An Overview on Quality Evaluation Constitution in Context of Big Data Application

Zhe Li\textsuperscript{1,a}, Sai Wu\textsuperscript{1,b}, Hongwei Zhou\textsuperscript{2,c}, Sheng Zou\textsuperscript{2,d} and Tingting Dong\textsuperscript{1,e}

\textsuperscript{1}China Electric Power Research Institute, Beijing, 100192, China
\textsuperscript{2}State Grid Jiangsu Economic Research Institute, Nanjing, 210000, China
\textsuperscript{a}Email: lizhe@epri.sgcc.com.cn
\textsuperscript{b}Email: wusai@epri.sgcc.com.cn
\textsuperscript{c}Email: zhouhongwei@js.sgcc.cn
\textsuperscript{d}Email: zus@js.sgcc.com.cn
\textsuperscript{e}Email: xintong-dongting@epri.sgcc.com.cn

Abstract. Big data applications have been applied in various domains. With the development of big data in electric power, the application of smart grid is developing towards a more intelligent and coordinated direction. However, except for the functionality realization, the non-functional requirement such as data quality become an more and more important issues. For the reason that it relate to the validation of final results of function in big data area. The criterial of quality evaluation is crucial to the evaluation for data quality, and also the standardizations the scientific management of big data in enterprises can be used as the guide the construction of new information infrastructure. This paper investigates the existing data quality evaluation dimensions, model and application domain, and combines with the practice in industry, analyses the standards, dimensions and methods and then gives suggestions on promoting the data quality evaluation in specific domain.

1. Introduction

In recent years, Data-Driven Systems have been widely used, but assumption hold for these systems is that the collected data is accurate. Unfortunately, this assumption does not always hold in the real world. There are many kinds of errors exist in the data stored in the information system, such as incorrectness, inconsistency, redundancy, imprecision, incompleteness or overdate.

Statistics shows that there are 1%-30% data errors and deviation in various American enterprises\cite{1}. As reported of data quality in medical databases, 13.6%-81% of the important data are incomplete or overdate\cite{2}. According to Gartner, among global Fortune 1000 companies more than 25% of important data are incorrect or inaccurate\cite{3}.

With the increasing improvement and perfection between data and real world, the concept of "data is assets" has been widely accepted, so that the importance of data has been raised to an unprecedented level. However, not all data can be assets, the value of data is mainly rely on the data quality.

Data quality emerges with the development of information system. The quality of data have impact on the operation quality of information system. Therefore, people realize the importance of data quality to information system, and thus open the prelude of data quality research. Over the past 30 years, data quality has gradually become a professional research field and many important research results have emerged.

The rest of the paper is organized as follows. Section2 show a general picture of data quality in Research and industry while Section 3 will review the dimension of data quality assessment. However,
the model build on the data quality dimension is illustrated in Section 4. After then in Section 5 the framework combine with the model and different dimension is discussed, at the end, Section 6 summarizes our conclusions and highlights future direction.

2. Overview of Data Quality in Research and Industry
Data quality has been paid lots attention in both research and industry area, for that it has become a key hinder which constraint the development of data industry. In this case, the research and industry practice on data quality become an important proposition in the application of big data. The research on data quality begin around the 1970s, after more than 30 years of development, a series of classical theories, technologies and methods have been formed. The development process can be roughly divided into three stages: the initial stage, the stable stage and the prosperity stage of data quality. Many important research results have emerged in each stage, such as the definition of data quality, the comprehensive data quality management proposed by MIT[4], the evaluation framework of data quality, and the ISO8000 data quality standard[5] etc. After 2010, new technologies such as Internet of Things (IOT), Cloud Computing and Big Data have a rapidly growth, which brings great challenges and opportunities to the research of data quality.

In research area, the research of data quality is more deep and concrete, many important theories and results have emerged. With the wide use of database and data warehouse, solving the problem of data quality on the use of database or data warehouse technology[6] become a research hotspot. In the field of database, data integrity, data accuracy, data reliability and data consistency, are the cornerstone of quality measurement, which can be used to prevent the existence of detrimental data in the database and avoid invalid operation of input/output caused by invalid data. A variety of methods to ensure data integrity, including primary keys, foreign keys, constraints, rules and triggers have been applied in the database. In addition to data integrity, users usually need to consider the relationship between attributes in database design, that is, dependency, which is a concept of semantic category and can only be determined according to the semantics of data. At present, the most commonly used dependencies are functional dependencies and inclusion dependencies, which can effectively avoid all kinds of data anomalies in the database. In 2007, Professor Fan Wenfei of Beihang proposed conditional functional dependency and conditional inclusion dependency, which enhanced the expressive ability of functional dependency and inclusion dependency to actual semantics through conditional table constraints[7]. After conditional functional dependency was proposed, it takes great influence among database researchers and also had a great impact on data cleaning and data quality improving.

3. Overview on the Dimension of Quality Assessment
In the 1990s, under the leadership of Professor Stuart Madnick and Professor Richard Y Wang, the data quality research group of MIT published their work on the theory of Total Data Quality Management (TDQM)[8], which is the first time that a complete theoretical system of data quality was published in the disciplinary level, and made a good theoretical basis for the comprehensive development of data quality in the future. The stage of this research focuses on define data quality, determining quality dimensions and enrich the TDQM theory. It has not yet formed a unified definition of data quality in academic area. However, the definition proposed by MIT is more acceptable one. They adopt the concept of fitness for use and define data quality as data is suitable for the use of data consumers. The judgment of data quality depends on the individual who uses the data, and the suitability of different people in different environments is different. Redman, an expert in data analysis, believes that data is of high quality if it can be used to meet the needs of customers in operation, decision-making and planning. According to this definition, the customer is the final arbiter of quality[9].

While the quality dimension is the basis of the data quality assessment, it is necessary to review the different dimension and it’s application in different domain. In this paper we review the dimension from the point of view by international organization and domestic organization as show in Table 1 and Table 2.
Table 1. Common Data Quality Dimensions of International Institutions and Government Departments

| International institutions or government departments | Data Quality Dimension |
|-------------------------------------------------------|------------------------|
| International Monetary Fund                           | Guarantee of integrity, soundness, accuracy and reliability of methods, applicability and accessibility |
| Eurostat                                              | Relevance, accuracy, comparability, coherence, timeliness, punctuality, accessibility and clarity |
| Food and Agriculture Organization of the United Nations| Relevance, accuracy, timeliness, punctuality, accessibility and clarity, comparability, consistency and integrity, and completeness of source data |
| Federal Government of the United States (Public Communication) | Practicability, objectivity (accuracy, reliability, clarity, completeness, no ambiguity) safety |
| U.S. Department of Commerce                           | Comparability, accuracy and applicability |
| U.S. Department of Defence                            | Accuracy, completeness, consistency, timeliness, uniqueness and effectiveness |
| Statistics Canada                                      | Accuracy, timeliness, applicability, accessibility, cohesion and interpretability |
| Australian Bureau of Balance of Payments Statistics   | Accuracy, timeliness, applicability, accessibility and scientific method |

After the overview of the international company, we give out the data quality dimension used in Chinese company as it is shown in Table 2:

Table 2. Several evaluation methods in Chinese company

| Industry                                 | Data Quality Dimension |
|------------------------------------------|------------------------|
| Tobacco                                  | Accuracy, completeness, consistency, timeliness, interpretability and accessibility |
| Meteorological Communication              | Scientifically, standardization, sharing, timeliness, stability and maintainability |
| Military                                 | Completeness, consistency, accuracy, uniqueness, timeliness and interpretability |
| Medical Care                             | Consistency, reliability, availability and applicability |
| Transportation                          | Integrity, validity, accuracy and real-time |
| Geographic Information System (GIS)      | Position accuracy, current situation, consistency, integrity and reliability |

As illustrated in the tables above, we can get the conclusion that the some of the dimensions in the table are mainly for data quality assessment in specific domain, while others are suitable for the evaluation of information systems or collaborative information systems within enterprises.

4. Overview on Data Quality Assessment Method

Data quality assessment methods are mainly divided into qualitative method, quantitative method and comprehensive method. Qualitative method mainly depends on the subjective judgment of the experts. However, quantitative method provides a systematic and objective quantitative analysis method, the results are more intuitive and concrete. The comprehensive method combines qualitative and quantitative methods to take both advantages of these two kinds of method.
Qualitative evaluation methods are based on certain evaluation criteria and requirements. According to the purpose of evaluation and the needs of users, data resources are described and evaluated qualitatively. Specifically, determine the relevant evaluation criteria, give the evaluation results by roughly evaluating the evaluation target. The evaluation results are expressed by hierarchy and percentage system. Generally, qualitative evaluation can be divided into user feedback method, expert evaluation method and third-party evaluation method.

Quantitative evaluation method refers to the optimization and evaluation of data quality dimensions from the perspective of objective quantification according to quantitative analysis method. Quantitative method provides a systematic and objective quantitative analysis method for the data quality assessment, and the results are more accurate and concrete.

The comprehensive method combines qualitative and quantitative methods organically and evaluates the quality of data resources from different aspects. The commonly used comprehensive evaluation methods are: Analytic Hierarchy Process (AHP)[10], Fuzzy Comprehensive Evaluation (FCE)[11], Cloud Model (CM)[12] and Defect Subtraction Score (DSS)[13]. However, the different assessment method may have different characteristic and can be applied in different scenario, it is important to review the related attribute for the future application, which are shown in the following table:

| Assessment Method | Difficulty Level | Used Model                  | Application Scenario                        | Scope of Fitness     |
|-------------------|------------------|-----------------------------|--------------------------------------------|----------------------|
| AHP               | simple           | Hierarchical structure model| The Weight of quality indicators are preset | No limitation        |
| FCE               | complex          | Membership function         | Fuzzy Quality Problem                      | No limitation        |
| CM                | complex          | Normal Cloud Model          | Fuzziness and Random of Quality Problems   | No limitation        |
| DSS               | simple           | No limitation               | Product quality                           | Specific areas of expertise |

AHP method needs to decompose complex problems into several levels, establish a hierarchical structure, and then carry out hierarchical ranking and consistency test to draw a conclusion. AHP is often used to determine the weight of indicators in the field of data quality for its simplicity and wide application scope. FCE method is based on fuzzy mathematics and is used to deal with the quality problems of fuzziness. By establishing membership function to make comprehensive evaluation of things, it can be applied in the domain of product quality evaluation, scientific and technological achievements evaluation, port environmental evaluation and so on. In the real world, many concepts of things are uncertain, and CM theory is more suitable for a quality problem with fuzziness and randomness coexisting. CM combines probability theory with fuzzy set theory, forms a transformation model between qualitative concepts and their quantitative representations, and reveals the intrinsic correlation between randomness and fuzziness. Defect deduction refers to the method of evaluating product quality by calculating the score value of unit product (data or information). However, in practice, this method also has many limitations and shortcomings. It can only be applied to the quality evaluation of structured data in some professional fields, such as spatial data, but not in comprehensive evaluation.

5. Overview on Data Quality Assessment Framework
During 2000 and 2009, the data quality research and application encountered a comprehensive development stage. In which, the data quality research goes to deeper, a large number of related data quality products have emerged, international organizations began to research and develop data quality
standards, and government departments promulgated data quality bill. In order to measure and evaluate data quality and information quality, MIT first proposed an Assessment Information Management framework; the International Monetary Fund (IMF) also proposed a Data Quality Assessment Framework (DQAF)[14], which can be widely used in the unification of member countries. Evaluate and improve the quality of data. After that, more than ten evaluation frameworks have emerged, enriching the relevant research results.

Table. 4 Evaluation Framework and Leader Scientists

| Abbreviation of Evaluation Framework | Full name | Major founders | Publish time |
|-------------------------------------|-----------|----------------|--------------|
| AIMQ[15]                            | A methodology for information quality assessment | Lee | 2002 |
| CIHI[16]                            | Canadian Institute for Health Information methodology | Long & Seko | 2002 |
| IQM[17]                             | Information Quality Measurement Activity-based Measuring and Evaluating of product information Quality(AMEQ) | Eppler & Munzenmaier | 2002 |
| AMEQ[18]                            | Evaluating of product information Quality(AMEQ) | Su & Jin | 2007 |
| DaQuinCIS[19]                       | Data Quality in Cooperative Information Systems | Scannapieco | 2004 |
| QAFD[20]                            | Methodology for the Quality Assessment of Financial Data | De Amicis & Batini | 2004 |
| CDQ[21]                             | Comprehensive methodology for Data Quality management | Batini & Scannapieco | 2006 |

As it is shown in Table. 4, AIMQ quality (AIMQ) to form a basis for Information quality (IQ) assessment and benchmarking, this model has four quadrants, depending on whether information is considered to be a product or a service, and on whether the improvements can be assessed against a formal specification or customer expectation[15]. In the CIHI-DQF framework, the concept of data quality is operationalized as a multi-level model and the framework includes an instrument and scoring algorithm for data quality measurement[16]. The information quality measurement (IQM) methodology should ensure that the measurement tools are used correctly, that is to say that they measure the rights things in the right manner[17].

6. Conclusion
This paper mainly focus on three facets, one is the dimension of the data quality, the other one is the model or method to assess the data quality in big data context, based on these dimension, the last facet is the framework about how to use these dimension and build a model to assess the data quality in different domain. The framework mentioned above is mainly focus on the information quality assessment, so that it can be an important reference of the practice in different industry area. However, in big data era, the content and relationship of big data change fast, how to fit the fast data quality assessment is an important issue can be studied in the future.

7. Acknowledgments
This work was supported by State Grid Technology Project Grant (B3441618K002).

8. References
[1] Best, M., The new competitive advantage: the renewal of American industry. OUP Catalogue, 2001.
[2] Cismondi, F., et al., *Missing data in medical databases: Impute, delete or classify?* Artificial intelligence in medicine, 2013. 58(1): p. 63-72.

[3] Swartz, N., *Gartner warns firms of ‘dirty data’*. Information Management Journal, 2007. 41(3): p. 6-7.

[4] Wang, R.Y., *A product perspective on total data quality management*. Communications of the ACM, 1998. 41(2): p. 58-66.

[5] Benson, P.R., *ISO 8000 data quality: The fundamentals, part 1*. Real-World Decision Support (RWDS) Journal, 2009. 3(4): p. 1-9.

[6] Ballou, D.P. and G.K. Tayi, *Enhancing data quality in data warehouse environments*. Communications of the ACM, 1999. 42(1): p. 73-74.

[7] Bohannon, P., et al. *Conditional functional dependencies for data cleaning*. in 2007 IEEE 23rd international conference on data engineering. 2007. IEEE.

[8] Wang, R.Y. and H.B. Kon. *Toward total data quality management (TDQM)*. in Information technology in action. 1993, Prentice-Hall, Inc.

[9] Redman, T.C., *The impact of poor data quality on the typical enterprise*. Communications of the ACM, 1998. 41(2): p. 79-83.

[10] Ertay, T., et al., *Quality function deployment implementation based on analytic network process with linguistic data: An application in automotive industry*. Journal of Intelligent & Fuzzy Systems, 2005. 16(3): p. 221-232.

[11] Yue-Hua, H., *Data Mining for Classroom Teaching Quality Based on Fuzzy Comprehensive Evaluation [J]*. Computer Science, 2008. 2.

[12] Hashem, I.A.T., et al., *The rise of “big data” on cloud computing: Review and open research issues*. Information systems, 2015. 47: p. 98-115.

[13] Sanchez, P.L., et al., *Quality index for quality assurance in software development*. 2010, Google Patents.

[14] Alexander, W.E., J. Cady, and J. Gonzalez-Garcia, *The IMF’s Data Dissemination Initiative After 10 Years*. 2008: International Monetary Fund.

[15] Lee, Y.W., et al., *AIMQ: a methodology for information quality assessment*. Information & management, 2002. 40(2): p. 133-146.

[16] Long, J. and C. Seko. *A New Method for Database Data Quality Evaluation at the Canadian Institute for Health Information (CIHI)*. in ICIQ. 2002.

[17] Eppler, M.J. and P. Muenzenmayer. *Measuring Information Quality in the Web Context: A Survey of State-of-the-Art Instruments and an Application Methodology*. in ICIQ. 2002. Citeseer.

[18] Su, Z. and Z. Jin, *A methodology for information quality assessment in the designing and manufacturing processes of mechanical products*, in Information Quality Management: Theory and Applications. 2007, IGI Global. p. 190-220.

[19] Scannapieco, M. *Data Quality in Cooperative Information Systems*. in Doctoral Poster at the 28th Very Large Databases Conference (VLDB 2002), Hong Kong. 2002.

[20] Batini, C., et al., *Methodologies for data quality assessment and improvement*. ACM computing surveys (CSUR), 2009. 41(3): p. 16.

[21] Batini, C., et al. *A comprehensive data quality methodology for web and structured data*. in 2006 1st International Conference on Digital Information Management. 2006. IEEE.