J Inf Process Syst, Vol.21, No.1, pp.13~27, February 2025
https://doi.org/10.3745/JIPS.04.0332

ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

DQA4AirQuality: TDQM-Based Data Quality Assessment Framework for Air Quality Datasets

Lina Zhang[1] and Sukhoon Lee[2,*]

Abstract

In recent years, open government data and big data analytic applications have become increasingly widespread. Without proper quality control, the rapid dissemination of data may jeopardize the reuse of datasets and exert negative effects. The current general frameworks for data quality management in literature are outdated and lack extensions to big and open data. In this work, a four-level data quality assessment dimension generation model was developed and applied for air quality datasets to measure the quality of air quality data from various data quality dimensions. This assessment framework was validated by comparing it with four air quality datasets from the World Health Organization (WHO), Beijing, Seoul, and Italy. The results show that the datasets published by the WHO have low quality due to their more complex sources.

Keywords

Air Quality, Big Data, Data Quality Assessment, Open Data

1. Introduction

Air pollution is the poisoning of the air people breathe, which may constitute a health hazard to human beings and ecosystems. Air quality index (AQI) is commonly used to quantitatively describe air quality conditions. Major air pollutants include particulate matter (PM), ozone O_3, nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and carbon monoxide (CO). $PM_{2.5}$ and PM_{10} can penetrate deep into the lungs, and $PM_{2.5}$ can even enter the bloodstream, which causes cardiovascular and respiratory diseases in severe cases [1]. From the characteristics of data on Internet of Things (IoT), the collection of air quality data contains multiple devices, including sensors of different environments, weather conditions, brands and types. Therefore, this causes the problem of data integration from heterogeneous sources. Besides, data quality assessment (DQA) for heterogeneous environmental integration is important and challenging. In this context, data quality (DQ) becomes a prerequisite to ensure the value of air quality datasets and the accuracy of AQI values. In particular, it is increasingly necessary to assess DQ via consensus-driven methods, and data quality management (DQM) plays an increasingly important role when DQ is used with open research data [2].

In the past two decades, various DQ methodologies have been developed in different papers. They include total data quality management (TDQM), total quality information management (TIQM), data

Manuscript received April 8, 2024; first revision July 16, 2024; accepted September 25, 2024.

* **Corresponding Author:** Sukhoon Lee (leha82@kunsan.ac.kr)

[1] School of Computer Science, Baoji University of Arts and Sciences, Baoji, China (zhanglina@bjwlxy.edu.cn)

[2] Department of Software Science & Engineering, Kunsan National University, Gunsan, South Korea (leha82@kunsan.ac.kr)

warehouse quality methodology, assessment of information quality methodology (AIMQ), heterogeneous information management quality assessment, and heterogeneous data quality management. International standards organizations also improve the efficiency of DQM among countries and organizations by developing unified DQ standards including ISO 8000 to facilitate data storage, delivery and sharing, and reduce decision errors caused by various DQ problems. Despite the gradual maturity of DQ research, different datasets should have diverse assessment dimensions with the development of big data technology, the growth of open datasets, and the widening application areas of IoT technology. In addition, there are few quality assessment studies related to air quality datasets.

By reviewing the existing literature on DQ, this study proposed a TDQM-based data quality assessment framework that can be applied to the quality assessment of air quality datasets. The advanced DQA framework focuses on refining and improving the definition and measurement phases of TDQM, which provides detailed guidance on its operation. Additionally, air quality datasets published by four different agencies were selected and evaluated for DQ in the present study.

2. Related Works

2.1 Development of DQM Frameworks

In the study [3], consideration was given to the DQM methodology collection of criteria and technology that starts from input messages which describes the specific application context and defines a reasonable procedure for evaluating and improving DQ. In the survey, DQM techniques applicable to IoT, including methodologies and frameworks, were discussed [4]. The frameworks mentioned in this study include methodologies, frameworks, methods and standards.

2.2 Comparative Perspectives

After a previous survey [4], four common generic methodologies were selected for comparison. TDQM has the advantages of comprehensiveness and generality [5], whereas the disadvantages are as follows. The DQ dimensions defined in the assessment phase are fixed and non-extendable; the proposed DQ dimensions are relatively old and cannot cover new data characteristics including big and open data, and accurately assess open and big datasets; and industry-specific management and assessment techniques are not mentioned [6-16].

Originally designed for use in data warehouses, TIQM possesses the advantage of having a detailed guidance process, which becomes one of the general-purpose DQMs at one time. Moreover, it also provides the benefit of the ability to estimate the cost issues corresponding to different data qualities from the perspective of managers in the management process. Notwithstanding its applicability to the context and data types of this study, TIQM was not used because the focus was not on the discussion of costs [6,17].

The main characteristic of AIMQ is the presence of two questionnaires in the assessment process [18]. To be specific, the former is to determine related DQ dimensions and attributes to be used as a benchmark. The latter is to obtain an information quality (IQ) measure. The advantage of AIMQ is that it is domain-independent and can be used for any type of data. Besides, the disadvantage of AIMQ is that its main management activities focus on assessment activities without providing guidelines, techniques and tools for improvement activities [6,19,20].

More IoT DQ studies based on the ISO 8000 standard have also emerged in recent years. The advantage of this standard is to allow for the international standardization of DQ in both structured and unstructured data, which provides standard guidelines for DQM and allows users to adapt the details to different domains. The disadvantage of this standard is that it is only an extension of ISO 9000 rather than a methodology, and proposes a process reference model which is more descriptive than operational and less actionable [21-24].

3. Flexible DQ Dimension Generation Model

3.1 DQA Dimension Generation Model

DQ is frequently characterized by multiple dimensions, each of which addresses a specific aspect of the overall quality of data [25]. In the current study, no universal DQ metric system can be applied to all but the most important few dimensions of datasets due to different industry domains, data types and application purposes. In this study, a flexible four-level dimension generation model was established, and a metric system was defined for corresponding dimensions. This model can be applied to specific data types and industry contexts.

This four-level model, which generates a set of DQ dimensions based on the type and domain of the dataset, is introduced in Fig. 1. The first level is to obtain generic DQ dimensions based on generic data characteristics. The second level determines whether the dimensions representing data characteristics are added to the set of DQ dimensions by determining the applicability of the dataset to big data analysis. Similarly, the third level determines whether open dataset characteristics can also be added to the set. The fourth level is the DQ dimensions corresponding to specific domain characteristics. In general, different application domains have more or less specific properties and a specific bias towards DQ, and this level is to add such dimensions to the set.

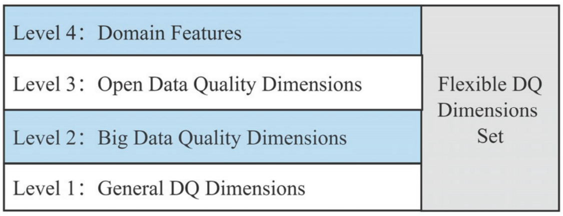
Level 4：Domain Features
Level 3：Open Data Quality Dimensions
Level 2：Big Data Quality Dimensions
Level 1：General DQ Dimensions
Flexible DQ Dimensions Set

Fig. 1. Dimension generation model.

3.2 General Dimensions

Some differences exist in the DQ dimensions used for an assortment of datasets depending on the data type and environment of the application. Based on previous research, however, some basic and important DQ dimensions can be employed for almost all DQAs, and here were referred to as the first level of generic DQ dimensions.

In [26] and [27], the authors concluded that six key dimensions of quality are essential and can be applied to most DQAs. ISO/IEC 25012 [28] defines DQ dimensions as data inherent and system dependent characteristics. The former refers to the characteristics that all data should have, and inherent characteristics include accuracy, completeness, consistency, trustworthiness and timeliness.

In 2013, DAMA UK issued a white paper that outlined six core dimensions of DQ [29]. Besides, a description was given to other characteristics that affect quality, such as usability, timing issues, flexibility, confidence and value. By summary, comparison and analysis, scores of studies focus on a fundamental suite of generic DQ dimensions. The model proposed in this study is consistent with DAMA UK, which sets the six properties of data as a common dataset dimension. The specific description of each dimension is shown in Table 1.

Table 1. General DQ dimensions

Dimension	Definition
Accuracy	The level of the data can correctly depict a "real-world" entity or activity.
Completeness	The percentage of stored data volume versus potential data volume.
Consistency	The differences between multiple representations and definitions of things are compared.
Uniqueness	No multiple instances of an entity should be recorded on the basis of satisfying object identification.
Timeliness	The level of performance of a data property is new or old for a given environment.
Validity	Data is considered as effective if it conforms to the grammar of its definition.

3.3 Dimensions for Big Data

What people commonly refer to as "big data" is broad and involves data, a complete conceptual and technical stack, as well as ways of data storage and management. Some specific data are collected according to the type of analysis and are organized in a particular way to handle different techniques [30]. This means that poor DQ can cause inaccurate results of big data analysis, which can bring about prediction or decision errors. Big data presents novel technologies and features that enable the use of DQM principles somewhat distinct from the application of traditional data.

Table 2. The quality characteristics of big data

Dimension	Definition
Accessibility	The difficulty of accessing data by users.
Authorization	Whether the person or group has permission to access the data.
Definition/Documentation	Data specification.
Credibility	It consists of both objective and subjective components for assessing data beyond numbers.
Metadata	Whether metadata describing different aspects of the dataset is provided.
Accuracy	It is determined by the application.
Integrity	The accuracy and consistency of data throughout its lifecycle are maintained and ensured.
Auditability	The stage of data use where the auditor fairly assesses the data.
Fitness	The extent to which the data matches the needs of the user.
Readability	The ability to interpret the content of the data.
Structure	Difficulty in converting semi-structured or unstructured data into structured data.
Validity	Data is considered effective if it conforms to the grammar of its definition.

In 2011, the three main dimensions of big data were proposed at the McKinney Global Institute [31], namely volume, variety and velocity, known as 3Vs. Merino et al. [32] proposed the "3As model of DQ in use," which includes three DQ characteristics—contextual, operational, and temporal adequacy—to evaluate the level of DQ for usage in big data projects, respectively. Cai and Zhu [33] developed a hierarchical DQ framework from the perspective of data users, including five big DQ dimensions, 14 quality characteristics and corresponding quality metrics. Due to more consistency with the one proposed in this study, the latter framework was integrated into the second layer of the dimensional generation model to obtain the quality level of incoming data usage in big data analysis. After excluding those that exist in general DQ dimensions, the remaining dimensions related to big DQ are shown in Table 2.

3.4 Dimensions for Open Data

Open data refers to the data that can be used, modified and shared by everyone. During the past few years, the dissemination of open government data has been very fast. The data has evolved from simple data analysis by data collectors based on purposes to current releases based on open data. Disclosure of data in the absence of adequate quality assurance exerts an adverse impact on users who use the data.

Vetro et al. [34] developed a framework to use a set of DQ dimensions to weigh the quality of open public data and measured and obtained quantitative DQ levels. The proposed framework includes seven dimensions, which are traceability, currentness, expiration, completeness, compliance, accuracy and understandability. After excluding those that exist in general DQ dimensions, the remaining dimensions related to open DQ are shown in Table 3.

Table 3. Dimensions for open data

Dimension	Definition
Traceability	Presence or absence of metadata related to the establishment and renewal of the dataset.
Understandability	Whether the columns or descriptive metadata are expressed in a format that is user-friendly.
Expiration	The proportion of the latency of a dataset's release beyond the expiry of its last edition to the time segment covered by the dataset.
Compliance	A five-star model is used to indicate the level of the dataset.

3.5 Data Domain Features

Similar to the definition of DQ, quality assessment has different application contexts and requirements, accompanied by focused dimensions. Lots of domains have established corresponding DQ standards, and some countries have also developed corresponding DQ requirements for specific domains.

In the domain of healthcare, the Canadian Institute for Health Information has defined a DQ framework to assess healthcare DQ. The framework model assesses DQ through three levels of indicators based on five dimensions, namely accuracy, timeliness, comparability, usability and relevance. The Institute of Hospital Management of the National Health and Wellness Commission of China published "Specific Requirements for DQ Assessment of Electronic Medical Records for Graded Evaluation-Revised Version 2022." The institute mentions that the DQA of electronic medical records mainly considers consistency, completeness, integration and timeliness. It also has characteristics including privacy and complex data types in addition to the characteristics of traditional data [35]. The use of global positioning system (GPS) data is more focused on validity, accuracy, real-timeliness and stability. Although they are all IoT systems, different application contexts require different DQ because of different environments and needs, which causes differences in dimensions. For example, medical systems may require higher device accuracy, while air quality monitoring for wide-scale deployment is sufficient by using common sensors. The six most commonly used DQ dimensions were summarized in IoT systems and analyzed in Section 3.4.

4. A Framework Based on TDQM

A DQM framework was proposed based on the classical TDQM methodology, and functional modules were added to evaluate big and open datasets. This framework inherits the four phases of TDQM, with a focus on improving and refining the definition and measurement phases. In the last level of the four-layer dimensional model, the data characteristics of the air quality domain are defined to be applied to the DQA of air quality datasets, which is called DQA4AirQuality. The main stages of the framework are summarized in Fig. 2. The four phases are definition, measurement, analysis, and improvement, which are repeated until the DQA results meet the requirements. This study refined the first stage by refining the four steps. They can identify the most common dimensions of DQA, determining whether it is big data or open data and special assessment dimensions that rely on domain characteristics.

The definition phase mainly uses the previously proposed flexible four-level model for generating DQ dimensions in four steps. The definition phase focuses on generating DQ dimensions in four steps by using the previously proposed flexible four-level model. Firstly, an inheritance assessment of generic data dimensions is performed. Secondly, whether the dataset is used for big data technologies is determined, and if so, big data related to measurement dimensions is added. Thirdly, whether the dataset is an open dataset is determined, and if so, measurement dimensions are added to evaluate the open dataset. Finally, the domain to which the dataset belongs and whether it contains some additional measurement dimensions is determined.

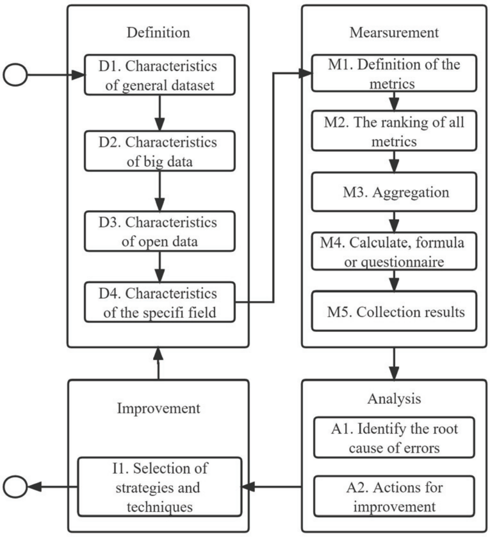
Definition
D1. Characteristics of general dataset
D2. Characteristics of big data
D3. Characteristics of open data
D4. Characteristics of the specifi field
Mearsurement
M1. Definition of the metrics
M2. The ranking of all metrics
M3. Aggregation
M4. Calculate, formula or questionnaire
M5. Collection results
Improvement
I1. Selection of strategies and techniques
Analysis
A1. Identify the root cause of errors
A2. Actions for improvement

Fig. 2. Lifecycle of DQA4AirQuality.

At the core of the four phases of DQM, measurement is the scientific assessment of DQ to determine whether it meets the needs of users or projects. A multifaceted assessment is performed by using the set of dimensions generated in the definition phase to measure the data quality. The DQA process can be classified into subjective and objective assessments. Specifically, subjective assessments are based on evaluations and surveys of users or experts. Objective assessments measure DQ attributes via formulas or calculations. Different users may define different assessment metrics but usually obtain quantitative results.

DQA results refer to the scores based on the quality performance indicators obtained during the measurement phase, with lower scores indicating more DQ problems and vice versa. In the poor-quality performance of the assessment results on a certain DQ dimension, it is indicated that some DQ problems exist in a certain aspect and DQ can be analyzed to investigate the root cause of DQ problems. In general, the heterogeneity of the IoT architecture increases as the complexity of the collected data increases, which causes a decrease in DQ. The data management capabilities of IoT platforms may face difficulties, resulting in an increased risk of poor DQ. This has numerous possible causes, as summarized in previous studies. The analysis phase concentrates on finding the root cause of DQ problems based on the assessment scores, which can provide a basis for the next step of quality improvement operations.

During the improvement phase, solution strategies and techniques are proposed to improve the degree of DQ. Nevertheless, not much attention is currently paid to the detailed exploration of improvement techniques and their practical applications in DQA. The TDQM methodology emphasizes the need to identify key areas for improvement. Various data sources often cause inconsistencies between interrelated fields. If DQ problems are caused by more than one data source, then the focus should be on checking whether these data sources are linked and investigating different data sources that provide the same data. DQ problems are also caused by manual operations, which require the investigation of data operators. The solution is to eliminate them by minimizing manual resolution and increasing automated operations.

5. A Case Study

The proposed framework focuses only on the DQM measurement phase. Three parts are included in the approach to a customized quality measurement framework for air quality datasets.

- Identify the most appropriate DQ model as the theoretical support for the measurement framework (Section 5.2)
- Select the measurement criteria (Section 5.3)
- Obtain the evaluation results (Section 5.4)

5.1 Datasets Analyzed

To investigate the quality of air quality datasets, four publicly available air quality datasets from WHO [1], Beijing [36], Seoul [37], and Italy [38] were selected. The information on these datasets is shown in Table 4. The four datasets have a greater variability—from global, Beijing, Seoul and an Italian town—to verify the generalizability of the model. The datasets come from different collection organizations, and may have different collection methods and techniques, different seasons, and different languages. Assessment and scoring of the quality of these diverse datasets can identify problems in the data monitoring process, and differences in the air quality monitoring stations worldwide, which can raise public awareness of the importance of openly available and high-quality data.

Table 4. Air quality datasets

Dataset	Institution	Years	Instances	Attributes	Time interval
A	World Health Organization	2000–2021	32,191	14	Annual average concentration
B	Beijing Municipal Environmental Monitoring Center	2013–2017	420,768	18	Every hour
C	Seoul Metropolitan Government	2017–2019	866,459	8	Every hour
D	ENEA	2004–2005	9,358	13	Every hour

Dataset A, the fifth air quality database published by the WHO, covers more than 6,000 cities in 117 countries. The database has been updated regularly every two to three years since 2011. The data include the annual average concentrations of PM, which are based either upon day-to-day observations or data that allow for summation into annual averages.

Dataset B includes air pollutant measurements per hour for 12 nationally coordinated air quality monitoring locations. The air quality data was offered by the Beijing Municipal Environmental Monitoring Center. The weather data from every air quality monitoring point was aligned with the closest meteorological station of the China Meteorological Administration. The timeframe is from March 1, 2013 to February 28, 2017. The missing data are indicated by NA.

Dataset C refers to the information on air pollution measurements for 25 districts in Seoul by the Seoul Institute of Health and Environment Air Pollution Measurement Seoul. The average measurement results for an hour are provided every five minutes, which offers the mean values of six contaminants (SO_2, NO_2, CO, O_3, PM_{10}, and $PM_{2.5}$).

Dataset D uses the public dataset of the machine learning repository of University of California, Irvine (UCI). The data was sampled from a heavily polluted area in Italy, with sensor arrays 9,358 installed next to roads. The data was recorded from March 2004 to January 2005. In particular, CO, non-methane nitrogen hydroxide, benzene, total nitrogen oxide and nitrogen dioxide concentrations were sampled hourly. Missing values occurred during data collection due to network, weather, disaster and sensor failure, and were marked by the dataset as -200.

5.2 DQA4AirQuality

The framework proposed in Section 4 is the most appropriate DQM for air quality datasets because of its generality and flexibility. The DQ dimension set used for air quality datasets first inherits the dimensions of the generic dataset as the smallest subset of the DQ dimension set. In the second step, whether this dataset is to be used for big data analysis is determined. Since this air quality dataset does not fit big data characteristics, it is unnecessary to merge the characteristics of big data. In the next step, an open dataset is uploaded to the UCI Machine Learning Repository, and it is necessary to inherit the characteristics of open data and merge corresponding DQ dimensions. In the last step, data volume and accessibility are selected from the characteristics of IoT DQ assessments. In air quality studies, AQI is usually used to represent the pollution level of air, which includes the quantity of six air pollutants such as ozone, nitrogen oxides, PM_{10}, $PM_{2.5}$, CO, and sulfur dioxide. Since some government $PM_{2.5}$ websites only provide the number of certain air pollutants, data volume is employed as one of the dimensions, which represents the number of records or attributes of this dataset. The more complete the dataset is, the more completely it can be used for research on air quality. Air quality data are mostly released by government departments. Many departments sometimes provide real-time AQI queries and also some governments encrypt data. The entrance to downloading the dataset is relatively difficult for ordinary users. Therefore, accessibility is also taken as one of the domain characteristics.

Table 5. DQA4AirQuality framework

Model level (1–4)	Dimensions	Subjective/Objective
General dimension	Accuracy	Subjective and objective
	Completeness/Consistency/Timeliness/Uniqueness/Validity	Objective
Dimensions for big data	None	None
Dimensions for open data	Traceability	Subjective
	Understandability/Expiration	Objective
	Compliance	Subjective
Dimensions for air quality	Data volume/Accessibility	Subjective

Accordingly, a flexible DQ dimension generation model is employed to obtain a DQ assessment framework that can be used to assess air quality datasets based on their characteristics. The framework is called DQA4AirQuality with 12 dimensions as described in Table 5. It is mainly applied in the measurement phase. The refinement of the DQA4AirQuality framework to TDQM is shown in Fig. 3.

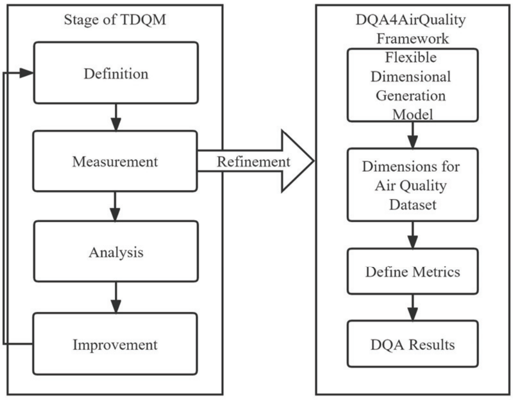

Fig. 3. Process of DQA4AirQuality framework.

5.3 Assessment Process

In the research, the assessment of the measurement phase integrated both subjective and objective evaluation approaches. Specifically, the objective evaluation approach drew on several studies, and some of their computational metrics were defined by authors [39]. The subjective assessment approach was based on a joint evaluation of the data dimensions by experts and users. A questionnaire was used to collect answers from experts and users who answered questions on such dimensions as usability, trustworthiness, reputation, relevance and confidentiality.

5.3.1 Definition of metrics

After defining DQA dimensions, the next matter is to define metrics for each dimension that quantify how well the dataset performs on such dimensions. Table 6 defines the corresponding metrics, both quantitative and qualitative, for each of the 12 dimensions in the DQA4AirQuality framework.

Table 6. Metrics definition and description

Dimensions	Metrics	Examples
Accuracy	The aggregate value of all sensor precision.	The numerical precision of the data values.
Completeness	The percentage of missing values/elements to the number of values/elements collected.	Null values.
Consistency	The percentage of outliers.	Outliers.
Timeliness	The frequency of data updates or number of not current rows.	Updated every hour.
Uniqueness	The percentage of records that have no duplicate data.	Two different monitoring records appeared in the same period.
Validity	The percentage of data formats and units in the expected range.	The measurement date does not exceed the date of release of the data and it is within a reasonable range.

on the next page

Table 6. continued

Dimensions	Metrics	Examples
Traceability	Track build and track renewal.	Experts make subjective decisions.
Understandability	The percentage of columns with metadata.	Loss of metadata.
Expiration	The proportion between the release time of the dataset after the expiration of its prior edition and the duration to which the dataset refers.	cd: current date; sd: starting date of the period of time referred to by the dataset; ed: end date of the period of time referred to by the dataset. 1–(cd–ed)/(ed–sd)
Compliance	Users determine 0–1 according to their requirements.	The representation of $PM_{2.5}$ does not meet international standards.
Data volume	The amount of raw data produced by the sensor with an initial value of 1; the number of attributes.	Experts make subjective decisions based on their experience.
Accessibility	The ease with which one can get to data.	Users determine 0–1 according to their experience.

5.3.2 Ranking of dimensions

The framework that has been defined up to this point contains 12 dimensions, and how these dimensions are weighted is the next question to be considered. Two options are offered as follows.

- Plan A: All dimensions are assumed to be of equal importance in the DQ score, with each dimension weighted equally. This is the simplest approach.
- Plan B: Using a multi-criteria decision-making approach, the 12 dimensions are ranked according to their importance, and a weight is assigned to each dimension.

In this study, plan B was used, and the analytic hierarchy process (AHP) was chosen for the DQ measurement process to select attributes and appropriate weights. Through this method, an importance judgment matrix was created based on the relative importance given by experts, and a weight ranking of all dimensions was finally obtained, as shown in Table 7. Accuracy has the highest weight, followed by completeness and consistency, while traceability and compliance have the lowest weight.

Table 7. AHP hierarchical analysis results

Dimensions	Eigenvector	Weight (%)
Accuracy	3.153	19.665
Completeness	2.817	17.573
Consistency	2.23	13.91
Timeliness	1.217	7.588
Uniqueness	0.991	6.183
Validity	1.757	10.958
Traceability	0.292	1.819
Understandability	0.559	3.489
Expiration	0.402	2.505
Compliance	0.292	1.819
Data volume	1.481	9.236
Accessibility	0.843	5.255

5.3.3 Aggregation method

The overall quality score of the dataset was obtained by weighting the sum of scores for each attribute. The following method was adopted to summarize the values of N univariate indicators.

$$C_{DQ} = \sum_{i=1}^{i} (a_i \cdot aM_i), \tag{1}$$

where a_i is a weighting factor, $a_i \in [0,1], and\ a_1 + a_2 + \cdots + a_n = 1$. M_i is a normalized value of the assessment of the i-th dimension.

5.3.4 Questionnaire design

In this study, 10 experts in the field of IoT or big data and 10 users were all invited to fill in a questionnaire, to obtain the results of a qualitative DQA. Questionnaires represent an important part of the measurement phase. The survey results were collected in an online format. An example of a user survey is shown in Table 8, where 10 questions were designed to obtain feedback from users and experts on the quality of the dataset, and users can score 0–10 based on their subjective experience.

5.4 Results

The consistency of dataset A is calculated as an example. As can be seen from Fig. 4, the data in this column of $PM_{2.5}$ show a serious right-skewed distribution, namely many great outliers. According to the characteristics of normal distribution, the data outside three standard deviations (3σ) can be treated as outliers. Then, the outliers greater than the upper limit and less than the lower limit were filtered out to obtain 198 outlines. Hence, the consistency of this column of $PM_{2.5}$ is 99.38%, and that of all columns can be calculated in turn. After the aggregation, the consistency of dataset A was obtained, namely 99.09%.

Table 8. Questionnaire for users and experts

Dimensions	Questions	Options
Accuracy	Q1: What is the average precision of commonly used sensors?	Open answer
	Q2: What is the degree of reliability and free of the error of the dataset?	0–1 scoring
Traceability	Q3: Is the information on the source of the data provided?	Yes or No
	Q4: How difficult is it to trace data sources?	0–1 scoring
Compliance	Q5: Is the data format compliant with the standard format?	Yes or No
	Q6: Is metadata missing?	Yes or No
Data volume	Q7: Is the amount and volume of datasets appropriate for your application?	Yes or No
	Q8: Are all 6 AQI-involved air quality data included?	Yes or No
	Q9: How sufficient is the amount of dataset instances?	0–1 scoring
Accessibility	Q10: How difficult is it for you to obtain the dataset?	0–1 scoring

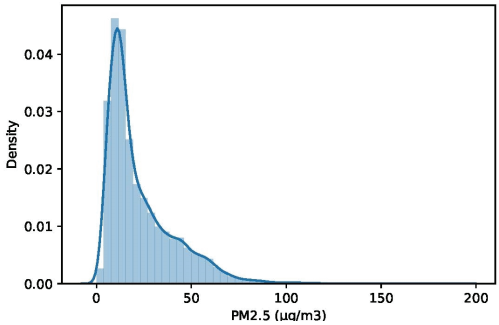

Fig. 4. Consistency of $PM_{2.5}$ in dataset A.

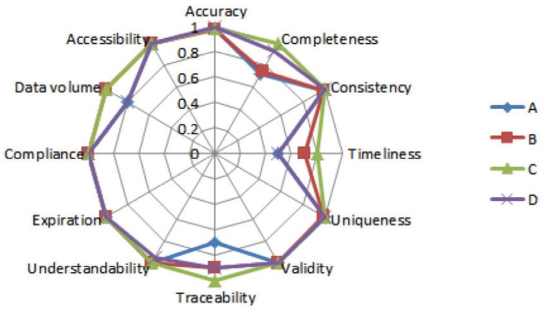

Fig. 5. DQ assessment result.

The results of the measurements obtained for each of the four datasets in 12 dimensions are reported in Fig. 5. Dataset C performs well in most dimensions, with the most variability in completeness, accuracy, traceability, currentness, and data volume. The other dimensions perform at comparable levels.

The results of the scores obtained using each dimension were aggregated to obtain the overall scores for the quality of the four datasets, as shown in Table 9. Datasets B, C, and D are of high quality and can be taken for subsequent research and use. Since air quality data in dataset A involve more countries and regions, more data are vacant with lower quality.

The next step is to analyze the reasons for lower scores in the dataset assessment results and consider how to adopt strategies and techniques to improve DQ. This is the work that should be done in the future.

Table 9. DQ score

	Dataset A	Dataset B	Dataset C	Dataset D
DQ score (%)	88.53	92.36	98.23	92.76

6. Conclusion

From the conclusion drawn from the measurements in Fig. 5 and the results in Table 9, it can be seen that data published centrally at the national level are better than data published decentrally. Datasets published by the WHO are mostly from official reports submitted by countries to the WHO or national agencies or government websites that report PM_{10} or $PM_{2.5}$ and NO_2 measurements. Other contributors of AQIs come from other United Nations institutions, development organizations, peer-reviewed journal papers and regional networks, including the electronic reports of the European Environment Agency on air quality [1]. This result is consistent with the limitations of dataset A summarized by the WHO, suggesting that the reasons for the poor quality of dataset A include: limited coverage, capturing only a small fraction of cities in some countries; omission of known data that cannot be accessed due to language barriers or limited accessibility; localized measures, with city averages based on ground measurement stations; high variability in measurement methods and techniques; different temporal coverage, where partial year data may not reflect the annual average due to seasonal variability; possible inclusion of ineligible data due to insufficient information to assess compliance; heterogeneous quality of measurements [40].

The results acknowledge the increasing efforts to expand the number of monitoring stations worldwide, highlighting the importance of publicly available, high-quality data. This can help raise awareness about the significance of accessible and reliable environmental information.

In addition, the specific metrics defined can be applied to more precisely un[derstand different air quality datasets in DQ. Based on these metrics, the results obtained for specific quality characteristics can be interpreted. The next step is to not only analyze the reasons behind lower scores for the dataset assessment results but also consider how to adopt strategies and techniques to improve DQ.

Conflict of Interest

The authors declare that they have no competing interests.

Funding

None.

References

[1] World Health Organization, "WHO's Fifth WHO Air Quality Database of over 6000 cities updated," 2024 [Online]. Available: https://www.who.int/news-room/questions-and-answers/item/who-s-fifth-who-air-quality-database-of-over-6000-cities-updated-april-2022.

[2] D. Zha, Z. P. Bhat, K. H. Lai, F. Yang, and X. Hu, "Data-centric AI: perspectives and challenges," in *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, Minneapolis-St. Paul Twin Cities, MN, USA, 2023, pp. 945-948. https://doi.org/10.1137/1.9781611977653.ch106

[3] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Computing Surveys (CSUR)*, vol. 41, no. 3, article no. 16, 2009. https://doi.org/10.1145/1541880.1541883

[4] L. Zhang, D. Jeong, and S. Lee, "Data quality management in the Internet of Things," *Sensors*, vol. 21, no. 17, article no. 5834, 2021. https://doi.org/10.3390/s21175834

[5] R. Y. Wang, "A product perspective on total data quality management," *Communications of the ACM*, vol. 41, no. 2, pp. 58-65, 1998. https://doi.org/10.1145/269012.269022

[6] C. Batini and M. Scannapieco, *Data Quality: Concepts, Methodologies and Techniques.* Cham, Switzerland: Springer, 2006. https://doi.org/10.1007/3-540-33173-5

[7] N. Wilantika and W. C. Wibowo, "Data quality management in educational data: a case study of statistics polytechnic," *Jurnal Sistem Informasi (Journal of Information System)*, vol. 15, no. 2, pp. 52-66, 2019. https://doi.org/10.21609/jsi.v15i2.848

[8] Z. Houhamdi and B. Athamena, "Impacts of information quality on decision-making," *Global Business and Economics Review*, vol. 21, no. 1, pp. 26-42, 2019. https://doi.org/10.1504/GBER.2019.096854

[9] D. Y. Siregar, H. Akbar, I. B. P. A. Pranidhana, A. N. Hidayanto, and Y. Ruldeviyani, "The importance of data quality to reinforce COVID-19 vaccination scheduling system: study case of Jakarta, Indonesia," in *Proceedings of 2022 2nd International Conference on Information Technology and Education (ICIT&E)*, Malang, Indonesia, 2022, pp. 262-268. https://doi.org/10.1109/ICITE54466.2022.9759880

[10] D. S. G. Gonzalez and J. N. P. Castillo, "Gestión total de la calidad de datos (TDQM) aplicada a la evaluación de calidad de datos abiertos de superficie de bosque natural de Colombia [Total Data Quality Management (TDQM) applied to the evaluation of the Colombian natural forest surface open data quality]," *Revista Ibérica de Sistemas e Tecnologias de Informação*, vol. 2021, no. Extra 40, pp. 454-469, 2021.

[11] Q. Liu, G. Feng, G. K. Tayi, and J. Tian, "Managing data quality of the data warehouse: a chance-constrained programming approach," *Information Systems Frontiers*, vol. 23, pp. 375-389, 2021. https://doi.org/10.1007/

s10796-019-09963-5

[12] Y. Heydarpour, P. Malekzadeh, R. Dimitri, and F. Tornabene, “Thermoelastic analysis of rotating multilayer FG-GPLRC truncated conical shells based on a coupled TDQM-NURBS scheme,” *Composite Structures*, vol. 235, article no. 111707, 2020. https://doi.org/10.1016/j.compstruct.2019.111707

[13] L. Poon, S. Farshidi, N. Li, and Z. Zhao, “Unsupervised anomaly detection in data quality control,” in *Proceedings of 2021 IEEE International Conference on Big Data (Big Data)*, Orlando, FL, USA, 2021, pp. 2327-2336. https://doi.org/10.1109/BigData52589.2021.9671672

[14] S. Loetpipatwanich and P. Vichitthamaros, “Sakdas: a Python package for data profiling and data quality auditing. In *2020 1st International Conference on Big Data Analytics and Practices (IBDAP)*, Bangkok, Thailand, 2020, pp. 1-4. https://doi.org/10.1109/IBDAP50342.2020.9245455

[15] R. T. Prasetyo, Y. Ruldeviyani, E. D. Purnamasari, and A. F. Wibowo, “Data quality assessment on lecturer primary data: a case study on higher education database at Ministry of Education and Culture Republic of Indonesia,” *IOP Conference Series: Materials Science and Engineering*, vol. 1077, no. 1, article no. 012036, 2021. https://doi.org/10.1088/1757-899X/1077/1/012036

[16] R. Vaziri, M. Mohsenzadeh, and J. Habibi, “Measuring data quality with weighted metrics,” *Total Quality Management & Business Excellence*, vol. 30, no. 5-6, pp. 708-720, 2019. https://doi.org/10.1080/14783363.2017.1332954

[17] W. Wijayanti, A. N. Hidayanto, N. Wilantika, I. R. Adawati, and S. B. Yudhoatmojo, “Data quality assessment on higher education: a case study of institute of statistics,” in *Proceedings of 2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, Yogyakarta, Indonesia, 2018, pp. 231-236. https://doi.org/10.1109/ISRITI.2018.8864476

[18] Y. W. Lee, D. M. Strong, B. K. Kahn, and R. Y. Wang, “AIMQ: a methodology for information quality assessment,” *Information & Management*, vol. 40, no. 2, pp. 133-146, 2002. https://doi.org/10.1016/S0378-7206(02)00043-5

[19] M. A. L. Pena and I. M. Fernandez, “SAT-IoT: an architectural model for a high-performance fog/edge/cloud IoT platform,” in *Proceedings of 2019 IEEE 5th World Forum on Internet of Things (WF-IoT)*, Limerick, Ireland, 2019, pp. 633-638. https://doi.org/10.1109/WF-IoT.2019.8767282

[20] S. G. Johnson, “A data quality framework for the secondary use of electronic health information,” Ph.D. dissertation, University of Minnesota, Minneapolis, MN, USA, 2016.

[21] R. Perez-Castillo, A. G. Carretero, I. Caballero, M. Rodriguez, M. Piattini, A. Mate, S. Kim, and D. Lee, “DAQUA-MASS: an ISO 8000-61 based data quality management methodology for sensor data,” *Sensors*, vol. 18, no. 9, article no. 3105, 2018. https://doi.org/10.3390/s18093105

[22] R. Perez-Castillo, A. G. Carretero, M. Rodriguez, I. Caballero, M. Piattini, A. Mate, S. Kim, and D. Lee, “Data quality best practices in IoT environments,” in *Proceedings of 2018 11th International Conference on the Quality of Information and Communications Technology (QUATIC)*, Coimbra, Portugal, 2018, pp. 272-275. https://doi.org/10.1109/QUATIC.2018.00048

[23] M. Joanna and G. Marek, “The concept of the qualitology and grey system theory application in marketing information quality cognition and assessment,” *Central European Journal of Operations Research*, vol. 28, no. 2, pp. 817-840, 2020. https://doi.org/10.1007/s10100-019-00635-y

[24] S. Kim, R. P. Del Castillo, I. Caballero, J. Lee, C. Lee, D. Lee, S. Lee, and A. Mate, “Extending data quality management for smart connected product operations,” *IEEE Access*, vol. 7, pp. 144663-144678, 2019. https://doi.org/10.1109/ACCESS.2019.2945124

[25] L. Ehrlinger and W. Woß, “A survey of data quality measurement and monitoring tools,” *Frontiers in Big Data*, vol. 5, article no. 850611, 2022. https://doi.org/10.3389/fdata.2022.850611

[26] Y. Wand and R. Y. Wang, “Anchoring data quality dimensions in ontological foundations,” *Communications of the ACM*, vol. 39, no. 11, pp. 86-95, 1996. https://doi.org/10.1145/240455.240479

[27] R. Y. Wang and D. M. Strong, “Beyond accuracy: what data quality means to data consumers,” *Journal of Management Information Systems*, vol. 12, no. 4, pp. 5-33, 1996. https://doi.org/10.1080/07421222.1996.11518099

[28] International Organization for Standardization, “ISO/IEC 25012:2008 Software engineering — Software

product Quality Requirements and Evaluation (SQuaRE) — Data quality model," 2008 [Online]. Available: https://www.iso.org/standard/35736.html.

[29] M. Scannapieco and T. Catarci, "Data quality under a computer science perspective," *Archivi & Computer*, vol. 2, pp. 1-15, 2002.

[30] A. E. Lewis, N. Weiskopf, Z. B. Abrams, R. Foraker, A. M. Lai, P. R. Payne, and A. Gupta, "Electronic health record data quality assessment and tools: a systematic review," *Journal of the American Medical Informatics Association*, vol. 30, no. 10, pp. 1730-1740, 2023. https://doi.org/10.1093/jamia/ocad120

[31] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, "Big data: the next frontier for innovation, competition, and productivity," 2018 [Online]. Available: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation.

[32] J. Merino, I. Caballero, B. Rivas, M. Serrano, and M. Piattini, "A data quality in use model for big data," *Future Generation Computer Systems*, vol. 63, pp. 123-130, 2016. https://doi.org/10.1016/j.future.2015.11.024

[33] L. Cai and Y. Zhu, "The challenges of data quality and data quality assessment in the big data era," *Data Science Journal*, vol. 14, article no. 2, 2015. https://doi.org/10.5334/dsj-2015-002

[34] A. Vetro, L. Canova, M. Torchiano, C. O. Minotas, R. Iemma, and F. Morando, "Open data quality measurement framework: definition and application to Open Government Data," *Government Information Quarterly*, vol. 33, no. 2, pp. 325-337, 2016. https://doi.org/10.1016/j.giq.2016.02.001

[35] N. G. Weiskopf and C. Weng, "Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research," *Journal of the American Medical Informatics Association*, vol. 20, no. 1, pp. 144-151, 2013. https://doi.org/10.1136/amiajnl-2011-000681

[36] UCI Machine Learning Repository, "Beijing Multi-Site Air-Quality Data," 2019 [Online]. Available: https://archive.ics.uci.edu/dataset/501/beijing+multi+site+air+quality+data.

[37] Kaggle, "Air Pollution in Seoul," 2020 [Online]. Available: https://www.kaggle.com/datasets/bappekim/air-pollution-in-seoul.

[38] UCI Machine Learning Repository, "Air quality," 2016 [Online]. Available: https://archive.ics.uci.edu/dataset/360/air+quality.

[39] L. Zhang, Y. Zhang, M. Zhu, L. Chen, and B. Wu, "A critical review on quantitative evaluation of aqueous toxicity in water quality assessment," *Chemosphere*, vol. 342, article no. 140159, 2023. https://doi.org/10.1016/j.chemosphere.2023.140159

[40] World Health Organization, "WHO ambient air quality database," 2024 [Online]. Available: https://www.who.int/data/gho/data/themes/air-pollution/who-air-quality-database.

Lina Zhang **https://orcid.org/0000-0003-1783-3458**

She received the M.S. and Ph.D. degrees in Computer System Architecture from Shaanxi Normal University, China, and Kunsan National University, South Korea, in 2012 and 2023, respectively. Since 2006, she has been working in her position at Baoji University of Arts and Sciences, China, where she is currently a lecturer in Artificial Intelligence. She is interested in Internet of Things, semantic web, and big data visualization.

Sukhoon Lee **https://orcid.org/0000-0002-3390-5602**

He received his M.S. and Ph.D. degrees in Computer and Radio Communications Engineering from Korea University, South Korea, 2011 and 2016 respectively. He was a research fellow in Department of Biomedical Informatics at Ajou University School of Medicine in 2016. He is currently an associate professor in Department of Software Science & Engineering at Kunsan National University. He is interested in data quality assessment, semantic web, data engineering, and sensor data platform.