality over time.

Data quality dimensions

Defining data quality and determining the most essential data quality dimensions, realizing the need for "free of defects" data that contains the right qualities for the task at hand still is difficult to perform.

Klein and Rossin (1999) note there is no single definition of data quality accepted by researchers, statisticians and data users. Some researcher suppose that data quality takes a user-focused view (users are people or groups who have experience in using data to make decisions) that qualitative data is 'data that is fit for use' (Loshin, 2001; English 1999, Redman 2001, Olson 2003; Wang, Strong, & Guarascio, 1996). Data quality is 'contextual'; the user defines what is good data quality for each intended use of the data, within its context of use (Pringle, Wilson, & Grol, 2002; Strong, Lee, & Wang, 1997). The intended use is commonly described as a multi-dimensional concept consisting of a set of quality attributes, called data quality dimensions which are determined by the data users (Wang & Strong 1996). There have been numerous theoretical studies on the identification of data quality dimensions (English 1999, Eppler 2006, Lee et al. 2006, Redman 1996, Wang et al. 1995, Wang et al 1996, Madnick et al 1999, Price & Shanks 2005), the outcome of which mainly were lists and categories of data quality dimensions. Here are some examples.

In the proposal of Wang and Strong, data quality dimensions have been selected by interviewing data consumers, as a result, nearly 179 data quality dimensions have been collected from the user's point of view by means of surveys. Out of those, the authors selected 15 different dimensions and grouped them under four different categories such as Intrinsic, Accessibility, Contextual, and Representational (Wang & Strong 1996).

Data quality dimensions proposed by Wang and Strong (Wang & Strong 1996)

Table 1

Data quality category	Data quality dimensions
Intrinsic data quality	Accuracy, Objectivity, Believability, Reputation
Accessibility data quality	Accessibility, Access security
Contextual data quality	Relevancy, Value-added, Timeliness, Completeness, Amount of data
Representational data quality	Interpretability, Ease of understanding, Concise representation, Consistent representation

The proposal of Bovee et al. (Bovee et al. 2001) considers data quality dimensions by taking the view of data consumers and developed a conceptual model consisting of 4 attributes, namely:

- **Accessibility:** To get information which we might find useful.
- **Interpretability:** To understand the information and find meaning from it.
- **Relevance:** To find it applicable to the domain and the context of interest.
- **Integrity:** To believe it free from defects.

The last attribute Integrity is further classified into four *sub attributes: Accuracy*, *Completeness*, *Consistency* and *Existence* where the last component *Existence* is absent in many studies.

In the proposal of Lee et al (Lee at al 2006), the authors consider a number of data quality dimensions that can be used to assess the data quality:

- **Free of error**- this dimension has been used to check whether the data is correct.
- **Completeness**- The completeness of the data is described with three different perspectives: schema completeness (refers to the degree to which the entities and the attributes are not missing from the schema), column completeness (refers to the missing value in a column of a table), population completeness (refers to the degree to which member of the population that should be present are not present).

- **Consistency-** is viewed in the proposal with difference perspectives such as consistency of redundant data in one table or in multiple tables, consistency between two related elements.
- **Believability-** is described as the extent to which the data is regarded as true and credible.
- **Appropriate amount of data-** this dimension is taken into account as one of the important dimensions, refers to the degree to which the amount of data should be neither too little nor too much.
- **Timeliness-** Timeliness is the extent to which the data is up-to-date with respect to the task for which it is used. Mostly timeliness is related with volatility and currency of the data.
- **Accessibility**- The accessibility dimension reflects the ease of attainability of the data. That is the extent to which the data is easily accessible for the required tasks.

The dimensions discussed in this model by Lee at al. can be of particular interest and importance to many organizations. Moreover, the authors define metrics to measure those dimensions.

Though the proposals discussed above do not cover a wide range of data quality dimensions, these are the commonly used dimensions by the majority of researchers to assess the level of data quality.

The author carried out an expert survey on data quality issues. In the next section, the author provide a set of data quality dimensions essential to assess the leval of data quality depending on the intended use of the data. Moreover, the author provides dimensions according to the perception of data users who have experience in using data to make decisions and statisticians who “produce” data. This survey results review would help to achieve a precise set of data quality dimensions.

Expert survey results

The expert survey questionnaire was paper-based. The survey covered 11 experts: 5 from the academic sector and 6 from the national governmental entities, 3 ot whome were NSO representatives.

In order to define target levels for individual data quality dimensions, the respective usage context for the data has to be analyzed.

The set of data quality dimensions has been tested with experts using four different data usage contexts: data for scientific research, data for decision-making, data for analysis the progress of research object during the reporting period, data for research object modeling and forecasting. Experts were asked to put in the order provided data quality dimensions from the most essential to the least essential, depending on intended use of data. The proposed data quality assessment framework contains 13 dimensions:

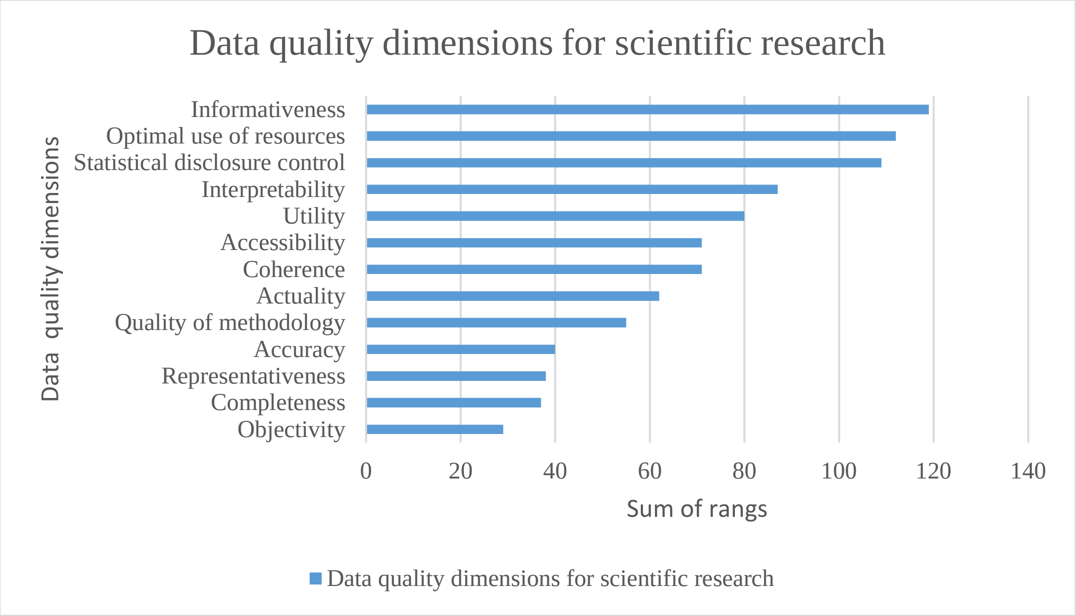
1. Data objectivity- the ability of statistical data to reflect the actual situation and its independence from the data users' interpretations or evaluations.
2. Data completeness- data meets user needs.
3. Data representativity- Sample data generalization capabilities.
4. Data accuracy- The degree of reflection of the actual situation.
5. Quality of methodology- Methodological scientific justification (including methodology approbation), correct use of methodology and unification level of methodology.
6. Coherence- Logical links between different statistical surveys findings, the data from different sources are comparable.
7. Actuality- Data collection and processing speed and frequency of renewal.
8. Data accessibility- Simplicity of data availability to the users.
9. Interpretability- statistical data collection and processing methodologies is available to the data users in order to make the correct interpretation of data.
10. Informativeness- Data presentation form that will enable data users to capture data quickly and easily navigate the data range.
11. Utility- Data users demand to the data.
12. Statistical disclosure control- Confidentiality of the information provided by respondents.
13. Optimal use of resources- Efficient use of existing resources for data collection and processing.

Data quality dimensions for scientific research

Experts indicated that the most important data quality dimensions for data used in scientific research are: data objectivity, data completeness, data representativeness and data accuracy. The least important are: statistical disclosure, optimal use of resources and informativeness.

Figure 1

Hierarchy of data quality dimensions for scientific research (sum of rangs)



Source: Author's construction

From statisticians point of view, the most significant dimensions to assess data quality used for scientific research are: completeness, representativity, objectivity and quality of methodology. Data users put objectivity on the first place, the next com accuracy, representativity and completeness. The least important dimension from both statisticians and data users view is data informativeness.

Hierarchy of data quality dimensions: comparison between statisticians and data users view

Table 2

Statisticians view	Data users view
Completeness	Objectivity
Representativity	Accuracy
Objectivity	Representativity
Quality of methodology	Completeness
Coherence	Actuality
Accessibility	Quality of methodology
Accuracy	Utility
Actuality	Accessibility

Statisticians view	Data users view
Interpretability Statistical disclosure control Optimal use of resources Utility Informativeness	Coherence Interpretability Statistical disclosure control Optimal use of resources Informativeness

Source: Results of the expert survey carried out by author

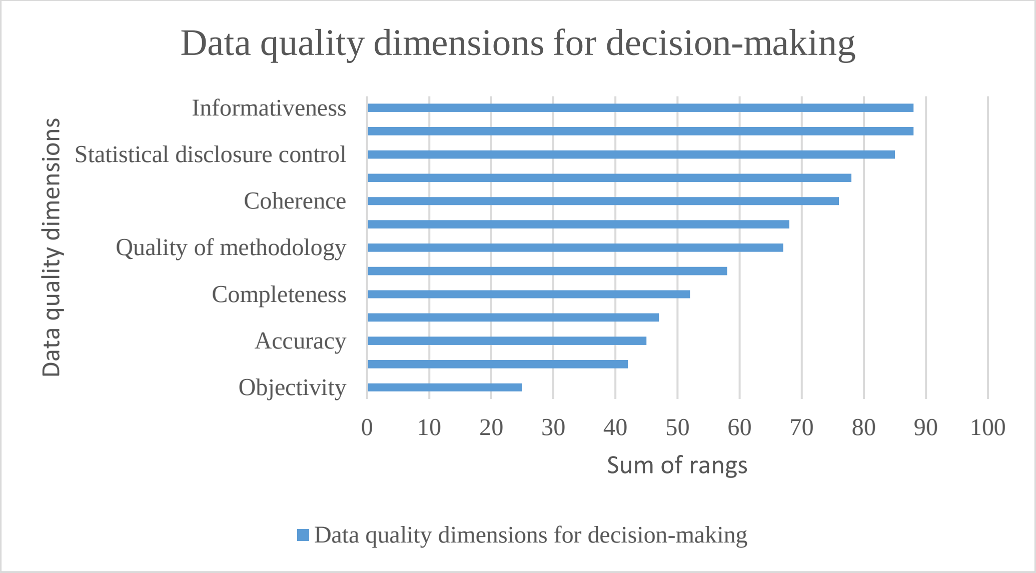
Kendall's coefficient of concordance for ranks (W) calculates agreements between rankers as they rank a number of subjects according to a particular characteristics. Kendall's coefficient of concordance is W=0,607692. $X^2_p=72,92308$ > $X^2_T=26,21697$ The conclusion is therefore that there is a significant agreement between experts.

Data quality dimensions for decision-making

Experts indicated that the most important data quality dimensions for data used for decision-making are: data objectivity, data atuality, data accuracy and data representativeness. The least important are: statistical disclosure, accessibility and informativeness.

Figure 2

Hierarchy of data quality dimensions for decision-making (sum of rangs)



Source: Author's construction

Both statisticians and data users put objectivity on the first place as the most essential dimension to assess data quality for decison-making. From statisticians point of view quality of methodology and accuracy are equally important, but data users put quality of methodology on the bottom of the list of dimensions.

Hierarchy of data quality dimensions: comparison between statisticians and data users view

Table 3

Statisticians view	Data users view
Objectivity	Objectivity
Actuality	Accuracy
Quality of methodology/Accuracy	Representativity
Representativity	Actuality
Informativeness	Completeness
Completeness	Utility
Interpretability	Optimal use of resources
Accessibility	Interpretability
Utility	Conherence
Statistical disclosure control	Quality of methodology
Conherence	Statistical disclosure control
Optimal use of resources	Accessibility
	Informativeness

Source: Results of the expert survey carried out by author

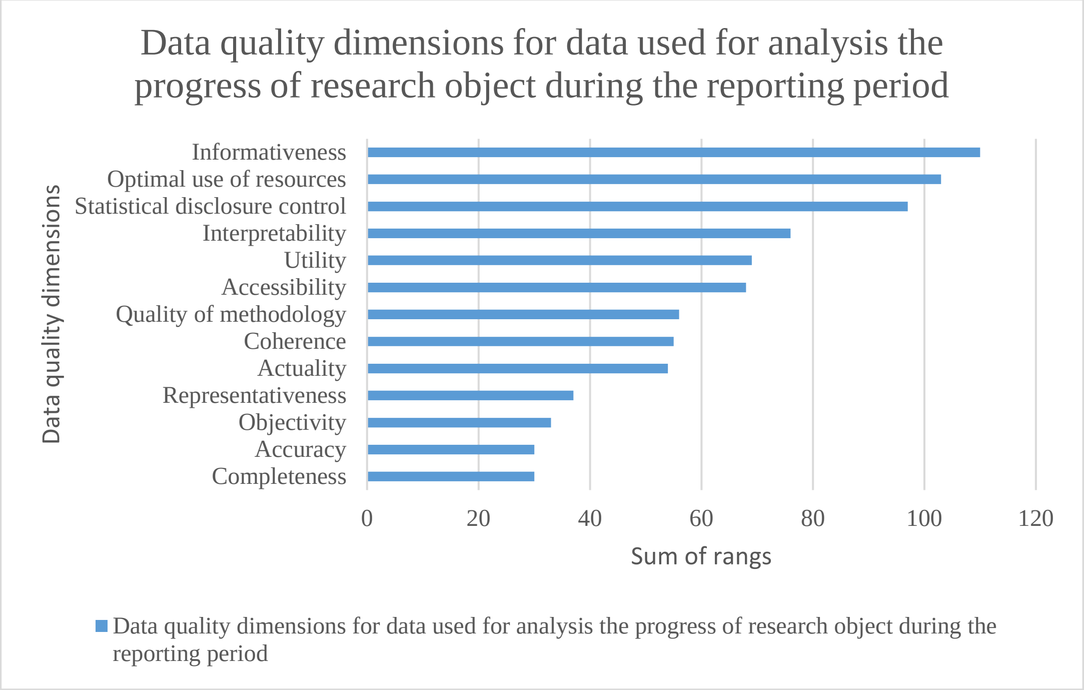
Kendall's coefficient of concordance is W=0,355055. X^2_p=42,60659 > X^2_T=26,21697 The conclusion is therefore that there is a significant agreement between experts.

Data quality dimensions for data used for analysis the progress of research object during the reporting period

Experts indicated that the most important data quality dimensions for data data used for analysis the progress of research object during the reporting period are the following: data completeness, data accuracy, objectivity, representativeness. The least important are: statistical disclosure control, optimal use of resources and data informativeness.

Figure 3

Hierarchy of data quality dimensions for data used for analysis the progress of research object during the reporting period (sum of rangs)



Source: Author's construction

In the context of data used for analysis the progress of research object during the reporting period the most significant differences in the answers of statisticians and users were identified. Statisticians consider quality of methodology as the most important dimension, data users argue that the most essential dimension is completeness. Data users put quality of methodology on the bottom of the list and believe that quality of methodology, accessibility and interpretability are equally less important dimensions.

Hierarchy of data quality dimensions: comparison between statisticians and data users view

Table 4

Statisticians view	Data users view
Quality of methodology Accuracy Completeness Objectivity Representativity Conherence Accessibility Actuality Interpretability/Statistical disclosure control Utility Informativeness Optimal use of resources	Completeness Objectivity/Accuracy Representativity Actuality Conherence Utility Accessibility/Interpretability/Quality of methodology Optimal use of resources Statistical disclosure control Informativeness

Source: Results of the expert survey carried out by author

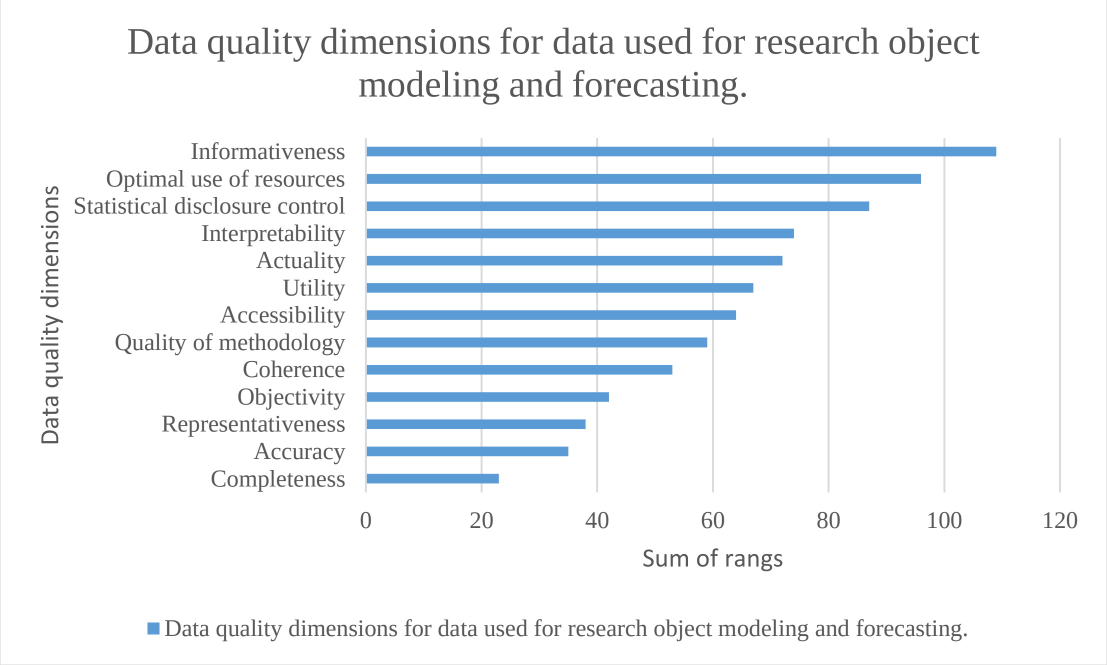
Kendall's coefficient of concordance is W=0,576429. X^2_p=69,17143 > X^2_T=26,21697 The conclusion is therefore that there is a significant agreement between experts.

Data quality dimensions for data used for research object modeling and forecasting.

Experts indicated that the most important data quality dimensions for data are the following: data completeness, data accuracy, data representativeness and data objectivity. The least important are: statistical disclosure control, optimal use of resources and data informativeness.

Figure 4

Hierarchy of data quality dimensions for research object modeling and forecasting (sum of rangs)



Source: Author's construction

Both statisticians and users agree on the most essential and the least essential data quality dimensions.

Hierarchy of data quality dimensions: comparison between statisticians and data users view

Table 5

Statisticians view	Data users view
Completeness	Completeness
Accuracy	Accuracy
Objectivity/Representativity	Representativity
Accessibility	Conherence
Quality of methodology	Objectivity
Conherence	Quality of methodology/Utility
Actuality/Statistical disclosure control	Interpretability/Actuality
Utility	Accessibility
Interpretability	Statistical disclosure control
Optimal use of resources	Optimal use of resources
Informativeness	Informativeness

Source: Results of the expert survey carried out by author

Kendall's coefficient of concordance is W=0,796. X^2_p=95,56 > X^2_T=26,21697 The conclusion is therefore that there is a significant agreement between experts.

Conclusions and future steps

To solve data quality problems effectively, both data users and data producers must use sufficient knowledge about solving data quality problems appropriate for their process areas. At minimum, statisticians must know what kind of data, how (this question includes mainly methodological issues), and why to collect the data; data users must know what data, how (what kind of analysis), and why (intended purpose) to use the data.

In sum, the two main actors mentioned above have roles in a data production process and should cooperate closely to improve statistical data quality. Involvement of both statisticians and data users in the process of identifying and solving possible drawbacks of data opens up new avenues for future research and practice.

References

Ballou, D. P., & Tayi, G. K. 1999. Enhancing data quality in data warehouse environments, Communications of the ACM, vol. 42 no. 1, pp. 73-78.

Ballou, D. P., Madnick, S., & Wang, R. (2004). Assuring information quality. *Journal of Management Information Systems*, 20, 9–11.

Bovee M., Srivastava R. R., Mak B.R (2001): A Conceptual Framework and Belief Function Approach to Assessing Overall Information Quality. In proceedings of the 6th International Conference on Information Quality

Chengalur-Smith, I., Ballou, D. P., & Pazer, H. 1999. The impact of data quality information on decision making: an exploratory analysis, IEE Transactions on Knowledge and Data Engineering vol. 11, pp. 853-864.

English, LP (1999) Improving Data Warehouse and Business Information Quality. John Wiley & Sons, Inc., New York, NY.

Eppler, M., & Helfert, M. (2004). A classification and analysis of data quality costs. MIT International Conference on Information Quality, November 5-6, 2004, Boston.

Eppler, M. J. Managing Information Quality, 2 ed. Springer, 2006

Even, A. Shankaranarayanan, G. Watts, S. 2006. Enhancing Decision Making with Process Metadata: Theoretical Framework, Research Tool, and Exploratory Examination. Proceedings of the 39th Hawaii International Conference on System Sciences

Kahn, B., Strong, D., & Wang, R. (2003). Information quality benchmarks: Product and service performance. *Communications of the ACM*, 45, 184-192.

Kim, W., & Choi, B. (2003). Towards Quantifying Data Quality Costs. *Journal of Object Technology*, 2(4), 69-76.

Kim, W. (2002). On Three Major Holes in Data Warehousing Today. *Journal of Object Technology*, 1(4), 39-47.

Klein, B., & Rossin, D. F. 1999. Data errors in neural network and linear regression models: An experimental comparison, Data Quality vol. 5, p. 25

Lee, Y. W., Pipino, L. L., Funk, J. D., Wang, R. Y. Journey to Data Quality. MIT Press. Boston, 2006

Leo, L., Pipino, L. Yang, W. L., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, 45(4), 211-218.

Levitin, A. V., & Redman, T. C. (1998). Data as a resource: Properties, implications, and prescriptions. *Sloan Management Review*, 40(1), 89-101.

Loshin, D. 2001. Enterprise Knowledge Management. The Data Quality Approach California: Academic Press. p. 493

S. Madnick, R. Wang, Y. W. Lee and H. Zhu, "Overview and Framework for Data and Information Quality Research", ACM Journal of Data and Information Quality, 1 (2009).

Marsh, R. (2005). Drowning in dirty data? It's time to sink or swim: A four-stage methodology for total data quality management. *Database Marketing & Customer Strategy Management*, 12(2), 105-112.

Olson, J (2003) Data Quality -The Accuracy Dimension. Morgan Kaufmann, San Francisco

R. Price and G. Shanks, "A semiotic information quality framework: development and comparative analysis", Journal of Information Technology, 20 (2005), pp. 88-102.

Pringle, M., Wilson, T., & Grol, R. 2002. Measuring "goodness" in individuals and healthcare systems, British Medical Journal. vol. 325, pp. 704-707.

Redman, T. C. Data Quality for the Information Age. Artech House. Boston, 1996

Redman, T.C. (1998). The impact of poor data quality on the typical enterprise. *Communications of the ACM*, 41(2), 79-82.

Redman, T. C. 2001. Data Quality. The Field Guide. Boston: Digital Press.

Ryu, K.-S., Park J.-S., & Park, J.-H. (2006). A data quality management maturity model. *ETRI Journal*, 28(2), 191-204.

Shankaranarayan, G., Ziad, M., & Wang, R. Y. 2003. Managing data quality in dynamic decision environments: an information product approach. Journal of Data Management, vol 14, no. 4, pp. 14-32

Strong, D. M., Lee, Y. W., & Wang, R. Y. 1997. Data quality in context, Communications of the ACM, vol. 40, pp. 103-110.

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

Wand, Y and Wang, RY (1996) Anchoring Data Quality Dimensions in Ontological Foundations. Communications of the ACM 39 (11), 86-95.

Wang, R. Y., Storey, V. C., Firth, C. P. "A Framework for Analysis of Data Quality Research." IEEE Transactions on Knowledge and Data Engineering, 7 (4). 1995. pp. 623-640.

Yang W. Lee Leo L. Pipino James D. Funk Richard Y. Wang: Journey to Data Quality, 2006 Massachusetts Institute of Technology