Data Quality Evaluation: Methodology and Key Factors

Ying Yang[1], Yuan Yuan[2], and Bo Li[2(✉)]

[1] Standardization Department, China Aero-Polytechnology Establishment, Aviation Industry of China, Beijing, China
yangyingpfy@163.com
[2] School of Computer Science and Engineering, Beihang University, Beijing, China
libo@act.buaa.edu.cn

Abstract. Data Quality Evaluation is becoming an institutionalized stage in data quality lifecycle. More and more practice is promoted by data management and user organization in specific fields especially in better informationalized application circumstance.

In order to improve the ability of data quality evaluation, the paper presents the key factors for data quality assessment and measurement. On the base of analyzing the main methodologies and standards on data quality management, the key factors includes objectives, general principles, characteristics, measurement function etc.

Keywords: Data quality evaluation · Measures · Standards · Characteristics

1 Introduction

Data Quality Evaluation is becoming an institutionalized stage in data quality lifecycle. More and more practice is promoted by data management and user organization in specific fields especially in better informationalized application circumstance.

Recently, the data quality assessment and measurement are utilized to improve the efficiency systematically. Methodology and key factors of data quality assessment and measurement related to the general concepts, terminology, objectives, procedure, model, principles, characteristics and measures.

Based on the achievement from the practice in authors' investigation project, the relative national or international standards focus on geographic, environmental and software field, the paper provide useful guidance for data quality evaluation scenarios over data lifecycle, and provide the fundamental input to Big Data service etc.

M. Qiu (Ed.): SmartCom 2017, LNCS 10699, pp. 222–230, 2018.
https://doi.org/10.1007/978-3-319-73830-7_22

2 Data Quality Management Methodology

2.1 Total Data Quality Management (TDQM) in Department of Defense (DoD)

DoD TDQM methodology [1] is intended to validate data quality problems, indentify root causes, and improve data quality and utility. TDQM is a process to support database migration, promote the use of data standards, and improve in conformance to business rules. TDQM approach conforms to TQM methodologies, integrates management techniques, improvement efforts, and technical tools to create and sustain a culture that is committed to continuous improvement. To attain TDQM objectives, data quality work includes four essential tasks: definition, measurement, analysis and improvement.

Definition phase: data quality problem are defined by establishment of the scope and objectives of the data quality management project, and by judging whether criteria conform to relative standards.

Measurement phase: data quality measurements should present qualitative indexes and quantitative indexes, judging whether conformance to standards or not, and flag exceptions or suspicious datum.

Analysis phase: identification, priority and validation of quality issues are the common analysis procedure. Providing relative recommendations to solve the issues of data quality problem.

Improvement phase: improvement projects are defined and choosed the opportunities to implement them. Improving data quality may lead to changing data entry procedures, updating data validation rules, using standards that prescribe a uniform representation throughout the DoD.

2.2 ISO 8000 Data Quality Management (DQM)

ISO 8000 DQM methodology is a family standards focus on systematic solution. The series standards include part 001-99 general principle, part 100-199 master data, part 200-299 business data quality, part 300-399 product data quality. Under ISO 8000, part 150 A Framework for Data Quality Management is very important which presents three principles. Firstly, data quality management is not merely a technology implementation. Secondly, effective management is based upon a number of key processes. Lastly, achieving continuous improvement of the data quality is the key objective. The methodology is summarized as nine box model [2] describing the roles, process and responsibility.

Figure 1 illustrates the roles, responsibilities and functions of data quality management reflected stakeholders of data quality.

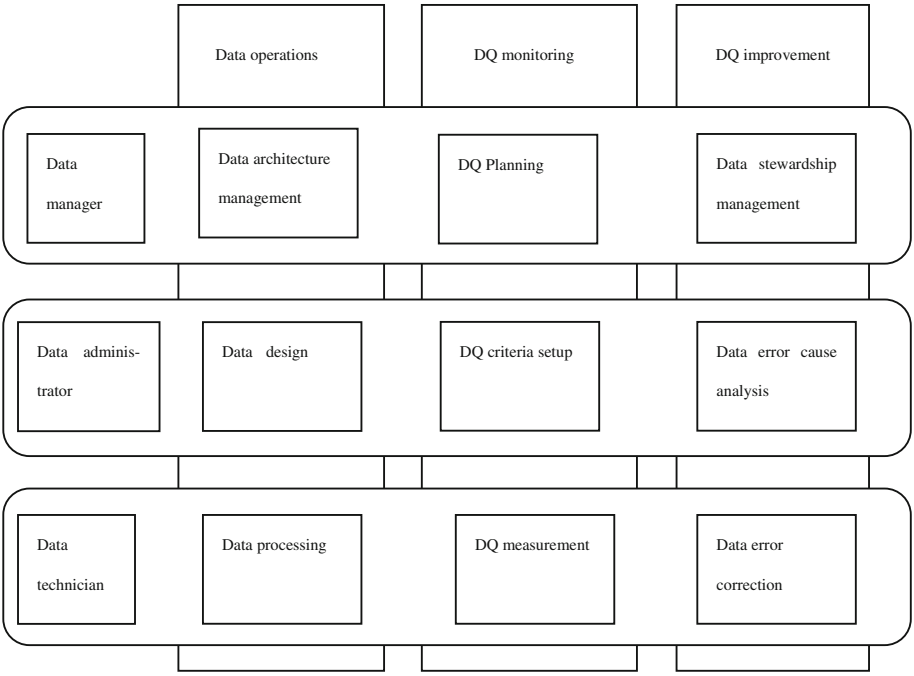

Fig. 1. The model of data quality management [2]

This model framework is divided into three processes and three roles. Three generic roles consist of data manager, data administrator and data technician. Three key processes and tasks are as follows:

Data operations consist of data architecture management, data design, data processing, focusing on the application.
Data quality monitoring consist of data quality planning, data quality criteria setup, data quality measurement, focusing on the assessment.
Data quality improvement consist of focus on data stewardship and flow management, data error cause analysis, data error correction, eliminating causes and correcting errors of data.

3 Key Factors of Data Quality Evaluation

3.1 The Objectives of Data Quality Evaluation

3.1.1 Compliance to Organizational Management Requirements

Satisfaction degree evaluation of data quality is provided according to measure the compliance to requirements. Baseline criteria are established within organizational information systems, then inspected periodically by specific regulations. The result of comparing with baseline is depended on attaining appropriate data or data sets.

3.1.2 Root Causes of Problems Within Organizational Environment

The evaluation offers opportunity to identify reoccurring issues damaging data quality. Some key questions can be answered just like: Do certain types of errors occur more frequently? What efforts are concentrated so as to get tgreater improvement in data quality? Determining root causes, there might be analysis from several points of view, including:

Process Problem: Data errors can be attributed to process problems. Checking the existing processes supported data entry, assignment and implementation of data quality responsibilities are suggested. These actions are recommend to correct deficiencies.
System Problem: Data problems frequently originate from system design deficiencies by poorly documented modifications, imperfect user training, or system beyond their original intent.
Policy and Procedure Problem: Data errors reveal lack of appropriate guidance, conflicting in existing directives, instructions and standards.
Data Design Problem: Data problem can be attributed to incomplete designation of database, errors of data values, incomplete specification of technical and business rules. For example, the inappropriate application of primary key constraints, referential integrity specifications, metadata specifications, null and not null data criterion etc.

3.1.3 Supporting Persistent Improvement of Data Quality

The recommendations of data quality improvement are categorized by four tasks reflecting multi-dimensions viewpoint of stakeholders. The recommendations are as follows:

Process Improvement: Concentration on the functional processes can change centralized data entry and data collection, in order to eliminate non value activities.
System Improvement: Data environment is depended on system. Software, hardware, and telecommunication changes can be utilized for improvement of data quality.
Policy and Procedure Improvement: Development of appropriate guidance for roles and responsibilities can be increase quality. Institutionalization such behaviors, for instance adding the periodic data quality examination into operating procedure, can promote business efficiency and data quality.
Data Design Improvement: Establishment and implementation of data standards aid in enhancing the overall data design capacity.

3.2 The General Principle of Data Quality Evaluation

3.2.1 The General Evaluation Principles

The general data quality evaluation principles include scientificity, intelligibility, objectivity and operationality.

Scientificity: Under the guidance of theories of quality management, data management, and standardization management, based on actual data collection, data acquisition, data storage and data exchange, the evaluation scheme shall be rational, well-arranged and operational.
Intelligibility: The evaluation scheme shall be concise, pertinent, plain and understandable, assuring that all concerned stakeholders and relevant personnels precisely comprehend, organize, and regulate the data quality assessment.
Objectivity: The evaluation scheme is supposed to be objective and reliable, authentically reflecting the fulfillment of need and the quality of data instead of blind enlargement and reduction of the evaluation scope.
Operationality: The evaluation scheme shall be specific, feasible for fear of unnecessary interferences. Data needed in assessment shall be easy to acquire and the evaluation process shall be objective, concise and convenient, maximizing the aid of automation tools and vehicles.

3.2.2 The Design Principles of Characteristics and Measures

The designation of data quality characteristics and measures conforms to the principle of integrity, openness, conciseness, and Orthogonality.

Integrity: In accordance with the characteristics and requirements of evaluation object, qualitative and quantitative meters shall be combined to form a comprehensive, systemic, and integral meter system.
Openness: Based on the actual situation, if accepted by the concerned stakeholders, the evaluation meter and weight can be adjusted, added and deleted and the evaluation process and its subprocess can be iterated.
Conciseness: Key points shall be simplified and highlighted, guaranteeing the selection of evaluation meter is sufficient and necessary.
Orthogonality: The design of meter and its corresponding data collection shall avoid ambiguity and overlap.

3.2.3 The Common Steps of Evaluation Procedure

The data quality evaluation tasks consist of four common steps [3]: definition of evaluation requirement, implementation of measurement, analysis result of assessment, implementation of improvement. It should be point out that iteration anyone step is permitted whenever needed, and implementation of improvement are performed by the data user organization after the prior evaluation according to current requirement and environment. Implementation of improvement response the effectiveness of the task of data quality evaluation to some extent.

Figure 2: illustrates the obligatory steps of data quality evaluation

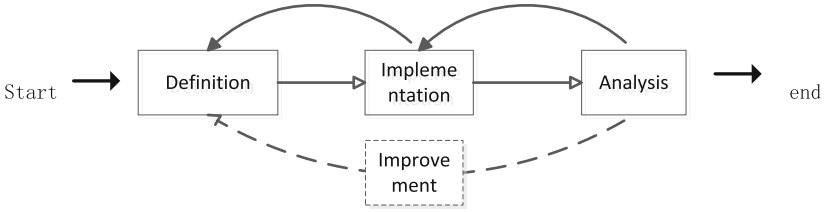

Fig. 2. The obligatory steps of data quality evaluation

3.3 The Data Quality Evaluation Dimensions

There are many methodologies and dimensions in data quality evaluation among specific field, especially in better informationalized circumstance. Abstraction from different data quality measurement practice, there are four categories characteristics: process, inherent, system dependent and user' satisfaction.

Figure 3: illustrates the relationships of process quality measures, data quality measures and quality in use measures.

Quality measures from inherent point of view: inherent characteristics may be applied to data itself, especially to data domain values and possible restrictions, for instance business rules governing the quality required for the characteristic, relationships of data values, metadata.

Quality measures from system dependent point of view: system dependent characteristics may be applied to quantify the influence on information technology applied in systems including software, hardware, network etc.

Relationship between types of quality measures from process and efficiency point of views: Data are expected to be correlated with other quality measures. High quality of the development and maintenance process is able to realize high quality of data, and data quality influences quality in use which represents the effect perceived by the final user [4].

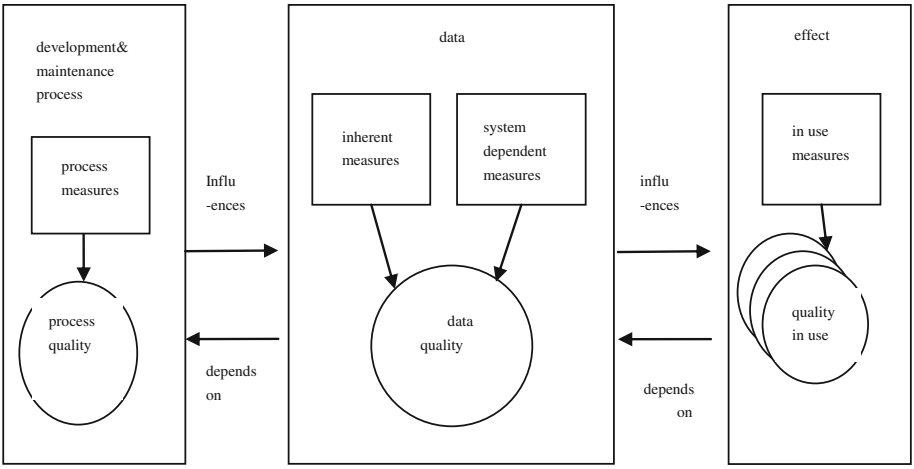

Fig. 3. Relationship between types of quality measures [4]

3.4 The Measures of Data Quality Characteristics [5]

According to the general principles of data quality evaluation, the data quality characteristics in this document are focus on 10 characteristics generally choosed to measurement in accordance with the inherent, system dependent points of view.

Completeness: Completeness measures provide the degree to which data associated with a target entity has values for all expected attributes in a specific context of use. Completeness includes attribute completeness, record completeness, data value completeness, data file completeness, metadata completeness etc.

Example: completeness of attribute for metadata [4] X
measurement function X = A/B
A = number of attributes with complete metadata within the data dictionary
B = number of attributes for which metadata are expected within the data dictionary

Consistency: Consistency measures provide the degree to which data attributes that are free from contradiction and coherent with other data. Consistency includes data format consistency, referential integrity, architecture consistency etc.

Accuracy: Accuracy measures provide the degree to which data has attribute that correctly represent the true value of the intended attributes. Accuracy includes data model accuracy, metadata accuracy, semantic data accuracy, syntactic data accuracy etc.

Example: metadata accuracy [4] X
measurement function X = A/B
A = number of metadata that provides the requested information
B = number of metadata defined within the specification of data

Compliance: Compliance measures provide the degree to which data has attributes that adhere to standards, regulations and conventions. Compliance includes regulatory compliance of value or format etc.
Currentness: Currentness measures provide the degree to which data has attributes that are of right stage. Currentness includes update frequency, timeliness of update etc.
Credibility: Credibility measures provide the degree to which data has attributes which considered as true and believable. Credibility includes values credibility, source credibility etc.

Example: source credibility [4] X
measurement function X = A/B
A = number of data values certified by a qualified organization.
B = number of data values for which source credibility can be defined.

Confidentiality: Confidentiality measures provide the degree to which data has attributes that only accessible by authorized users. Confidentiality includes encryption usage, privacy protection etc.
Traceability: Traceability measures provide the degree to which data has attributes that support an audit trail access of changes made to data. Traceability includes users access traceability etc.
Efficiency: Efficiency measures provide the degree to which data has attributes that can be processed the expected levels of performance. Efficiency includes usable efficiency etc.

Example: s usable efficiency [4] X
measurement function X = A/B
A = number of data values that intended users evaluate as "easily used"
B = number of data values evaluated by users.

4 Conclusion

This paper provides the fundamental solution of data quality evaluation that focus on the methodology, principle, procedure, model, characteristics and measures. The data quality assessment and measurement are becoming more and more popular in data management and quality management in order to improve the efficiency systematically. The concerning factors in this paper are abstracted from the practice in author's investigation project, and on the basis of achievement proposed in industrial standard s, national standards, international standards, that focus on geographi, environmental and software field etc. The information in this paper will provide useful guidance for data quality evaluation scenarios over data lifecycle, and provide the fundamental input to Big Data service etc.

References

1. DOD Guidelines on Data Quality Management (Summary), 31 July 2003
2. ISO/TS 8000:150 – A Framework for Data Quality Management (2011). http://www.dpadvantage.co.uk
3. McGilvray, D.: Executing Data Quality Project, Ten Steps to Quality Data and Trusted Information (2008)
4. ISO/IEC DIS 25024 – Systems and Software Engineering - Systems and Software Quality Requirements and Evaluation - Measurement of Data Quality (2015)
5. Loshin, D.: The Practitioner's Guide to Data Quality Improvement (2011)