Compliance to Confidence: A Data Quality Model for Central Banks

by Debasis Nandi and Sujeesh Kumar ^

In a complex data ecosystem, ensuring the quality of data becomes challenging for central bankers, particularly in a regulatory landscape. As measuring data quality is contextual and subjective, it is emphasized that the need for tailored measurement strategies to develop various data quality dimensions to measure the quality of data effectively. This article provides an approach for constructing a data quality index (DQI) to evaluate the quality of the data submitted by the regulated entities. The article outlines a stepwise approach for constructing the data quality index at various levels. The proposed DQI framework enables central banks and regulators to monitor and improve data quality systematically enhancing institutional credibility, regulatory and supervisory efficiency, and public trust.

Introduction

Central Banks play an important role, *inter alia*, in maintaining financial stability and ensuring the health of the banking sector. Daily operations of the banking system produce large amount of data. Such data is an important asset for institutions like central banks, multilateral bodies, and many other organizations particularly entrusted with data collection and its maintenance to support data driven policymaking. The ultimate objective of data management is the production and dissemination of quality data–a precious asset in the present world.

A key part of the functioning of the central banks involves the collection of huge volume of banking and financial sector data from the regulated entities (REs). In a complex and dynamic data ecosystem, measuring data quality also challenging. Automation has an important role in ensuring the quality of data being collected, processed, and maintained at the data repository of the central bank. It will enable to effectively monitor various economic indicators, obtaining regulatory and supervisory insights and facilitating data driven policies for the well-being of the public.

Over the years, the data collection and dissemination process has undergone several transformational changes due to the rapid technological advancements witnessed across the globe. Organizations, particularly central banks have adopted advanced statistical techniques and technology tools to validate the data generation process and best efforts have been made to address emerging data gaps and challenges. While addressing the challenges, an evolving multitude of non-traditional data is adding more complexity in the data ecosystem. Girard (2020) noted that one of the of organizational challenges for managing data in the (AI) era is equipping staff to the latest tools and technologies.

Measuring data quality is subjective in nature as the measurement involves various techniques and depends on the type of data produced. A well-defined structure of data quality framework with suitable dimensions is an appropriate way to measure the quality of data (Van G.B. 2023). Motivated by this fact, a structural approach for measuring various dimensions of data quality and deriving data quality indices for the data generation and collection process have been attempted. With this backdrop, this article has the primary objective of providing an approach to measure the data quality considering various data quality dimensions defined in the literature. The existing data quality frameworks by various organisations does not provide specific formula for calculating various measures of data quality dimensions, while it

^ The authors are from the Department of Statistics and Information Management, Reserve Bank of India. The views expressed in this article are those of the authors and do not represent the views of the Reserve Bank of India.

specifies broad guidelines for assessing various quality dimensions. Besides, frameworks do not suggest a data quality index (DQI) measure for data collection process and dissemination process separately.

In the Indian context, the Reserve Bank of India (RBI) has recently published supervisory data quality index (sDQI) scores for the supervised entities based on four data quality dimensions viz., accuracy, completeness, timeliness and consistency. The sDQI provides a measure of the supervisory data quality, forming the basis for supervisory examinations (RBI,2025). The sDQI is intended to measure the data quality of select supervisory returns in a supervisory data collection perspective.

In the central banking context, data provided by the regulated entities not merely for the supervisory purpose, while it is being used for regulatory, policy formulation, statistical data dissemination, research and various other purposes. The approach outlines in this article is not limited to supervisory data; it encompasses all types of data collected from regulatory entities through prescribed returns. Furthermore, the DQI presented in this article extends beyond the four data quality dimensions, covering data collection and dissemination aspects. This article thus helps bridge existing gaps in this aspect.

The rest of the paper is structured into five sections. The next section presents a brief review of the literature. In Section III a description of various data quality dimensions is given, while Section IV outlines the method to construct a data quality index using several dimensions discussed in section III. Finally, section V concludes the article.

2. Review of Literature

Data quality has been defined differently across the literature. Data quality is the extent to which the data satisfies the users' needs (Wang, 1998). One of the widely accepted definitions of data quality by Wang and Strong (1996) is 'fitness for use'. While it was argued by Strong et al. (1997), that fitness for usage varies for different users under different circumstances and therefore data quality is relative and cannot be evaluated independent of users. Federal Committee on Statistical Methodology (FCSM) defined data quality as the degree to which data capture the desired information using appropriate methodology in a manner that sustains public trust. The importance of data quality was highlighted by several authors. Poor quality of data leads to wrong conclusions and substandard decision making leading to financial losses. It can result flawed risk assessments and negatively affects the organizational performance, demanding data governance strategies (Redman, 2008; Kharti & Browm, 2010; Lee *et al.*, 2004)

Researchers have pointed out several challenges in measuring data quality. As the data changes over time, data categorized as 'high-quality' today may not remain the same in the future. This dynamic nature of data requires continuous monitoring and frequent reassessments (Batini *et al.*, 2009). Measuring data quality becomes complex and tedious when it comes to large volumes of data, particularly in a big data environment. It will be more difficult to track all dimensions of data quality (Muller *et al.*, 2012). In a dynamic data environment, the selection of dimensions may be contextual and relevant. A framework developed by Fadahunsi *et al.* (2009) addressed these challenges to a certain extent. For instance, the information quality framework categorizes dimensions into intrinsic, contextual, representational, and accessibility aspects providing a comprehensive approach for evaluating data quality. Furthermore, data consistency becomes a sizable issue when the number of data source increases, regardless 'fitness of use' for any particular purpose. Therefore, a standardized approach is appropriate when there is a disagreement regarding data quality between different domains of people. Brining quality process under a data governance structure would be a solution for ensuring data quality (Smallwood, 2014).

Subjectivity in the selection of various quality matrices is also another challenge in measuring data quality. These subjective measures can vary between users, leading to inconsistencies in quality assessments (Redman, 2008). High-quality data is a fundamental requirement for any information system, while inaccurate data costs organizations deeply in correction activities, lost customers, missed opportunities, and incorrect decisions. Issues like wrong data entry can reduce the accuracy of the data. This can also lead to wrong information when data is inappropriately reported or used (Olson,2003). Garreitt *et.al.,* (2014) investigated the association between organizational trust in the management and financial reporting aspects like accruals quality, mis statements, and internal control quality. They found that trust is significantly associated with financial reporting quality and varies relatively in decentralized firms, while not those in a centralized data environment.

Measuring data quality is an important step in the data quality management process. Various approaches for measuring data quality dimensions have been discussed in the literature, while the quality dimensions considered appear to be common across the literature. These approaches are mainly quantitative, where the quality of data is measured based on statistical techniques or simple arithmetical ratio or using some modeling approach that consider various quantifiable ratios and measures (Rahm & Do, 2000 and Chandola *et al.*, 2009). However, certain quality dimensions are not directly measurable from the systems and processes. Such dimensions are measured qualitatively using survey methods (Zhang *et al.*, 2005; Lemire *et al.*, 2009). Some authors have suggested benchmarking methods, mainly for a quality comparison, wherein data quality is evaluated against benchmarks from the leading organizations. This enables the organization to understand where their data quality stands in comparison to their peers (Batini *et.al.*, 2009).

A third approach would be a hybrid approach, which is the combination of both quantitative and qualitative methods. This approach recognizes that no single method can fully address all the complexities of data quality measurement, particularly in a heterogeneous and dynamic data environment. Hybrid models are adaptable and flexible, and they can easily integrate real-time data quality monitoring with ongoing feedback from users.

Several data quality frameworks have been proposed by organizations like the IMF, World Bank and many other multinational institutions. Such frameworks often integrate multiple dimensions and provide a structured approach to evaluate data quality across several aspects. A notable framework introduced by the IMF is known as Data Quality Assessment Framework (DQAF), which provides a comprehensive model for evaluating data quality. It incorporates dimensions such as accuracy, reliability and timeliness, and is specifically designed for assessing statistical data in government and international organizations (IMF, 2003). Another seminal framework proposed by Wang and Strong (1996) is known as Information Quality Framework (IQF), which mainly emphasizes that data quality is a multi-dimensional concept that includes both technical aspects like accuracy and consistency and user perceptions of data quality. Calculating a data quality index (DQI), based on suitable quality dimensions is a typical approach, which has been used in various fields such as healthcare, industries and finance, facilitating a single composite measure for data quality. Such composite measure enables organizations to track changes in data quality over time.

3. Data Quality Dimensions

There are several quality dimensions defined in the literature. A recent and comprehensive survey on data quality dimensions across the various disciplines was conducted by Carvalhoa, *et.al.,* (2025). They

surveyed and listed almost all quality dimensions in their paper, as there is no consensus on the determination of data quality dimensions. They identified around 66 quality dimensions, allowing for users to suitably selecting the dimensions and the development of data quality frameworks. Research in this field suggests that different dimensions provide different perspectives or have different dependencies based on the purpose of the data. Central banks and multilateral institutions have specified around 14 quality dimensions in their data quality frameworks. Some of the commonly used data quality dimensions by various organizations across different countries listed in Table-1.

'Fitness for use' is a broader definition of data quality which depends on the purpose, needs and priorities, and user perspectives of the required data. These requirements can vary across group of users. Even though data is accurate, it need not be of good quality if they produced too late to be useful, or cannot be easily accessed, or appear to conflict with other data. Thus, quality is viewed as a multi-faceted concept (Enrico and Ward,2004). Therefore, organizations defined their data quality dimensions depends on data collection and dissemination needs and usage. The number of quality dimensions also can vary according to the nature and type of data collected or disseminated. These quality criteria or dimension reflects an inclusive approach to quality definition and assessment.

Based on the data quality frameworks and quality dimensions reviewed across the literature, this article presented eight data quality dimensions which are comprehensive and takes care of data quality issues largely, encompassing collection and dissemination of data. Central banks or the regulators typically design a data collection format with a clear objective to gather, analyze and monitor economic and financial data. They also keep the purpose and goals of data collection in mind, with proper identification of relevant sources

Table1: Various Dimensions of Data Quality

Sr.No	Dimensions of quality	Bank of England	European Central Bank#	OECD	Australian bureau of statistics	Federal statistical office of Germany	Government of Canada
1	Punctuality and timeliness	✓	✓	✓	✓	✓	✓
2	Accuracy	✓	✓	✓	✓	✓	✓
3	Credibility			✓			
4	Accessibility/clarity	✓		✓	✓	✓	✓
5	Consistency						✓
6	Interpretability			✓	✓		✓
7	Relevance	✓			✓	✓	✓
8	Coherence	✓		✓	✓		✓
9	Completeness		✓				✓
10	Stability		✓				
11	Plausibility		✓				
12	Reliability		✓				✓
13	Comparability	✓				✓	
14	Cost-efficiency*			✓			

Notes: i. Some organisations are used punctuality and timeliness together and some are used separately in their data quality framework.
ii. * OECD does not consider cost-efficiency as a dimension of quality, while it is a factor taken into account in any analysis of quality as it can affect quality in all dimensions.
iii. # Dimensions ae relating to supervisory data quality farmwork
Source: Compiled by the authors from the websites of the various organisations.

and variables. The nomenclature of the data formats often differs from country to country. For instance, the European Central Bank named their data format as 'reporting templates' or 'data templates'. The Federal Reserve uses the term 'call report', and the Bank of England's data format is 'statistical return'. The Reserve Bank of Australia uses 'statistical forms' to collect economic and financial data which are then used to produce various statistical releases and tables. In the Indian context, the Reserve Bank uses a template called 'Return' for collecting statistical or regulatory data for the REs. For convenience, the terminology 'Return' is used throughout this article. With this background, the following quality dimensions provides an inclusive assessment of data quality.

3.1 Timeliness

The timeliness dimension is sometimes used interchangeably with punctuality, or both are used together as 'punctuality and timeliness'. In either case, the important aspect of data quality is the timely availability of the data. The timeliness dimension mainly evaluates whether the statistics intended to be collected by the organization have been received on time as per the prescribed timeline. Adherence to deadline for filing the data by the REs is vital, as the timely availability of data is important – especially when a particular data set relates to any other data, or it is to be read along with another set of data.

Sometimes timeliness referred to as how up to date the data is or how current the data is when produced or reported–connected to relevance of the data. Both approaches are used to assess whether data is provided or reported at the expected time. Typically, this measure is expressed as a ratio considering the timely availability of data relative to the total data being collected. Organizations generally prescribe timelines for submitting returns. The return may be of any form–supervisory, regulatory, or statistical.

In a standard data quality framework, the following formula may be used for measuring the timeliness dimension:

$$Timeliness\ (T) = \frac{n_r}{n_d + \mathrm{n_r}} \times 100 \qquad (1)$$

where, n_r is the number of returns submitted within the prescribed time and n_d is the number of returns submitted with a delay, *i.e.* after the prescribed time. This percentage measure should be weighted appropriately in the data quality index calculation.

3.2 Accuracy

The accuracy dimension of data quality is generally measured by the correctness or exactness of the of the data submitted by the REs. In a data filing process, the accuracy of the data determines the quality of the overall data filed by the entities. Accuracy reflects the real data which should have desirable characteristics such as being free from errors and deviations, closeness to the true value, and high precision. However, this is difficult to measure, as it is theoretically defined as the difference between estimated values and the true (unknown) values. Data revisions can give a good assessment of accuracy since they provide a mechanism for determining how estimates change over time as they approach their 'final' value (OECD, 2003). This approach is particularly suitable for capturing accuracy of the data filed by the REs, as revisions are common in data reporting, especially in banking or financial sector.

The extent of revision determines the quality of the data–whether the change is minimal or substantial. If the change is significant from the initial filing of the same data, then it is certainly a quality issue. It is also important to determine whether change is genuine or due to a data error. Moreover, the refiling or resubmission of data is not necessarily due to the actual revisions. Validation failure can sometime lead to failure in data filing, requiring the REs to resubmit the file.

In a data quality framework, the accuracy dimension should check the number of times the data is revised and the magnitude of revision during a reporting period based on a key indicator. A revision is defined as the difference between a later and an earlier estimate of the same key item. Considering these aspects, a formula for accuracy dimension may be defined as follows:

Let '*u*' denote the total number of times a particular reporting entity had to resubmit a specific return in a given reporting period, '*v*' denotes the number of times validation failures occurred for a particular return for a particular reporting entity, and 'w' denote the number of resubmissions not due to validation failures, such that $u = v + w$.

Relative mean absolute revision (RMAR) is calculated for all resubmissions *u* (resubmission due to validation failure, *v* and other than validation failure, *w*) using the below mentioned formula:

$$RMAR_u = \frac{\sum_u |Z_f - Z_l|}{\sum_u |Z_f|} \quad (2)$$

where, Z_f is the value reported in the first submission and Z_l is the last value (final value submitted for the key aggregate *Z*

$$\text{Then, Accuracy (Ac)} = \begin{cases} 0 & \text{if } RMAR_u > 1 \\ 100 - (RMAR_u \times 100) & \text{otherwise} \end{cases} \quad (3)$$

3.3 Credibility

Credibility measures the degree of trustworthiness of the entire data generation and submission process of a return. It assesses whether all data have been produced in an automated manner without manual intervention. The credibility of the data provided by the REs depends mainly on three aspects of data process.

(i) the extent of automation in granular level data capturing mechanisms,

(ii) automation of data aggregation and calculations process to meet the regulatory requirements; and

(iii) automation of data transmission process.

Generally, regulators have control over the third process, as the data submission channels are provided by the regulator. However, the first two processes are often not visible to the regulator. To assess them, data auditors visit entities or request information via surveys.

The granular data is captured through the online transaction processing systems (OLTP) or other automated systems such as core banking systems (CBS), treasury operations systems (TOS), *etc.* which are linked to the data warehouse (DW) of the REs. The data aggregation or return generation process occurs either in the DW or through management information system (MIS) using various programs with business logics to extract the data. Part of the aggregation sometimes manually performed by punching data into predefined data templates. These processes are expected to be in an automatic manner to increase the credibility of the data collection mechanism. The third level of data process is the data transmission level where various channels being used for filing returns. These includes system-to-system channel, file upload channel, application programming interface (API) based channel, and web based or screen-based submission channels. Among these, system-to-system and API based channels ensure fully automated data submission process offering the most credible means for data submission. The credibility measure is qualitative in nature and is derived based on the scores given to the REs for their return filing process as described above. A scoring matrix suggested for measuring credibility is given in Table-2.

Return wise scores for DGP and GAP may be obtained from the REs while the DTP score can be obtained from the data submission system provided by the regulators. Finally, a weighted average score may be derived for determining the credibility dimension.

3.4 Consistency

The consistency dimension of the data quality checks the violation of various validation rules

Table 2: Credibility Scoring Matrix

Level of Automation				
Category	**< 30 percent**	**30-50 percent**	**50-80 percent**	**> 80 percent**
Data generation process (DGP)	30	50	80	100
Data aggregation process (DAP)	30	50	80	100
Channels of Data Submission				
Data transmission process (DTP)	**system-to-system**	**API**	**File upload**	**Web based/others**
Scores	100	100	80	60

Notes: i. The scores are need not be fixed and can vary according to importance /levels of automation sets by the organisations.
ii. Percent of automation is to be obtained based on the number of returns automated in each process (DGP and DAP)

including business validations. The data items can be relational or static in a data file (Batini & Scannapieca, 2006). A data template is typically relational[1] in nature implying several numbers/cells are interconnected and involve calculations. The consistency dimension checks whether the data appearing across the format follows the logical and arithmetic operations and whether the requisite data point is reported across multiple sheets or returns. These are termed integrity constrains, which are properties that must be satisfied by all instances of a database schema.

In a data submission process–when same data point or data element is required to be submitted in different returns, and the data pertains to same reporting period, then it is expected that the same value is reported across all returns. Here the data point reported may be consistent across the returns.

Let c_e be the number of datapoints which are reported across multiple returns. If c_t is the number of datapoints (out of c_e) which are reported during data submission which not matching across the returns.

Thean a return consistency (*Co*) may be arrived as follows:

$$Co = \frac{c_e}{c_t} \times 100 \qquad (4)$$

[1] Even if the data are not relational, consistency rules can be defined. For instance, in the case of a questionnaire format, semantic rules are defined in a way similar to relational constrains (Atzeni, & De Antonellis,1993; Batini & Scannapieca, 2006).

3.5 Completeness

Completeness is a qualitative measure of data quality which describes the extent to which data values are sufficiently populated using the given information/guidelines/ definitions, *etc.* The data populating process consist of data aggregation which involves arithmetical or logical calculations. Primarily, REs populates the required data through an automated process or with some manual intervention. As the REs operate at different levels of technological environments, proper guidance for return preparation is very essential for them to streamline their return preparation activities. Generally, regulator provides necessary guidelines, data definitions, compilation manuals, updates on regulatory changes and changes in data requirements etc., through circulars and press releases. The REs is also expected to maintain such documents, and track the information provided to them on various returns. It is advisable to maintain a compilation manual or procedural document for each return preparation process. This document also serves as a business continuity document for the REs. Considering all these aspects and availability of requisite documents at the REs, a qualitative measure of completeness dimension can be developed, providing appropriate scores to the REs.

Completeness can be also measured quantitatively considering data gaps, missing observations, calculation errors, *etc.* Here it refers to the extent to which users receive all the data without missing

templates and missing values and the data are accompanied by related metadata. This includes both the dataset and additional information that helps users to understand the dataset in their specific contexts. The qualitative completeness dimension is maninly applicable to the evaluation of the quality of data dissemination process

Both types of measures assess different aspects of completeness, and they serve to provide a more holistic understanding of how data is complete or incomplete. In the case of quantitative measures, some of the characteristics of completeness one should look at are empty records, attribute completeness and entity completeness. Weighted completeness can be arrived giving appropriate weights to each characteristic of completeness. Technical score relating to each quality aspects of completeness may be assessed by data auditors in the organisation

3.6 Relevance

Another important dimension of data quality is the relevance dimension, which refers to the degree to which the data is appropriate, useful, and its applicability for a specific purpose. If data produced or disseminated by a central bank is not relevant for the intended users, it cannot effectively support policy making and analysis. This dimension is used to evaluate the quality of the data disseminated by the organization. The relevance of data depends on whether it provides useful insights to the users who wants to obtain their desired level of information. The relevance dimension is often qualitative, and it evaluates how well the data meets the needs of users or stakeholders. The disseminated data should align with the context– which means that data must relate to the domain and purpose of the analysis. The data must be up-to-date, usable, and actionable for decision making. Using this dimension of quality, the data managers or auditors can make a qualitative assessment of the data being collected and check whether data is relevant depending on the situation and the user's needs. Regular data user surveys and interaction will provide input to the data managers or auditors who may provide score for this dimension by building appropriate scoring matrix.

3.7 Stability

The stability dimension of data quality indicates how data remain consistent and reliable over time. It reflects the ability of the data maintains its integrity and usefulness over various time periods, ensuring changes in the dataset are tracked and controlled without affecting its quality. Stability dimension is measured either qualitatively or using quantitative metrics depending on the context and type of data. Qualitative measures can be arrived based user feedback or expert assessments. For instance, some of the characteristics like usability, traceability[2], and how well the data has performed or used in real applications or analytical exercise of the data may be assessed. If the data users continuously find that data is reliable and consistent, the data is likely to be considered as stable.

In a quantitative aspect, measures such as data drift, consistency ratio, and change rate can be used to assess the stability dimension of data. These simple measures often involve numerical calculations using formulas. For example, the data drift indicating the change in data over time measured by comparing the data distributions at different points of time using the divergence measures like Kullback–Leibler divergence or Jensen–Shannon divergence (Csiszar, I. 1975; Nielsen, F. 2021). Similarly, data consistency ratio provides the proportion of consistent data points over time. An alternative measure would be the change rate which is measured by the ratio of number of data changes during a period and total data point for the same period. This measure will tell the user that how quickly data changes over time.

[2] Traceability means availability of time series data implying the ability to track the history of data from its origin to its present period.

3.8 Accessibility

The accessibility is another important dimension of data quality which refers to the ease with which data can be accessed, retrieved, and utilized by the users when needed. It also refers to the metadata availability to users, including the form or medium through which information is accessed, data security features, and interoperability. The assistance provided to users may be adequate to get the complete information about data and its accessibility. Although the data possess the other quality dimensions like accuracy, timeliness, and completeness indicating high-quality data, it is not valuable unless it is accessible to the data users in an easy manner whenever required. This dimension is commonly measured using qualitative criteria–conducting periodic feedback surveys by the data mangers or data auditors. Data users' feedback is very important criteria to arrive the accessibility dimension of data quality. Feedback surveys can be conducted for the users of data (both dissemination and collection). This dimension also checks for the availability of support to the users on the data portal, ease of access and easiness navigating around the data portal.

4. Data Quality Index- Methodology

Using the various data quality dimension estimated, one can arrive at a weighted measure of data quality in the form of an index. Construction of such an index enable central banks to monitor data quality progress, identify areas for improvement for each return/publication, and ensure reliable decision-making. The quality dimensions can be weighted according to the importance of each dimension estimated. The weights need not be fixed and can vary according to the importance of the data quality dimensions set by the organizations. The applicability of the dimensions is distinct for data collection and data dissemination processes. The organization can select suitable quality dimensions for the construction of a data quality indices for both data collection and dissemination processes.

Typically, returns are submitted by the REs at different frequencies, *i.e.*, weekly, fortnightly, monthly, half-yearly, *etc.* The regulator needs to decide the frequency of the DQI to be calculated *i.e.*, either monthly or quarterly. All returns falling in the desired period may be considered for calculating DQI. If someone has to calculate DQI on a monthly or quarterly basis, all returns which are falling in that month or quarter irrespective of the frequency of the returns may be considered.

Let $(f_1,f_2,...,f_p)$ be the set of p different frequencies (weekly, fortnightly, monthly, quarterly, *etc.*) of returns/publications falling in a month. The data quality index to be constructed should cover all these returns of the p frequencies in a month.

Let E_i be the set of REs submitting return to the organization, $i = 1,2, ...,l.$, R_j is the set of returns filed by the REs, $j = 1,2,...,m.$, and D_k is the set of quality dimensions under consideration, k = 1,2,...,n. For the calculation of a data quality index, the data analyst has to determine the triplet: (E_i,R_j,D_k); $i = 1,2,...,l$; $j = 1,2,...,m$; $k = 1,2,...,n$.

The quality dimension scores for triplet (E_i ,R_j,D_k) measured for i^{th} entity, j^{th} return, and k^{th} dimension is denoted by d_{ijk}. These scores are then aggregated with appropriate weights w_k for arriving am entity-return data quality index, DQI (E_iR_j) and is defined as follows:

$$DQI(E_i, R_j) = \frac{1}{N}\sum_{k=1}^{n} w_k d_{ijk} \qquad (5)$$

where $N = \sum w_k$, appropriate weights for each dimension depend on the regulators data collection process and systems and their relative importance. In the Indian context, according to Verma & Nandi (2017), accuracy was the most important data quality dimension (31.25%), followed by consistency (21.25%), timeliness (20%) and completeness (11.25%). The study also considered uniqueness with a weightage of (16.25%) which is closely related to credibility dimension.

4.1 Entity Level DQI

The entity level DQI may be arrived by aggregating returns quality indices with appropriate weights based on the number of datapoints/cells of a particular return[3]. The entity level DQI is denoted by $DQI(E_i)$ and is defined as follows:

$$DQI(E_i) = \frac{1}{N}\sum_{j=1}^{m} \alpha_j DQI(E_i R_j) \quad (6)$$

where α_j are the weights based on the number data points/cells submitted by an entity E for a return R. Entities which are filing more data points will have proportionate weights in α_j. $N = \sum \alpha_j$.

4.2 Return Level DQI

The return level DQI may be arrived by weighting the overall business profile of the entity. The weights may be derived using the share of total banking business undertaken by the entity, E to the overall banking business, is a key indicator to give relative importance to the entity[4] E.

The return level $DQI(R_j)$ may be calculated as follows:

$$DQI(R_j) = \frac{1}{N}\sum_{i=1}^{l} \beta_i DQI\,(E_i R_j) \quad (7)$$

where β_i is the weights based on the entities business, $N = \sum \beta_i$.

4.3 Computation of DQI at the Regulators

An enterprise level data quality index can be derived by aggregating either entity-level DQI or return level-DQI. Accordingly, an enterprise level DQI, denoted by $DQI_{EP\ level}$ and is defined as follows:

$$DQI_{EP\ level} = \frac{1}{N}\sum_{h=1}^{(l,m)} \delta_h DQI_h\,(E,R)_{(h)} \quad (8)$$

where $N = \sum \delta_h$, depending on the choice of entity weights or returns weights for arriving an enterprise level DQI.

[3] If same set of returns are applicable for all REs. In case the returns are different for different entities, return weights may be calculated for the returns applicable to a particular entity only.

[4] This may be proxied by the sum of aggregate deposit and total credit from the previous financial year for banks.

Even though the data collection process of a Central bank is operating in a centralized environment, there are multiple departments/verticals/domains that take care of different sets of data. For example, foreign exchange market data is collected and published by the foreign exchange department, which is the domain owner of forex related data and returns. Similarly, banking data is collected and disseminated by the banking department or regulatory department. The department-level or domain-level data quality indices can be also estimated by grouping the returns which are applicable to a department or domain. Accordingly, a weighted average DQI can be derived considering the number of data points/cells submitted by a regulatory entity to a department or vertical. Such a department-level data quality index can be used for comparisons between different departments or domains. This will enable monitoring of data collection quality concerning divergent returns handled by different domains/departments of the central banks.

Following the DQI methodology mentioned in this article, dissemination quality indices can also be derived considering each statistical tables (similar to return) or for a publication (consisting multiple tables) using appropriate dimensions and weights. Adopting the approaches provided in this article may be useful for the organizations to institutionalize their data quality measurements and enhance overall data quality framework and enhancing overall data governance.

4. 4 Interpretation of the DQI

The DQI can provide a single measure of overall data quality, considering the importance of each dimension and the frequency with which the data is used. It is desirable to have thresholds for the DQI to categorize the data. A DQI score closer to 100 ($\geq$ 80) suggests excellent data quality, while lower scores indicate areas for improvement. If $70 \leq DQI < 80$, then data quality is good, while if $DQI < 70$, the

organization needs improvement in their data quality. The same criteria can also be used for any dimensions or any levels of DQI

5. Conclusion

This paper reviews various data quality dimensions across the literature and provides a robust and scalable framework for selecting contextual and content-dependent data quality dimensions and their estimation. This will facilitate central banks or organizations to adopt and implement suitable data quality dimensions and a data quality index at various levels for monitoring and improving their data quality. Even though, the article suggests eight quality dimensions and two distinct approaches for the data collection and dissemination processes, organizations may employ either process depending on their domain of operations.

Additional information at the organizations/ department/vertical levels can also incorporated into the data quality dimensions with appropriate weights. The weighting patterns given in the article are not strictly applicable to organizations, it is left to the organizations to decide upon their processes and systems.

While the data quality management is a continuous process, the framework provided in this article can serve as a benchmark for the other financial institutions or data-driven policymakers aiming to integrate data quality into their data governance strategies. This article contributes to the ongoing discourse on the enhancement of data quality framework within central banks and other data-driven organizations.

References

Atzeni, P. and De Antonellis, V. (1993), "Relational Database Theory", The Benjamin Publishing Company.

Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. (2009), "Methodologies for Data Quality Assessment and Improvement", *ACM Computing Surveys*, 41(3):1-41.

Batini,C. and Scannapieca, M.(2006), "Data Quality Concepts, Methodologies and Techniques", Springer-Verlag, Berlin Heidelberg.

Carvalhoa,A.M., Soaresb, S., Montenegrob, J. and Conceiçaob,L. (2025). "Data Quality: revisiting dimensions towards new framework development", *Procedia Computer Science* 253, 247–256.

Chandola, V., Banerjee, A., and Kumar, V. (2009), "Anomaly Detection: A Survey", *ACM Computing Surveys*, 41(3):1-58.

Csiszar, I (1975). I-Divergence Geometry of Probability Distributions and Minimization Problems". Annals of Probability. 3 (1): 146–158.

Enrico, G. and Ward, D. (2004). "Quality framework for OECD statistics getting our own house in order", paper presented in the conference on data quality for international organizations, Germany, May 2004.

Fadahunsi, K. P., Akinlua, J. T., O Connor, S., Wark, P. A., and Gallagher, J. (2019), "Protocol for a systematic review and qualitative synthesis of information quality frameworks in eHealth", *BMJ Open*, 9(3).

Garreitt, J., Hoitash. R and Prawitt,D.F. (2014), "Trust and Financial Reporting Quality" *Journal of Accounting Research*, 52 (5).

Girard, M. (2020), "Helping Organizations Master Data Governance", *Policy Brief No. 163*, Centre for International Governance Innovation.

IMF (2003), "Data quality Assessment Framework and Data Quality Program", International Monetary Fund, Washington.

Khatri, V. and Brown, C. V. (2010), "Designing data governance", *Communications of the ACM*,53(1):148-152.

Lee, Y.W., Pipino,L. , Strong, D.M., and Wang, R.Y.(2004), "Process embedded data integrity", *Journal of Database Management*,15(1):87-103.

Lemire, D., MacLellan, C., and Kargupta, H. (2009), "Task-Dependent Data Quality", *IEEE Transactions on Data Engineering*, 31(4):205-221.

Muller, H. J., Rojas, R. G., and Wilke, G. (2012), "Big data analytics and the role of data quality", *Information Systems and E-Business Management*, 10(1):37-52.

Nielsen, F. (2021). On a variational definition for the Jensen-Shannon symmetrization of distances based on the information radius. Entropy. 23 (4).

OECD (2003), "Quality Framework and Guidelines for OECD Statistical Activities", OECD, Paris.

Olson, J. E (2003), "Data Quality: The Accuracy Dimension", *The Morgan Kaufmann Series in Data Management Systems*, 3-23.

Rahm, E., and Do, H. H. (2000), "Data Cleaning: Problems and Current Approaches", *IEEE Transactions on Knowledge and Data Engineering*, 11(4):147-162.

RBI (2025). "Supervisory Data Quality Index for Scheduled Commercial Banks", Reserve Bank of India, Press release March 2025.

Redman, T. C. (2008), "Data Quality: The Field Guide", Digital Press.

Smallwood, R.F. (2014), "Information Governance: Concepts, Strategies, and Best Practices", John Wiley and Sons.

Strong, D. M. Yang W. L, and Wang, R.Y. (1997), "Data Quality in Context", *Communications of the ACM*, 40(5).

Van Gils, B. (2023), "Data in Context-Models as Enablers for Managing and Using Data". The Enterprise Engineering Series. Springer.

Verma P. and Nandi, D (2017), "Data Quality of Data Warehouse: A Case Study", *International Journal of Advances in Electronics and Computer Science*, 4(9).

Wang, R.Y (1998), "A Product Perspective on Total Data", *Communications of the ACM*, 41(2).

Wang, R.Y. and Strong, D.M. (1996), "Beyond accuracy: What data quality means to data consumers", *Journal of Management Information System*, 12(4).

Zhang, S., Lee, K. P., and Chen, D. (2005), "Measuring Perceived Data Quality", *Data and Knowledge Engineering*, 55(3): 289-319.