A Research Review of Issues Related to Data Governance

Guang Chen[1,a *],Xinda Li[1,b], Di Wang[1,c] and Lin Huang[2,d]

[1]State Grid Energy Research Institute Co. Ltd., Beijing 102209, China

[2]State Grid Sichuan Electric Power Co., Ltd., Chengdu 610041, China

[a]chenguang@sgeri.sgcc.com.cn, [b]lixinda@sgeri.sgcc.com.cn, [c]wangdi1@sgeri.sgcc.com.cn, [d]19798860@qq.com

[*]corresponding author

Keywords: Data Governance; Data Quality; Metadata

Abstract: Data governance is an emerging and developing discipline that is not uniformly defined by the industry. This paper reviews the problem of data governance, first reviews the origin and development process of data governance problems, and analyzes the current situation and latest progress of data governance research from foreign and domestic dimensions respectively, in order to provide help and reference to relevant researchers.

1. A historical Review of Data Governance Issues

Data governance is an emerging and developing discipline that is not uniformly defined by the industry. One definition holds that data governance refers to a process from the use of scattered data to the use of unified master data, from little or no organization and process governance to comprehensive data governance enterprise-wide, from trying to handle master data chaos to master data organized [1]. Another definition is that data governance is initiated and implemented by the Data Management Committee of Enterprise Senior Management, and is a series of policies and procedures on how to conduct the commercial application and technology management of data throughout the enterprise. Data governance is a set of continuous improvement management mechanism, which usually includes organizational structure, policy system, data standards, technical tools, operation process, assessment and supervision. The industry prefers the second definition, to build a data management system from management and technology, for unified storage and management of data.

Data quality is a key task in data governance and is directly related to the availability and understandability of the data. Data Quality Management Data Quality issues, the statistical field began in the late 1960 s, the management field in the early 1980 s and the computer field in the early 1990 s. In the 1990 s, the United States began data quality control, data quality analysis and other related theoretical research, starting with the social insurance number that corrected mistakes in the United States. With the continuous development of business and information technology, the systematic data quality research began to appear [2].

ISO9000 series standard is a widely accepted quality management standard, data quality standard was first defined in the ISO9000-9004 series, formulated by ISO Quality Management and Quality Assurance Technical Committee, which provides quality management guidelines from technology and management, at the core is the satisfaction of user requirements, establishment of functional responsibilities, assessment of potential risks and benefits, ISO9000 series provides a complete architecture for data quality research. ISO8000 data quality standard is a developing ISO standard, developed by ISO Technical Committee TC 184. ISO8000 data quality content involves data cleaning, data governance, data integrity research, data verification, data management, master data management, metadata management and data quality evaluation. ISO8000 clearly points out the concept of "data is product" and draws from mature product quality management system and theory to guide enterprises in data quality management [3]. At present, this method has begun to carry out

verification in many enterprises and achieved practical results.

According to the needs of data quality management practice, relevant international organizations have developed various international standards of data quality, among which the most representative are the research results of the International Monetary Fund (IMF) and Eurostat (Eurostat) [4]. In order to guide and regulate their member practices in data production, release and quality evaluation to ultimately improve the data quality of their member States, IMF developed a series of Data Quality Assessment Framework from 2001 to 2003 (DQAF), which finally completed the elaboration of data quality standards from five preconditions and quality (including ensuring integrity, method integrity, accuracy and reliability, accuracy and reliability, and availability). Eurostat established the Data Quality Assessment Working Group, which developed the definition of data quality and its six elements, namely, correlation, accuracy, timeliness and punctuality, accessibility and clarity, comparability, and consistency. Around the definition of data quality, Eurostat developed or funded the development of a series of data quality management tools, mainly including European Code of Statistical Practice, Standard Quality IndicSet, Data Quality Assessment Methods and Tool Manual, Quality Improvement Manual Based on Process Variable Analysis, Self-Assessment List of Survey Leaders, Quality Report Manual, Quality Reporting Standard, etc. These data quality management tools present a complete set of guidelines or approaches for user consultation and customer satisfaction surveys, self-assessment, quality indicators, quality reports, statistical review and quality certification, and quality improvement measures to guide data quality practices in EU Member States [5].

After nearly more than 20 years of development, many foreign research institutions or companies have achieved comparative data quality management research and systematic results. MIT Comprehensive Data Quality Management (TDQM) research proposed comprehensive data quality management methods, dividing data quality process into four stages: definition, measurement, analysis and improvement, proposed data product concept, compared data processing, storage and use with general industrial products, and thought that data quality control is similar to general product quality control process; proposed data product quality evaluation methodology, including data quality evaluation model, evaluation data collection method and evaluation methods. Trillium in the United States has proposed best practice methods to manage data status throughout the data lifecycle, implement data governance during data discovery, process development, and deployment management stages, and achieve faster results in multiple data domains. Trillium provides summary statistics and analysis information through data analysis and uses the results to establish automated processes to continuously evaluate data elements and sensitive information conditions in the production system [6]. For big data quality management, mainly in the traditional data quality practice, according to the characteristics of big data related technical research, for example, research more suitable for big data than traditional enterprise data, data quality research, IBM, Oracle and other companies have preliminary big data management solutions, mainly involves to deal with big data objects, data organization responsibilities, data quality strategy, data quality specific means, life cycle management, metadata management and other data management content.

2. The Latest Progress on Data Governance Research Abroad

(1) Data quality analysis

In the development of data quality and its evaluation methods, a large number of foreign statistical researchers have made significant contributions. T.Dalenius proposed in 1983 the concept of "measurement vectors" for the quality of statistical data including statistical accuracy, data detail, relevance, timeliness, timeliness, economics, confidentiality, etc. Groves conducted a detailed study on several aspects of the sources of error from the statistical error perspective. CurioBatini proposed systems that to compare different data quality evaluation methods in 2009. In 2000, John Cornish,Lee Dongmyeong et al. established an index system for data quality evaluation and improved the evaluation method. Kahn MG in 2012 proposed a conceptual model approach to the quality assessment of single-site and multi-site data, and applied it to medical data quality

evaluation, achieving useful results. PhilipWoodall has many different requirements for data quality evaluation and proposed a mixed data quality evaluation method. In this process, different scholars put forward their own data quality evaluation models, with the following data models: Aebi data quality index measurement description, Kon data quality measurement description, Motro measurement evaluation, Reddy data quality index measurement evaluation, Parssian data quality index measurement evaluation, Naumann integrity index measurement evaluation, Scannapieco integrity index evaluation, Even abstract measurement description, Ballou data quality measurement description [7].

After nearly more than 20 years of development, many foreign research institutions or companies have achieved comparative data quality management research and systematic results. MIT Comprehensive Data Quality Management (TDQM) research proposed comprehensive data quality management methods, dividing data quality process into four stages: definition, measurement, analysis and improvement, proposed data product concept, compared data processing, storage and use with general industrial products, and thought that data quality control is similar to general product quality control process; proposed data product quality evaluation methodology, including data quality evaluation model, evaluation data collection method and evaluation methods. Trillium in the United States has proposed best practice methods to manage data status throughout the data lifecycle, implement data governance during data discovery, process development, and deployment management stages, and achieve faster results in multiple data domains. Trillium provides summary statistics and analysis information through data analysis and uses the results to establish automated processes to continuously evaluate data elements and sensitive information conditions in the production system. SallieMae, the largest student loan company in the US, conducts business-driven data quality management practices designed by 11 parts, including data quality vision, quality strategy, data quality organization, data quality service and constraint models, metrics of data quality [8], data quality framework and methodology. In August 2014, the US Department of Defense, the American Defense Industry Association and the Software Engineering Research Center of Carnegie Mellon University officially launched the Data Management maturity Assessment Model (DMM). Data Quality Management is one of the six DMM management domains, used to guide enterprises to improve the data quality management ability and evaluate and improve the data quality management level of the organization. For big data quality management, mainly in the traditional data quality practice, according to the characteristics of big data related technical research, for example, research more suitable for big data than traditional enterprise data, data quality research, IBM, Oracle and other companies have preliminary big data management solutions, mainly involves to deal with big data objects, data organization responsibilities, data quality strategy, data quality specific means, life cycle management, metadata management and other data management content.

(2) Data governance system

At present, the main organizations of academia and industry at home and abroad include the international data management association, data management expert organization, etc. IBM is typical of data governance. Many years ago, there were many problems in data governance, no clear data sources and data owners, and data governance was low. With data governance, simplifying infrastructure and reducing management complexity, IBM in 2007 dropped from 128 in 1992 and data centers from 155 to 6. As management and data centers decrease, data management becomes more comprehensive and reasonable. In 2011, the Enterprise Data World Conference in Chicago proposed that data governance investments must be targeted at advancing business goals and raising the bottom line. The chairman of the Enterprise Data Management and Data Governance Association pointed out that the data governance plan is clear, but it is a long process to identify the cause of the problem and give quantitative analysis. The models in the field of data governance include the data governance maturity model proposed by IBM. The mainstream data governance methods mainly include data asset management, data quality management, main data management and data use case management. At present, the popular data management products in the industry include IBM Industry Models、Identity Manager Data Governance、Collibra Data Governance

Center、Cloudera Enterprise Data Hub and other products, which all provide metadata management and data quality management functions, and some products also provide data model management, data warehouse management and other characteristics [9].

Metadata is the basis of data management. There are two main ideas for metadata management: establish a metadata storage, provide metadata access and metadata life cycle management; establish a way of metadata exchange, through which metadata in different systems can access each other to integrate distributed and heterogeneous systems together to realize metadata management function. Management methods based on metadata warehousing require the definition and implementation of metadata standards and the modeling language for this standard, and current mainstream standards include MOF metadata warehousing structure of OMG and OIM's warehousing model. The metadata management strategy based on exchange channels needs to establish metadata exchange standards and provide metadata bridge. Communicate different tools or applications through the metadata bridge, so that they can access their respective metadata with each other to achieve the purpose of metadata integration. Currently, the mainstream standards are CDIF, XMI, etc. In recent years, with the increasing system scale, the traditional artificial annotation-based method cannot effectively solve the problem of metadata management of multi-source heterogeneous data in the background of big data. Therefore, the semi-automatic metadata data collection method has emerged. Semi-automatic metadata mobile phone method is divided into two categories: metadata extraction and metadata collection. Through the artificial strategy-assisted method, semi-automatic metadata data is found.

3. Recent Progress in Domestic Data Governance Research

In the research of data quality management and its evaluation methods, Chinese research institutions and scholars have made a lot of exploration. The National Engineering Center of Basic Software of Chinese Academy of Sciences compared and analyzed the two main aspects involved in data quality research, namely, data quality evaluation and data quality improvement technology, and introduced the representative data quality improvement tools, and finally put forward an evaluation-driven data quality improvement framework. Fang Youlin, Yang Dongqing and others described the data quality quantitative elements with quantitative methods, while Meng Wei put forward the definition of "quality factor", mainly using the calculation method to act the quantitative elements on the data warehouse objects, so as to form a complete set of quality evaluation system. Fang Yulin, Yang Dongqing and others also used qualitative methods to study some non-quantitative elements. Zhang Fang established a three-layer quality evaluation structure system to improve the government's fuzzy statistical comprehensive evaluation of data quality. Peking University in the group led by Professor Tang Shiwei with the six yuan method of the model to evaluate the data quality, including data set, rules, expectation factors, the model not only explains the calculation method and technology, but also has a unique innovation is to borrow quantitative indicators to complete the whole system or part of the data quality model evaluation. Chen Fenglan and Wang Xiuqin proposed to improve the quality management system of statistical data products, and improve the integrity and accuracy of statistical data. Liu Hong and Huang Yan established a combined model according to the changing characteristics of the time series data, and used the method of trend simulation to evaluate the accuracy of the GDP data in China. According to the reliability of data quality, Luanpo selected Chinese GDP data for data quality evaluation using steady MM estimation method. Xu Dilong targeted the methods of data quality evaluation, discussed six types of evaluation methods in detail, and learned the EU's evaluation theory of data quality and its practice, and suggested that China establish a data quality evaluation framework. Li Tinghui evaluated time series matching in China for the accuracy of GDP data. Meng Lian and Wang Xiaolu used the production functions to evaluate the Chinese GDP method for the data quality, and concluded that the GDP data was indeed distorted. In order to comprehensively improve the data quality, Huang Jianqi put forward the data quality control standards and data quality control technology for each stage of the statistical work, to evaluate the quality of the

obtained data. Zeng Wuyi proposed that statistical data should be accurate, timeliness, comparable, applicability and availability, and defined the quality connotation of statistical data to each generation link of statistical data [10].

The common problem in domestic data quality management research is the lack of scale organization and systematic research results, no relevant data management authority, the participation in international data quality standards is not enough, and the implementation and application of data quality standards is still in the initial stage. However, in recent years, index many industries in our country, data is increasing year by year, the number of source system, information system collected information increased year by year, and data integration environment is complex, make data difficult to maintain, difficult to guarantee quality, data quality management has received more and more attention, China's finance, telecommunications, energy and other industries in data quality management and data quality evaluation has done a lot of research and practical activities.

Acknowledgments

This work is supported by the science and technology project of State Grid Corporation of China: “Research on data governance and knowledge mining technology of power IOT based on Artificial Intelligence” (Grand No. 5700-202058184A-0-0-00)

References

[1] Liu Guifeng, Qian Jinlin, Lu Zhangping. Research progress in data governance at home and abroad: Connotation, elements, model and framework [J]. Book Intelligence Work, 2017, 61 (21): 137-144.

[2] Zheng Daqing, Huang Lihua, Zhang Chenghong, Zhang Shaohua. The concept of big data governance and its Reference architecture [J]. Research and Development Management, 2017, 29 (04): 65-72.

[3] Zhang Ning, Yuan Qinjian. Review on Data Governance Research [J]. Intelligence Journal, 2017, 36 (05): 129-134 + 163.

[4]Zhang Kangzhi. Data Governance: Direction of understanding and construction [J]. Egovernment, 2018, {4} (01): 2-13.

[5] tube Chi ili.Government data opening focuses on collaborative governance [N].Learning Times, 2021-07-16 (003).

[6] Kou Xiaofang, Huang Hua, Xu Yijun, Zu set sail.Online monitoring data management based on the big data algorithm [J].Environmental Ecology, 2021,3 (07): 93-98.

[7] Hu Zhengkun, Guo Feng.Global data governance: situational analysis and trend outlook [J].Information Security and Communication Security, 2021 (07): 11-18.

[8] Yang Rongjun.On the Value Objective, Right Ownership and the Legal Guarantee of Government Data Governance [J].Southeast Academic, 2021 (04): 113-124 + 247.

[9] Wang Weiling.Global data governance: realistic motivation, dual circumstances, and advance path [J].International Trade, 2021 (06): 73-80.

[10] Feng Yi.Improve the government data governance institutions to build a data governance pattern [J].China Information Industry, 2021 (03): 78-81.