The National Data Quality Framework

Kirsti Pohjanpää[1]

[1]*Statistics Finland*

Abstract

The volume of data is continually growing, as is the data managed and disclosed by government institutions. It's crucial to establish consistent methods for recognizing and defining the quality of public sector data. As we aim to repurpose data for uses beyond its original intent (secondary use), ensuring its quality becomes increasingly important. Questions regarding reliability, quality, and usability are pivotal in the decision-making process. With the abundance of information, assessing data quality becomes even more significant. Therefore, widely accepted rules and criteria are imperative.

The National Data Quality Framework is composed of data quality criteria, indicators, as well as models and tools to support their implementation and management. Developed collaboratively with numerous stakeholders, these criteria and indicators have undergone testing in multiple pilots, reflecting the current understanding of data quality.

But what exactly does data quality entail? We've established eleven criteria to answer this question, divided into three groups. The first group assesses how accurately information reflects reality, the second focuses on how information has been described, and the last addresses how the information can be utilized.

Implementing these data quality criteria necessitates simultaneous action at three levels: nationally, within organizations, and for specific datasets.

Keywords: data quality, data quality framework, data ecosystem

1. Introduction

The volume of data is constantly increasing, as is the data opened and managed by government institutions. It is crucial to uniformly recognize and describe the quality of public sector data. Furthermore, the more we intend to repurpose data for purposes other than its original one (secondary use), the more essential it becomes to verify its quality. Questions regarding reliability, quality, and usability of data are essential in the decision-making process. With the abundance of information, assessing the quality of data becomes increasingly important. Therefore, widely shared rules and criteria are necessary.

Quality criteria can be used to describe and assess the quality of datasets. These criteria are part of the Data Quality Framework, assisting users in evaluating whether a dataset is of sufficient quality for its intended purpose. Over time, quality criteria support the enhancement of the quality of datasets and data repositories.

Quality criteria are intended to be a flexible tool; not all criteria, especially indicators, may be relevant in all situations or for all datasets. Additionally, it's important to note that the

intended use determines the level of quality described by the quality criteria. Quality criteria, especially their indicators, focus on structured data.

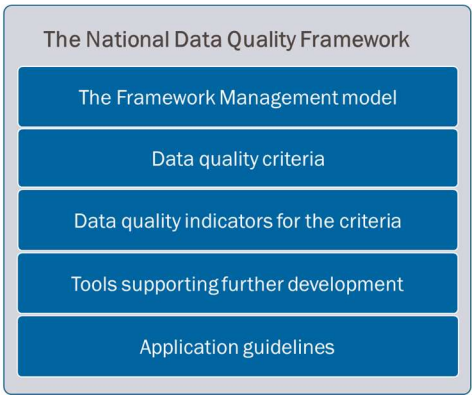
Quality criteria, along with their indicators, form a hierarchical structure, but they are interrelated and interconnected. Improving the quality with respect to one quality criterion may even weaken the quality of data described by another quality criterion. For example, if the goal is to achieve comprehensive coverage or particularly high accuracy of attribute data in a dataset, data timeliness declines.

This work is based on the Finnish Government's objectives regarding information policy. The project on opening and utilizing public data (TiHA) set up by the Ministry of Finance implements these objectives by promoting the wider and more efficient use of public data throughout society. The project term was from April 2020 to December 2022, followed by a year-long adoption promotion project. The TiHA project consists of four work packages: Strategic Aims (WP1), Opening Up Data, including common practices (WP2), Data Quality (WP3), and Semantic and Technical Interoperability, including guidelines for APIs (WP4). Statistics Finland led the work package Data Quality (WP3). The entire project was implemented and funded by the Finnish Ministry of Finance from 2020 to 2023.

2. The National Quality Framework for Public Sector

The project developed a national tool for assessing data quality. The data quality criteria and indicators, along with models and tools supporting their implementation and management, collectively form the National Data Quality Framework. It's essential to gather user feedback and periodically review the currency of the criteria and indicators. (Figure 1).

Figure 1: The National Quality Framework

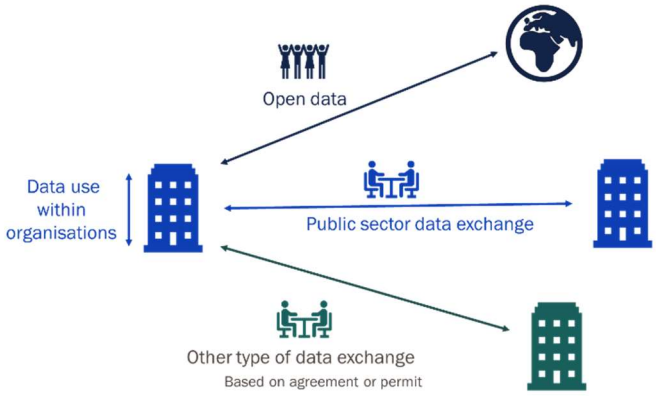


Additionally, the project created an online course to facilitate the use of these indicators. The course is freely accessible to everyone and is available in three languages (Finnish,

Swedish, and English, Data Quality - Significance and Description of Quality). In a bid to boost motivation, a brief promotional video was also produced for YouTube.

Quality criteria can serve various purposes. They prove valuable in evaluating data quality within and across organizations. Additionally, quality criteria aid in assessing the quality and usability of open data. (Figure 2.)

Figure 2: Broad scope of the Data Quality Framework



The data quality criteria and indicators were developed in collaboration with a significant number of stakeholders in Finland. Eleven agencies, including the Tax Administration, Natural Resources Institute Finland, Finnish Patent and Registration Office, and Finnish Customs, participated in this work in various capacities and at different times.

These criteria and indicators underwent testing in several pilot projects, reflecting the current understanding of data quality. In this paper, I primarily focus on describing the data quality criteria and their corresponding indicators (Chapter 3).

Unfortunately, the implementation of these criteria in Finland has been slower than anticipated and desired. The financial constraints faced by public organizations, including Statistics Finland, have posed challenges to the implementation process.

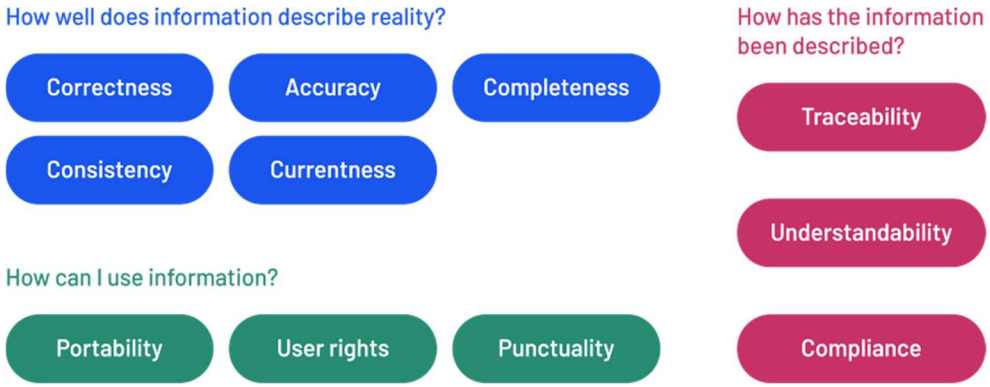
3. Data Quality Criteria

The data quality criteria and indicators serve as a tool for describing and assessing data quality, answering the fundamental question: *What does data quality mean?* The principles guiding the development of these criteria include being standard-based and easily applicable.

We've established eleven criteria to answer this question, divided into three groups. The first group assesses how accurately information reflects reality, the second focuses on how information has been described, and the last addresses how the information can be utilized.

The development methods employed involve discussion, collaboration, piloting, and iteration.

Figure 3: Data Quality Criteria



3.1 How well does information describe reality?

In the first category of data quality criteria, there are five criteria. This category includes criteria that answer the question of how well information describes reality.

Firstly, **correctness** evaluates how well the data in the dataset correspond to reality and helps identify systematic distortions. Indicators include *methodically produced values, incorrect values,* and *incorrect values*. For instance, data used for operational decisions represent the best understanding of accurate data. Accuracy, for example, is achieved when the declared salary for tax purposes matches the salary paid.

Secondly, **accuracy** describes how well the data in the dataset correspond to what is being sought. It describes how well the data hit the mark. Indicators for accuracy are *standard deviation,* and *outliers*. Accuracy describes, for example, the dispersion of indicator values, the proportion of outliers in the dataset, the accuracy of the classification and the scale of measurement (e.g. decimals, time, coordinates).

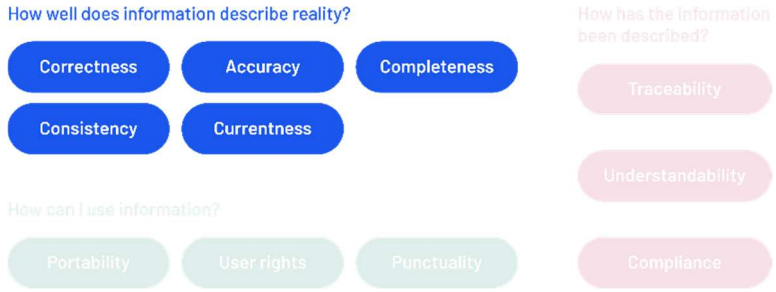
Completeness describes the extent to which the data cover the temporal and regional targets, as well as the target units and characteristics data. It also indicates the degree to which the dataset contains the desired data. Indicators include *temporal target coverage, regional target coverage, target units, shortcomings in characteristics, missing units, additional units, incomplete units,* and *incomplete characteristics*. For example, the dataset covers all units in a defined area, e.g. all enterprises in Finland. Regional coverage indicates whether all the target regions are included in the dataset (e.g. all Finnish municipalities), and if the dataset also covers Åland.

Consistency indicates that the data are consistent and non-contradictory. The indicator can also be used to describe the consistency between different datasets. There is only one indicator: *logic of data reviewed*. For example, there is an inconsistency when there are no dwellings in a residential building, or a person's date of marriage is earlier than their date of birth. Data consistency can be checked by means of validation/qualification rules.

Finaly, **currentness** describes the timeframe of the data in the dataset. The closer the data baseline period is to the present, the more current the data are. The baseline period refers to the point in time to which the data apply.

Indicators include *baseline period, creation period, review period,* and *change period*. For example, the baseline period associated with the dataset is provided with the data, allowing determination of data freshness. It can be the period between the beginning and the end of the year or a specific day. During data production, checking the data review and change periods is important.

Figure 4: How well does information describe reality?



3.2 How has the information been described?

There are three criteria for the second category. The main question is how the information has been described. The data quality criteria are organized into three types from the perspective of the information user.

First, **traceability** ensures that changes made to the dataset and its data can be traced, and the origin of the data is known. Indicators include *data source, data lifecycle*, and *change management*. For example, the origin of the data and its change history are documented, with timestamps available for each change. This enables verification and establishes the dataset's reliability.

Secondly, **understandability** assesses the extent to which a dataset contains metadata to aid users in understanding the data being utilized. Indicators are *dataset descriptions, definitions of concepts, data descriptions of characteristics,* and *customer feedback on comprehensibility*.

For instance, metadata descriptions provide sufficient detail about the dataset and its characteristics to facilitate understanding of the data content and significance. Code lists used for data characteristics are documented and consistent with the data, with descriptions available through accessible links. Essential concepts are explained, and links to relevant glossaries are provided in metadata descriptions.

And last, **compliance** ensures that the dataset and its characteristics adhere to established standards, practices, and regulations, which are specified in the dataset description. Indicators: *involve regulations and standards to be complied with*. For example, national conformity can be ensured by employing uniform national terminology and code lists when designing datasets. International conformity can be achieved by utilizing standard classifications endorsed by the EU and ISO language codes.

Figure 5: How has the information been described?

3.3 How can I use the information?

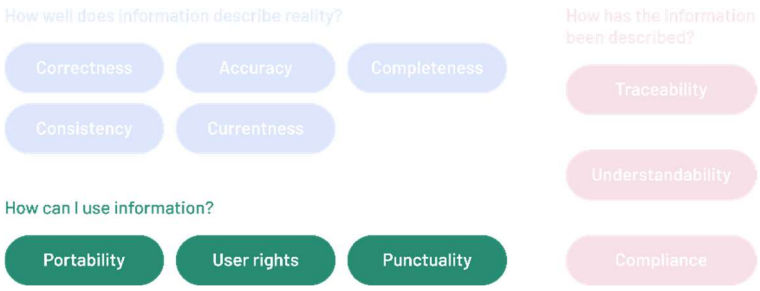
In the last group of criteria, we consider how we can utilize the information, focusing on three dimensions: portability, user rights, and punctuality.

Portability assesses whether the dataset is structured to enable automated processing across different information systems. Indicators include *the dataset's data model, permanent identifier of the target unit,* and *customer feedback on portability*. For instance, the dataset is structured in formats like .csv, .json, or .xml, with its structure documented using a schema.

User rights outline the rights granted to users regarding data usage and purposes, and indicators encompass *access rights*, and *restrictions on use*. For example, a dataset may be available for scientific research under specific conditions, and open data are typically licensed.

Lastly, **punctuality** ensures that the dataset is released and updated as scheduled, with adequate frequency to reflect changes. Indicators involve *compliance with due dates, frequency of updates,* and *identification of values changed in updates*. For instance, the publication schedule and frequency of updates are clearly defined and adhered to.

Figure 6: How can I use the information?



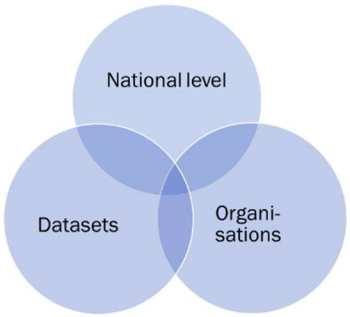
4. Implementation

Quality criteria are of no use unless they are widely adopted. Implementation of the data quality framework demands concerted efforts across all three identified levels: national, organizational, and dataset-specific (see Figure 7). At each level, numerous actors play crucial roles, underscoring the necessity for collaboration and support among them.

Recognizing that organizations differ in maturity and datasets vary in nature, it's imperative to approach implementation step by step.

Collaboration and support from diverse actors at all levels are essential for the successful adoption and integration of the data quality framework into existing processes and practices. By fostering collaboration and providing necessary support, stakeholders can collectively advance the quality and usability of data across various domains and sectors.

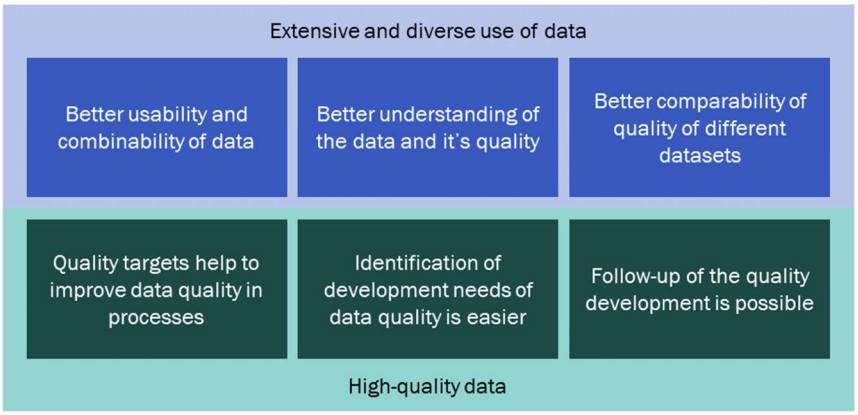
Figure 7: The three levels of implementation of data quality framework



The implementation of the data quality framework is of great significance for the functionality of the entire information ecosystem (Figure 8). Through extensive and diverse use of data, we can see numerous benefits, including enhanced usability and combinability, deeper understanding of data and its quality, and improved comparability across different datasets. These advancements culminate in the delivery of high-quality data that meets the needs of stakeholders and supports informed decision-making.

By setting quality targets, organizations can systematically enhance data quality throughout processes. Additionally, these targets facilitate the identification of development needs and enable ongoing monitoring of quality improvements. Thus, quality targets serve as guiding beacons, steering organizations towards continuous improvement in data quality and ensuring alignment with broader objectives.

Figure 8: The Role of Quality Assessment in Data Ecosystem



Acknowledgment

The work has been conducted with the support of the Finnish Ministry of Finance.

The work has been carried out extensively in collaboration. At Statistics Finland, the project has been actively promoted by, among others, Essi Kaukonen, Mervi Haakana, Jarmo Ranki, and Janika Tarkoma.