2023 6th International Conference on Data Science and Information Technology (DSIT) | 979-8-3503-0444-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/DSIT60026.2023.00038

Multi-source Open Data Quality Evaluation Model in the Web 3.0 Era

Tianan Yan
Engineering Center
Institute of Scientific and Technical Information of China
Beijing, China

Zhaosen You
Schools of Software Engineering
University of Science and Technology of China
Suzhou, China

Yifan Zhang
Institute of Statistics
Xi’an University of Finance and Economics
Xi’an, China

Ruikai Hua*
JIANGNAN RURAL COMMERCIAL BANK
Changzhou, China

***Abstract*—In the Web 3.0 era, information sources were distributed in different locations, and data collectors collected data with issues such as inconsistent data formats and missing data fields, resulting in the inability to conduct data analysis and provide accurate decision-making services. According to the actual process of multi-source open data collection, the evaluation process is divided into three parts: multi-source monitoring quality, original open data quality and processing open data quality. According to the open data quality evaluation indexes set up at different stages, the weight value of each evaluation index is calculated according to the scoring situation, and the open data quality evaluation model oriented to the characteristics of multi-source is constructed. It is found that the multi-source monitoring quality and the open data itself quality should be ensured first, and then is the quality of open data after data processing, in order to better improve the overall multi-source open data quality and provide support for data analysis and intelligence services.**

Keywords-component; data quality; multi-source; open data; evaluation model; Web 3.0

I. INTRODUCTION

In the Web 3.0 era, information sources are all kinds of open data subjects, including government departments, scientific research institutes and other entities for information disclosure and data sharing, as well as individuals providing data [1].Multiple information sources are the combination of multiple information sources and data sources. Because related research and business work need to obtain a large number of open data distributed in different sources, and the diversity of sources will cause problems such as difficult data acquisition, low value of original data and difficult processing and integration of original data. Now, different scholars have carried out a lot of research on data quality evaluation for different data use scenarios and data processing links.However, there is no research on multi-source open quality data for the whole process cycle from source monitoring to data acquisition and processing. Therefore, when the data is expected to have high-value characteristics, it is necessary to evaluate the data quality of multiple sources, and implement corresponding measures in the whole process of multi-source open data from source quality to data quality to ensure the quality of multi-source open data.

II. MULTI-SOURCE OPEN DATA QUALITY

According to the existing understanding of the quality of multi-source open data, the essence of multi-source different from traditional single source is the diversity of its data sources, the complexity of different source open data and the heterogeneity of different source open data. Open data is the content of data disclosure by the data source according to relevant laws and regulations. The open data of multiple sources needs to be obtained legally. According to the perspective of the whole process of open data collection, combined with the five key links of "task planning, collection and acquisition, processing, development and dissemination" in tcped model [2], the quality of open data of multiple sources needs to be monitored, including the multi-source monitoring quality, original open data quality and the processing open data quality.

Multi-source monitoring quality. It mainly includes the determination of key monitoring objects, source authority and other self attributes, as well as the efficiency of source monitoring. The determination of key monitoring countries mainly covers the countries with the characteristics of collecting open data. Focus on monitoring the determination of relevant national open data departments. Including scientific research institutions, universities, companies, social organizations and other open data release subjects. Different sources need to be monitored, including the determination of monitoring objects, their own attributes and data monitoring process[3].

Original open data quality. There are different open data generation methods for open data from different sources. The open data in the source shows the characteristics of the source. The characteristics of the open data of the source include the integrity, relevance, availability, timeliness, security of the original open data and the legitimacy of data. Source disclosure data is usually the processed data released

979-8-3503-0444-2/23/$31.00 ©2023 IEEE
DOI 10.1109/DSIT60026.2023.00038

by the data disclosure subject, which can completely convey a certain meaning. There will be some connection between the relevant open data collected from multiple sources, or they are all data in some field. Through the quality evaluation of the original open data, we can timely understand the quality of the collected data and adjust the monitoring object and mode of the source in time[4].

Processing open data quality. Processing open data quality refers to the original open data quality, which is the data quality after data processing, data governance, data visualization and other data governance steps, including the value, accuracy, availability, intuitiveness, effectiveness and rationality of processing open data[5]. The quality of open data after processing should be related to the user experience. For example, the value of public processing data is determined by the user's use of the data. The intuitiveness and effectiveness of processing data are closely related to the user experience.

III. Analysis of Evaluation Index of Multi-Source Open Data Quality

The data quality of multi-source disclosure will be affected by the authority of the source itself, source screening methods, data processing, data collection, data analysis and prediction, and user needs. The difference in source selection, data collection methods, data analysis and prediction methods will affect the data quality. Next, the evaluation indicators of multi-source open data quality are analyzed from three dimensions:Multi-source monitoring quality, original open data quality and processing open data quality..

A. Multi-source monitoring quality

For multi-source monitoring, from the determination of the source monitoring object to the evaluation of the source's own authority and security stability, and then to the standardization of the source monitoring process and the source monitoring efficiency, through the evaluation of the whole process of source monitoring, we can comprehensively grasp the quality of multi-source monitoring. The quality of multi-source monitoring is mainly measured by five indicators: the rationality of the monitoring object, the authority of the monitoring source, the standardization of the source monitoring process, the security and stability of the source and the efficiency of the source monitoring.

Rationality of monitoring objects. The determination of monitoring objects includes the determination of monitoring countries and the determination of relevant open data departments of monitoring countries[6]. Monitor the reasonableness of national determinations. After determining the key monitoring countries through relevant papers and patents, it is necessary to carry out the rationality of the selection of key monitoring countries at the first time, evaluate whether the selected countries meet the data needs, and ensure the data quality from the source. Monitor the necessity determined by relevant open data departments of the state. The determination of the necessity of the open data Department determines the necessity of pre-collecting the open data Department, which can ensure that the data collected by the Department can meet the needs of users.

The authority of the source. The information source needs to have certain authority, which determines the influence of the information source. It is necessary to judge whether the information source has the qualification related to data disclosure and the ability related to data governance[5]. Before selecting the monitoring source, it is necessary to investigate the establishment background, data source, data mining tools and other data related to the original open data. For sources with low authority, it should be considered whether to keep them in the monitoring list.

The security and stability of the source. Source security and stability should be considered from two aspects: source security and source. Source security[6]. The monitoring information source needs to meet the security. In the process of checking the monitoring information source, the information source itself must have certain security measures, certain security protection ability and emergency remedial measures.The stability of the monitoring source determines the continuity of open data collection. During data acquisition, it is necessary to continuously and stably collect open data, which requires a stable source as the premise.

Standardization of monitoring process. The standardization of monitoring process refers to whether each source monitoring step is standardized in the source monitoring process [7]. The standardization of the process will affect the standardization of collecting the open data of the source. A standardized source monitoring process should be established in advance and standardized monitoring steps should be formulated in the monitoring process.

Efficiency of source monitoring. The efficiency of source monitoring is the determination of monitoring and the efficiency evaluation of continuous monitoring process[8]. The evaluation of monitoring efficiency can ensure that the monitoring of key sources can be improved in the source monitoring stage. It is also evaluated by monitoring the open data acquisition difficulty of information sources. Before monitoring, it is necessary to evaluate the difficulty of open data acquisition of the source, and timely adjust the monitoring personnel, data acquisition personnel and relevant equipment of the source.

B. Original open data quality

The original open data refers to the unprocessed data directly collected from the source. Through the evaluation results of the quality of the original open data, we can judge the quality of the source data disclosure and provide data optimization strategies to facilitate the subsequent processing of the original open data. The evaluation of the quality of the original open data is mainly carried out from the five perspectives of the internal characteristics and use level of the original open data, namely the integrity, relevance, availability, timeliness and legitimacy of the data.

Integrity of original open data. The original open data shall include all the open data disclosed by the source, including the main open data contents such as data disclosure title, time, disclosure department and content[4]. Ensure the overall data collection during data collection. At the same time, monitor the lack of open data fields and contents in the process of collecting the original open data, and judge the integrity of the original open data through statistical analysis.

The relevance of the original open data. Data correlation is the relationship between different data[10,18].The relevance of the original open data is the relationship between the original open data in multiple sources. Because the open data of multiple sources have different data types and data structures, it is necessary to obtain the open data for a specific demand. After obtaining the original open data, we need to pay attention to the relevance of the open data to avoid hidden dangers for subsequent data analysis.

Availability of original open data. The availability of data indicates the availability of data, which can directly affect the value of data[5]. The availability of the original disclosure includes not only whether the user can use the original open data directly after it is obtained, but also whether the original open data can be processed again and the complexity of processing.

Timeliness of original open data. Data has a life cycle, and the value of data changes with time[8]. Due to the timeliness of the data, we also need to pay attention to the timeliness of the original open data after collection. The timeliness of the original open data is mainly reflected in the lack of data value in the analysis and prediction results of the original open data beyond a certain time range, that is, the prediction results are inconsistent with the facts. Legitimacy of data. Data use needs to comply with. The legitimacy of the use of the original open data needs to be in accordance with the regulations issued by the state in advance[9]. The source has set up some agreements for the open data to avoid disputes in the subsequent use of the original open data.

C. *Processing open data quality*

In the original open data, the data level fusion and field level fusion of multi-source open data are carried out through different data processing methods. In order to ensure that the data value is not reduced, the meaning of data expression is more accurate and the effectiveness of data processing is better, the quality of processed open data needs to be evaluated. Specifically, it is evaluated from five perspectives: the value of processing open data, the accuracy of processing open data, the intuition of processing open data, the effectiveness of data processing and the rationality of processing methods.

Processing open data has value. Data value evaluation needs to investigate users' use of processing open data, and evaluate the value of processing open data through users' actual experience, so as to provide support for the adjustment and optimization of open data processing methods[9].

Accuracy of processing open data. Compared with other data, the accuracy of data directly affects whether the data quality can meet the requirements of data analysis. For the evaluation of the accuracy of processing open data, we can judge whether the data is more accurate and can meet the needs of users compared with the original open data[5,6].

Intuitiveness of processing open data. The intuitive performance of data directly affects the value of data, thus affecting the quality of data. It can be evaluated by whether the processing open data can visually display the key data and key fields, as well as the visualization degree of the processing open data. Relevant statistics and analysis technologies can be directly used to help users directly understand the content of the processed open data[4].

Effectiveness of data processing. Whether the data processing is effective or not needs to be compared according to the original data and processing data, and the proportion of processing open data and original open data should be counted[5,10]. If the proportion of processing open data is low, it means that the loss degree in the process of data processing is too high and the effectiveness of data processing is low.

Rationality of data processing methods. The processing method determines the quality of the processing open data. According to the characteristics of the original open data and the specific application scenarios of the data processing method and model, we should choose to use the data statistics method and data mining method to convert the original open data into the data that can meet the use needs.

IV. Construction of Multi-Source Open Data Quality Evaluation Model

Based on the analysis of multi-source open data quality evaluation indicators, this paper constructs a multi-source open data quality evaluation model from three dimensions: Multi-source monitoring quality, original open data quality and processing open data quality. The whole process of multi-source open data acquisition quality cannot obtain a unified way. It is necessary to consider the specific indicators of three dimensions at the same time to find the problems existing in the existing data quality.

TABLE I. Multi-source open data quality evaluation index score situation

Variable name	Max	Min	Avg	Median
Rationality of monitoring objects	7	5	6.2	6
The authority of the source	7	4	6	6
The security and stability of the source	7	3	5.77	6
Standardization of monitoring process	7	3	5.53	6
Efficiency of source monitoring	7	3	4.97	5
Integrity of original open data	7	3	5.1	5
The relevance of the original open data	7	4	5.8	5.5
Availability of original open data	7	4	5.23	6
Timeliness of original open data	6	3	4.67	5
Legitimacy of data	7	4	5.8	6
Processing open data has value	7	3	5.4	5
Accuracy of processing open data	7	5	6.1	6
Intuitiveness of processing open data	7	4	4.5	4
Effectiveness of data processing	7	3	5.13	5
Rationality of data processing methods	7	4	5.6	5

By using the questionnaire survey method to evaluate the importance of various indicators in the three dimensions, a questionnaire was distributed to 60 experts and scholars who carried out multi-source monitoring research and open data collection, and the scores of the importance of multi-source open data quality evaluation indicators were obtained, as shown in Table 1.

According to the data in Table 1, among the five evaluation indexes of multi-source monitoring quality, the average score of employees is between 5.4 and 6.2, which belongs to a important range.The importance of multi-source monitoring quality evaluation indexes is as follows: the rationality of monitoring object determination (6.2), the authority of information source (6), the safety and stability of information source (5.77), the standardization of information source monitoring process (5.53) Efficiency of source monitoring (4.97). It reflects that the quality of multi-source monitoring is highly important in the determination and ideal of monitoring objects and the authority of monitoring sources. It is highly important in the security and stability of sources and the standardization of source monitoring process, but slightly less important in the efficiency of source monitoring. From the median score, the efficiency of source monitoring (5) is less than other test items in the quality of multi-source monitoring. It shows that the decisive factor lies in the determination and authority of the monitoring object. It is necessary to ensure that the data is safe and stable, and set up a standardized monitoring process. The monitoring efficiency will affect the cost , and has a low impact on the quality of source monitoring.

Among the five indicators of the quality of the original open data, the mean value of importance is 4.67 ~ 5.8, which belongs to a very important range. The evaluation indicators of the quality of multi-source monitoring are in order of importance: the relevance of the original open data (5.8), the legitimacy of the use of the original open data (5.8), the availability of the original open data (5.23), the integrity of the original open data (5.1) Timeliness of original open data (4.67). It can be found that the relevance and legitimacy are of high importance to the data quality and the timeliness of the original open data is of low importance. From the median score, the median scores of usability (6) and legitimacy (6) are higher, and the median scores of integrity (5) and timeliness (5) are lower. It can be seen that most practitioners believe that the availability and legitimacy of the original open data should be taken into account.

Among the five indexes of processing open data quality, the mean value of importance is 4.5 ~ 6.1, and the range of importance is large. The evaluation indexes of processing open data quality are in order of importance: accuracy of processing open data (6.1), rationality of data processing methods (5.6), value of processing open data (5.4), effectiveness of data processing (5.13) and intuitiveness of processing open data (4.5). It can be concluded that data accuracy is very important for the quality of processing open data, the rationality of data processing methods and the importance of the value of processing open data are high, and the importance of the intuitiveness of processing open data is low. From the median score of employees, it is found that the accuracy of processing open data (6) is still an important factor affecting the data quality. The value of processing open data (5), the rationality of data processing methods (5) and effectiveness (5) are the same, while the intuitiveness of processing open data (4) is the same as the mean, and the median score is not different from the mean, It reflects that most practitioners believe that the most important factor affecting the quality of processing open data is accuracy.

Through the analysis of the overall importance, the mean value scores of 15 evaluation indicators are processed by deviation standardization to obtain the index weights of relative importance, which are as follows: the weight values of each index of source monitoring quality are the rationality of the monitoring object (0.119), the authority of the source (0.105), the security and stability of the source (0.089), the standardization of the source monitoring process (0.072) and the efficiency of source monitoring (0.033); The weight values of each index of the quality of the original open data are the integrity of the original open data (0.042), the relevance of the original open data (0.091), the availability of the original open data (0.051), the timeliness of the original open data (0.012) and the legitimacy of the use of the original open data (0.091); The weight values of each index of processing open data quality are the value of processing open data (0.063), the accuracy of processing open data (0.112), the intuitiveness of processing open data (0), the effectiveness of data processing (0.044) and the rationality of data processing method (0.077). Due to the intuitiveness of the processing open data, the normalized importance value is 0, which can be deleted from the evaluation model, and finally the multi-source open data quality evaluation model is obtained, as shown in Figure 1.

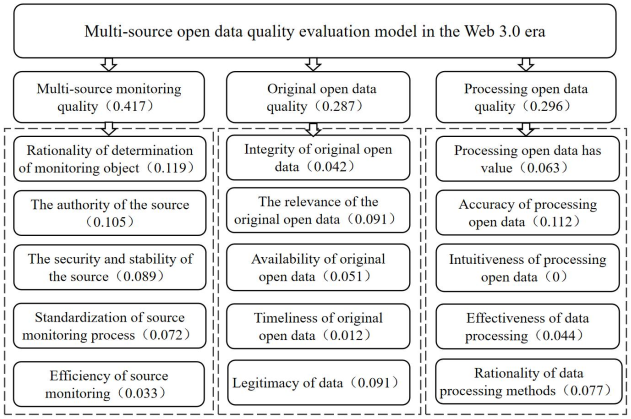
Multi-source open data quality evaluation model in the Web 3.0 era
Multi-source monitoring quality（0.417）
Original open data quality（0.287）
Processing open data quality（0.296）
Rationality of determination of monitoring object（0.119）
The authority of the source（0.105）
The security and stability of the source（0.089）
Standardization of source monitoring process（0.072）
Efficiency of source monitoring（0.033）
Integrity of original open data（0.042）
The relevance of the original open data（0.091）
Availability of original open data（0.051）
Timeliness of original open data（0.012）
Legitimacy of data（0.091）
Processing open data has value（0.063）
Accuracy of processing open data（0.112）
Intuitiveness of processing open data（0）
Effectiveness of data processing（0.044）
Rationality of data processing methods（0.077）

Figure 1. Multi-source open data quality evaluation model

By calculating the weight values of the three-dimensional dispersion standardization, they are multi-source monitoring quality (0.417), original open data quality (0.287) and processing open data quality (0.296). It can be found that the weight value of multi-source monitoring quality is significantly higher than the original open data quality and processing open data quality, indicating that to improve the quality of multi-source open data monitoring, the focus is to improve the quality of multi-source monitoring. Of course, the quality of open data is the foundation of everything, and improving the quality of open data itself should be as important as the quality of monitoring.At the same time, it also relies on the joint action of processing open data quality to improve the overall quality of multi-source open data.

V. Conclusion

This article obtains multi-source open data quality evaluation model based on the calculation of the importance of evaluation dimensions. After obtaining the quantified weights of various evaluation indicators, the overall quality of multi source public data can be quantified by evaluating the entire process of multi-source public data. The multi-source public data quality assessment model provides a theoretical framework for the quality assessment of multi-source monitoring, public data collection and processing in the Web 3.0 era, and also provides experience for future research on data analysis and intelligence services.

REFERENCES

[1] Christensen E W , Bailey J R. 1997. A Source Accessibility Effect on Media Selection. Management Communication Quarterly An International Journal.https://doi.org/10.1177/0893318997010003005

[2] Nissen M E, Gallup S P. 2012. Knowledge to the tactical edge: Enhancing the TCPED process through new metrics to inform ConOps development and IT system acquisition. Monterey, CA: Naval Postgraduate School.

[3] Jo, HW., Kim, SW. 2011. A Service Quality Model for the Public Information Service. In: , et al. U- and E-Service, Science and Technology. UNESST 2011. Communications in Computer and Information Science, vol 264. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-27210-3_44

[4] Wei Yanhua. 2017. Research on the Authoritative Calculation Method of Weibo Posts for Information Retrieval. Central China Normal University.https://cdmd.cnki.com.cn/Article/CDMD-10511-1017264750.htm

[5] Zeng Yan, Zhang Jianyong. 2009. Data Processing Specification for Literature Database - Chapter 1 Compilation Principles and Methods. Specification for Data Processing in Literature Database,.http://ir.las.ac.cn/handle/12502/4457

[6] Li Ziyang, et al. 2018. Reasonability testing of dam monitoring data based on statistical diagnosis. Progress in Water Resources and Hydropower Technology.https://doi.org/10.3880/j.issn.1006-7647.2018.05.013

[7] Nyberg A, Palmgren S.2011. Using Indicators for Technology Monitoring. Steps toward a proposed framework.

[8] Zhang Xiang. 2020.The Transformation Path of Cross border Enterprise Data Compliance Governance on the TikTok Incident.China Information Security.

[9] Liu Wenjun, et al. 2022. A Quality Evaluation System for Monitoring Data Based on Anomaly Detection Integrated Algorithm. Grid and Clean Energy .http://qikan.cqvip.com/Qikan/Article/Detail?id=7107890204

[10] Zhang Tiewei, et al. 2022. Governance Path for Financial Data Compliance Driven by Privacy Technology. Business Research.http://qikan.cqvip.com/Qikan/Article/Detail?id=7108892728