Data quality criteria and indicators – a proposal for a recommendation

Data Quality Framework, TiHA WP3
11 February 2022, updated on 19 December 2022

Statistics Finland

Contents

1 Introduction

This proposal for a recommendation on the data quality criteria and indicators for describing and assessing the quality of public administration data has been prepared as part of Work Package 3 (WP3) in the Project on Opening Up and Using Public Data (TiHA, VN/5386/2020), set up by the Ministry of Finance. The project ran from 30 April 2020 to 31 December 2022. The draft recommendation was made available for public comment between 6 September and 4 October 2021.

This recommendation describes a set of quality criteria and indicators that forms a clear and accessible tool for describing and assessing public administration datasets, specifically structured data, in a uniform and user-oriented manner. The recommendation is targeted at data producers working on data quality issues, as well as for data users interested in data quality. The aim is to provide, for the first time, a common national language and terminology for data quality experts and those interested in data quality.

The recommendation does not cover the adoption of the quality criteria and indicators in organisations or the maintenance and management of the tool developed. The document addresses the main aspects of applying the quality criteria and indicators, but the aim is also to produce separate application guidelines at a later stage. Solutions for communicating and disseminating information on data quality are also beyond the scope of this recommendation. On the other hand, the quality approach should also extend to developing data products and systems and customer service. Achieving versatility and combinability of data resources also requires investment in the availability of interfaces and electronic systems and improving awareness of and access to services. However, these aspects are not examined in this recommendation.

Data quality is a complex concept. This set of quality criteria and indicators aims to describe what quality means in the case of public administration datasets, and which aspects should be considered when describing data quality. The aim has been to produce a comprehensive and concise set of quality criteria that takes into account the different aspects of data quality. The project participants, who represented various perspectives on data design, processing, delivery, analysis and quality, played an essential role in achieving this goal. The indicators were also jointly developed, and they were tested with actual datasets in several pilot projects.

The quality criteria are largely based on the ISO 25012 standard, which is also used in several other countries as a model for describing data quality. In this project, this model has been supplemented with the criterion of *correctness*, which is important for base registers, as well as with a customer perspective. The criteria are also aligned with the European Interoperability Framework (EIF), the FAIR principles and the European Statistics Code of Practice. The set of indicators supports the application of the quality criteria. Some of the quality indicators are more closely linked than others to common standards or indicators used elsewhere.

Participants in the practical development work included Statistics Finland, the National Land Survey of Finland, the Tax Administration, Finnish Customs, the Natural Resources Institute Finland, the State Treasury, and the Social Insurance Institution of Finland. In addition to the parties above, the project team included participants from the Digital and Population Data Services Agency, the Finnish National Agency for Education, the Finnish Patent and Registration Office, and the Finnish Institute of Occupational Health. The quality criteria have also been presented to stakeholders such as the local government sector at TiHA project events and stakeholder meetings. They have also been made available for public comment on Statistics Finland’s website prior to the piloting.

2 Describing data quality

2.1 Objectives

The aim of these quality criteria and indicators is to facilitate the identification and description of data quality, specifically structured data, in different data exchange situations, as well as to promote a common approach to the comparison and development of data quality in public administration. The quality criteria and indicators form a uniform framework for describing data quality across organisational and sectoral boundaries. They help users assess whether a dataset is suitable for the intended use. Besides ensuring high data quality, the criteria have been selected to ensure the findability, combinability and interoperability of data and a smooth user experience.

In the design of the quality criteria and indicators, the aim has been to describe quality understandably, so that even users without previous experience of the dataset or expertise in determining data quality can also assess the data suitability. The aim has also been to make the implementation of the quality criteria and indicators as easy as possible.

3 Identified benefits

The objectives described in the previous section will facilitate the assessment of the suitability of data for different uses and thus also support a more diversified use of public data resources. The quality criteria and indicators provide a tool that promotes the opening up, interoperability and use of data.

Other benefits were also identified during the project work. For example, the indicators produce temporally comparable information about data quality that can be useful in the monitoring and improvement of data quality and process guidance. The expectation is that the indicators will also facilitate communication on data quality within organisations.

It is expected that the jointly developed quality criteria and indicators will increase national awareness and knowledge of data quality and help build a common language to describe and discuss data quality. They can also be used to determine the required data quality in development projects or when drafting legislation, for example. The quality criteria are intended as a tool for public administration, but they can also be useful in other areas of society like public procurement.

In the longer term, the quality criteria and indicators could be further developed to allow for national guidance and monitoring of the quality of public administration data if necessary.

At the national level, there are no previous similar models for determining data quality. Internationally, this recommendation is also one of the first national models for determining data quality. Therefore, it also contributes to the international debate and development in the field.

3.1 Boundaries

The following boundaries were set for the development of the quality criteria and indicators:

- The quality criteria and indicators are only applicable to structured data.
- The primary aim of the indicators is to serve a variety of data exchange situations, and the users of the data are mainly external users.
- The quality criteria and indicators do not address the issue of the desired quality level.

3.2 Data quality criteria increase understanding of data quality

Data quality is a broad and complex concept. Data quality is generally defined in terms of how well the data are suited to the needs of the data user. The quality criteria and indicators developed in the project address the different dimensions of data quality, primarily in the context of various data exchange situations within public administration and the different aspects of quality that are relevant from the data user’s perspective.

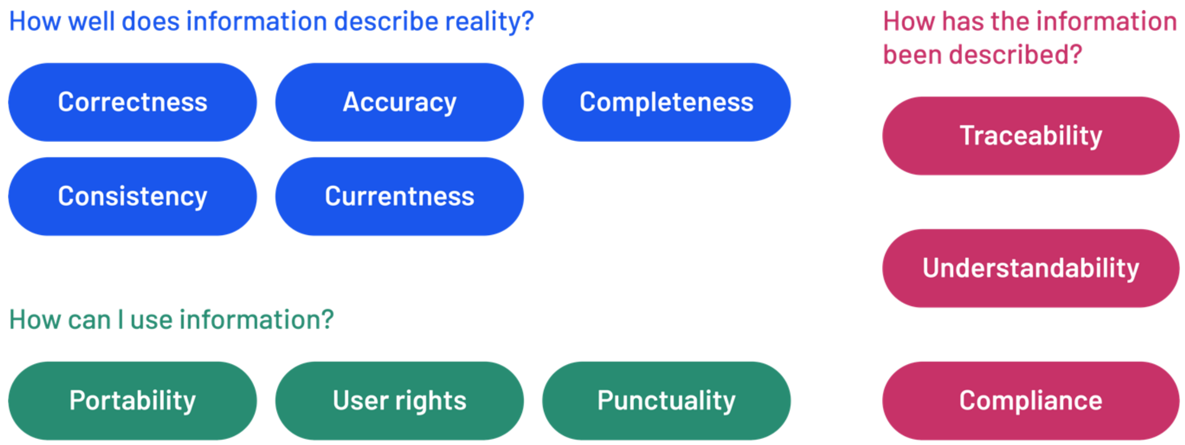

Figure 1. The quality criteria by category. When considering data quality, the key aspects tell the user what the data are about, and how well the data reflect reality. In terms of data usability, it is essential to know how the data can be used, and how well they are described.

In data exchange situations, the main challenge is to achieve an understanding of the data needs of the data users and the promises made by the data producers. The quality criteria can help identify which aspects are important for data quality. The user needs to have a comprehensive and accurate picture of the data content, the accuracy and relevance of the content and its description, and the accessibility of the data. The different aspects of quality are examined below through user-oriented questions and the associated quality criteria.

The quality criteria are divided into three groups, each with three to five criteria: “How has the information been described?” The quality criteria for this category are *traceability*, *understandability* and *compliance*. “How well does information describe reality?” The quality criteria for this category are *currentness*, *consistency*, *correctness*, *accuracy* and *completeness*. “How can I use information?” The quality criteria for this category are *portability*, *user rights* and *punctuality*.

Metadata are data that describe some other set of data. Data quality descriptions can therefore be considered part of the dataset metadata, i.e. the dataset description. Values produced by the data quality descriptions and indicators must be supported with other elements in the dataset description, such as the description of the target population and its location. Access to comprehensive descriptive data is essential for interpreting and understanding the values produced by the quality indicators. It was therefore decided that this first version of the quality description should also include descriptive elements that might belong to another metadata element than quality in a broader metadata model (such as the Metadata of Register Data model (JHS 201) or organisations’ internal metadata models). These include some of the indicators associated with the quality criterion of punctuality. The set also includes several other indicators for which key metadata must be determined and made available. At a minimum, such metadata should be provided to the data user in the quality description.

The structure of the quality description follows a strict and clear hierarchical model (Figure 3), which, together with the indicators, supports the understanding of the quality criteria. During the development of the quality criteria, and especially the indicators, many differences were identified in how quality is described between different organisations, types of data and sectors. The basic set of indicators was assembled so that the indicators could be applied as widely as possible. However, sectoral and other indicators can also be used to further support the quality description. These are treated less strictly in the hierarchy.

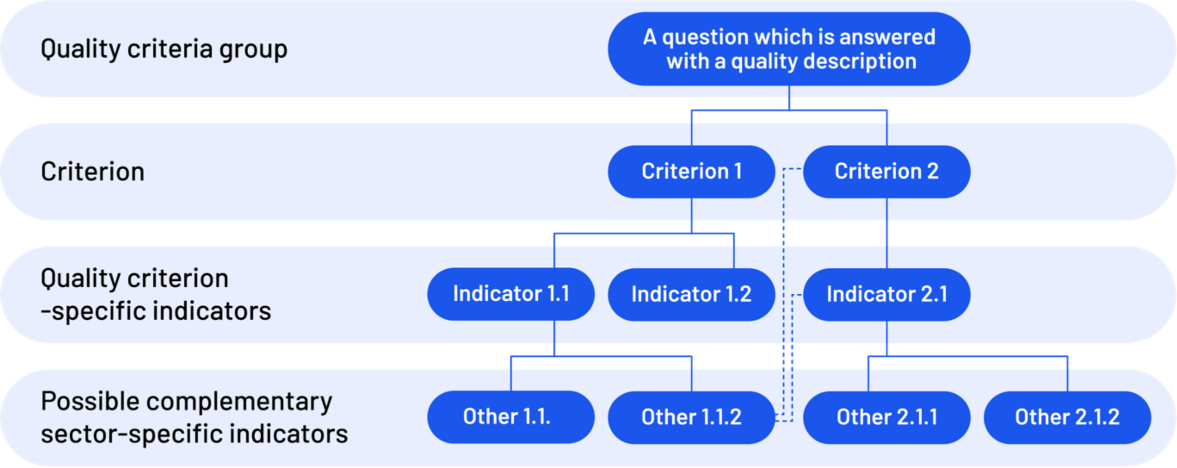
Quality criteria group
A question which is answered with a quality description
Criterion
Criterion 1
Criterion 2
Quality criterion -specific indicators
Indicator 1.1
Indicator 1.2
Indicator 2.1
Possible complementary sector-specific indicators
Other 1.1.
Other 1.1.2
Other 2.1.1
Other 2.1.2

Figure 2. Diagram of the structure of the quality description.

Despite the hierarchical structure, the quality criteria and indicators form a whole in which the different quality elements interact. An improvement in the quality of one criterion may even reduce the quality of another criterion. For example, if the aim is to achieve perfect data completeness or accuracy of the characteristics, the punctuality of the data will typically be reduced.

The quality criteria and related indicators are presented in more detail by quality criteria category in sections 3 to 5. In addition, the indicators are presented in a table in Appendix 3. The terminology used in this document (Appendix 1) plays an important role and should be developed further. Continuing this work and aiming to describe the concepts as precisely as possible will help us use common concepts to discuss data quality and achieve a shared understanding of what it means. This is essential for the further development of the quality criteria and indicators, as well as for comparing quality.

3.3 Attention to key aspects with the support of indicators

The set of indicators complements and supports the understanding of the quality criteria and the determining of data quality by giving concrete content to the criteria. The number of indicators varies between the criteria, and while some of the criteria describe quality in a relatively simple manner, others approach it from multiple perspectives. The set includes both quantitative indicators and descriptive indicators (yes/no). For the descriptive indicators, it is assumed that the data user has access to additional data on the subject of the indicator in other documentation of the dataset. The goal is that the indicators provide data users with a tool to compare the quality of public data resources.

When selecting the indicators, particular attention was paid to the following aspects:

- The indicators should provide a detailed description of the aspect of data quality measured by the associated quality criterion.
- The indicators should describe the quality, and especially the quality criterion, clearly, unambiguously and understandably.
- The indicators should be common, suitable for all users and reasonably easy to produce from a variety of datasets.
- Each quality criterion should have at least one indicator.
- For each indicator, a standard recommended format or ratio should be agreed.

The quality criteria and the indicators associated with each criterion are described in sections 3 to 5. The descriptions provided in this recommendation are concise and introductory but cover the main aspects of the indicators' application. Examples are also provided for illustrative purposes. In addition, the following general principles and advice should be taken into account when applying the indicators:

- The quality criteria and indicators are intended to be a flexible tool – not all the criteria, and therefore not all the indicators, are necessarily relevant in all situations. However, it is recommended that quality be assessed based on the overall quality description, possibly also from new perspectives, and focusing on the needs of the data user.
- It is important to determine the data users (i.e. the customers). Different branches of government have different customers. Consider who the customers are, and how you could integrate the customer perspective into the quality review. Are there any customer groups with specific quality-related needs?
- The indicators are applied to the dataset at the level at which the data are described. For example, if company data are aggregated to the industry level, the missing units indicator is applied at the level of missing industries.

- If it is difficult to produce an accurate description (e.g. a numerical value), an estimate can also be used. This is better than not filling in any value for an indicator relevant to the quality of the data under review. When using an estimate, this should be mentioned, with the reasons for the decision.
- The quality criteria and indicators form a tight package. The different elements are interlinked, and an improvement in quality in one area can mean a deterioration in another. For example, if the aim is to achieve perfect data completeness or accuracy of the characteristics, the punctuality of the data will typically be reduced.
- The use of complementary indicators is allowed and even recommended.

4 Category 1: “How well does information describe reality?”

The starting point for the data review is the phenomenon for which the data are to be used. The usability of the data is strongly linked to the substantive objectives of the data, i.e. which aspects of the phenomenon under review the data should describe. The data quality is determined by the degree to which the data satisfy the desired content. High-quality data describe the target phenomenon as accurately and correctly as possible. The data should also be as current as possible.

Data completeness describes the target population whose characteristics the dataset is intended to illustrate and how well the target population is represented in the dataset. The dataset description (i.e. the metadata) typically contains extensive data about the intended content of the dataset, while the data quality description highlights only the main aspects of the intended content. The criterion of completeness also examines the degree to which the target characteristics are included in the dataset.

The criterion of currentness covers several aspects that help assess the freshness of the data. The data should be as close as possible to the baseline period, i.e. the point in time to which the data apply. On the other hand, data that have not been recently updated are not necessarily of poor quality if no changes have taken place in the characteristics.

High-quality data describe reality accurately and correctly. This means that systematic biases or other sources of error have been identified, and their effect has been corrected. Consistent data have no internal conflicts.

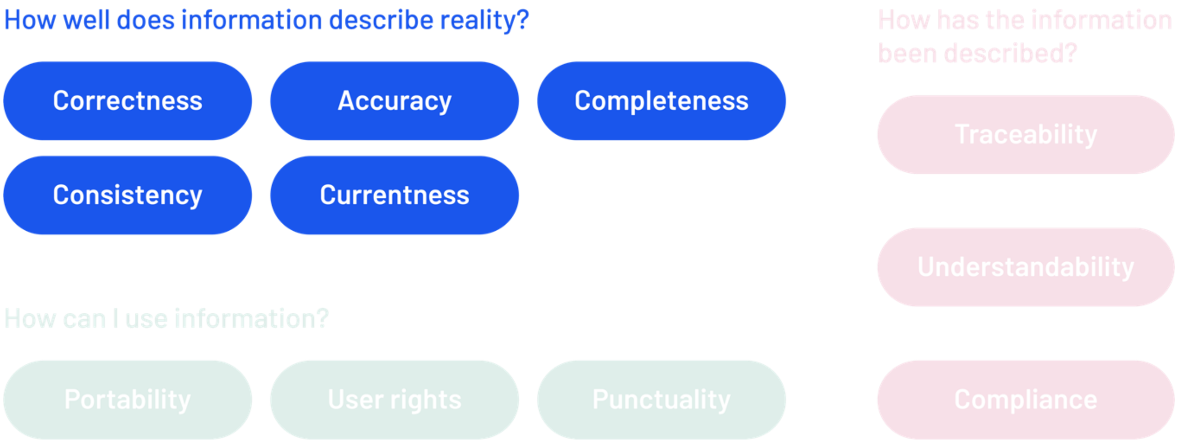

Figure 3. The quality criteria for the "How well does information describe reality?" category. Of these, correctness and accuracy are closely connected, if not overlapping, aspects. The other quality criteria in this group are completeness, currentness and consistency.

4.1 Quality criterion: Correctness

Synonym: *accuracy*

Description: *Correctness* describes how the data in the dataset correspond to reality. It also helps to identify systematic distortions in the dataset.

Example: The data leading to an operational decision represent the best understanding of the accurate data. For example, the data are considered accurate when the salary declared for tax purposes corresponds to the salary paid.

4.1.1 Methodically produced values

The *methodically produced values* indicator describes the proportion of values for a characteristic produced methodically or using surrogate data to all the values for that characteristic in the dataset.

- Level of assessment: characteristic
- Format: percentage / not relevant

The value of this indicator is calculated as the proportion of values for a characteristic produced methodically (i.e. by imputation) or using surrogate data to all the values for that characteristic in the dataset. This means the values produced using surrogate data or by imputation are not precisely the same as the actual values received by the target unit. For example, the indicator also covers situations where data from the previous year are used to fill in missing data. The indicator does not cover values that are corrected using actual values obtained directly from the data supplier.

Examples: Supplementing income data by using donor imputation and using sex, age, education and occupation as the criteria for selecting the donor. Data describing the activities of small enterprises are updated less frequently than once a year. The values to be used are therefore based on the data obtained in the previous survey.

4.1.2 Incorrect values

Incorrect values describes the proportion of target units with an incorrect characteristic value to the total number of target units in the dataset.

- Level of assessment: characteristic
- Format: percentage / not relevant
- Background: ISO 19157 (id 63 JHS 160)

The value of the indicator is expressed as the proportion of target units with an incorrect characteristic value to the total number of target units in the dataset.

The value can be produced using comparable datasets or by carrying out quality control on a sample basis. If it proves challenging to calculate the exact value, the percentage can be estimated based on experience.

Example: A dataset is known to include data that have not been updated. Based on experience, it can be estimated that five per cent of the values for the characteristic need updating.

4.1.3 Misclassification

Misclassification describes the proportion of target units with incorrectly classified characteristic values to the total number of target units in the dataset.

- Level of assessment: characteristic
- Format: percentage / not relevant
- Background: ISO 19157 (id 63 JHS 160)

The value of the indicator is expressed as the proportion of target units with incorrectly classified characteristic values to the total number of target units in the dataset. This indicator can also include target units for which it has not been possible to correct the missing data but whose values are not structurally missing (this means the characteristic is relevant for the target unit).

The value can be produced using comparable datasets or by carrying out quality control on a sample basis. If it proves challenging to calculate the exact value, the percentage can be estimated based on experience.

Example: The intended use of a building is incorrectly determined in n per cent of all buildings.

4.2 Quality criterion: Accuracy

Synonym: *unbiasedness*

Description: *Accuracy* describes how well the data in the dataset correspond to what is being sought. It describes how well the data hit the mark.

Examples: Accuracy describes 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).

4.2.1 Standard deviation

Standard deviation describes how spread out the characteristic values are relative to the mean. The purpose of this indicator is to give the data user an idea of the dispersion of the characteristic values.

- Level of assessment: characteristic
- Format: standard deviation

The value of this indicator provides the data user with information about the extent of the dispersion of the characteristic values.

It is also recommended to provide the mean alongside the standard deviation to make it easier to examine the dispersion of the values. Without the mean, the dispersion lacks a scale. However, the mean in itself does not describe accuracy so it is not included in the set as a separate indicator. Producing the value is also essential because both the mean and the standard deviation are needed to identify outliers.

Examples: When the standard deviation is low, the values of the characteristic are concentrated close to the mean, and when the standard deviation is high, the values are more dispersed. This may be due to inaccuracies, or simply be typical for the characteristic in question.

4.2.2 Outliers

The *outliers* indicator describes the proportion of outliers to the total number of target units in the dataset.

- Level of assessment: characteristic
- Format: percentage
- Background: ISO 25024

The value of the indicator is expressed as the proportion of outliers to the total number of target units in the dataset. An outlier refers to a target unit that receives a value that differs significantly from the majority of the other values.

A typical cut-off value for an outlier is considered to be 2.5 times the standard deviation from the mean. This cut-off value is also used for this indicator. This means that values that are smaller than the value obtained by subtracting 2.5 times the standard deviation from the mean or greater than the value obtained by adding 2.5 times the standard deviation to the mean are considered to be outliers. Many statistical software applications include automated tools for detecting outliers.

It should also be noted that an outlier can be a true or false outlier.

Example: When examining income data, individuals with significantly high income are the outliers in the dataset. The income data of high earners could cause problems later in the analysis.

4.3 Quality criterion: Consistency

Synonyms: *regularity, logical integrity of data*

Description: *Consistency* indicates that the data are consistent and non-contradictory. The indicator can also be used to describe the consistency between different datasets.

Examples: 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.

4.3.1 Logic of data reviewed

The *logic of data reviewed* indicator describes whether the data have been subjected to logic checks when compiling or processing the dataset.

- Assessment level: characteristic and dataset
- Format: yes/no

When applying this indicator, it should be stated whether the data have been checked using logical conditions or qualification rules.

It would also be useful to provide a more detailed description of the extent to which logical conditions have been used, especially at the dataset level. The data user is also interested in the details of the logical conditions used, and these should be highlighted in the dataset description.

Example: For the “address of establishment” characteristic, a logical condition has been used to determine whether the postal code of the address of the establishment corresponds to the municipality in which the establishment is located.

4.4 Quality criterion: Currentness

Description: *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 is the point in time to which the data apply.

Examples: The baseline period associated with the dataset is provided with the data. It can be used to determine the freshness of the data. The baseline period can be the period between the beginning and the end of the year or a particular day, for example. In data production, it is also important to check the data review and change periods.

4.4.1 Baseline period

Baseline period shows the point in time when the data in the dataset were collected, i.e. the point in time to which the data apply. Data processing always causes some delay and the data may have been collected before the dataset is ready for use by the data user.

- Assessment level: dataset
- Format: time period/not relevant

The value of this indicator shows how far in the past the events underlying the dataset have occurred. It also enables the determination of the delay between the processing of the data and their publication or availability.

Example: The baseline period of statistics may be several months ago because of the time needed to collect and process the data before their publication.

4.4.2 Creation period

Creation period indicates the time of creation of the target unit or characteristic. Another purpose of this indicator is to provide information on the period for which data are available.

- Assessment level: characteristic and dataset
- Format: time period

The value of this indicator tells the time the characteristic or dataset was created. It also indicates the date when the characteristic was included in the dataset, or the date when the compilation of the dataset was started. The creation period of a target unit is not necessarily the same as the creation period of the entire dataset.

The creation period is usually specified in the metadata. Besides the creation period, it is important to include the periods of comparable data in the dataset description.

4.4.3 Review period

Review period indicates the time the target unit or characteristic was revised.

- Assessment level: characteristic and dataset
- Format: time period / not relevant

The value of this indicator tells the latest revision date of the characteristic values or data in the dataset. It also shows when the data were last reviewed.

The review period is usually specified in the metadata.

4.4.4 Change period

Change period indicates the time of change of the target unit or characteristic.

- Assessment level: characteristic and dataset
- Format: time period/not relevant

The value of this indicator tells the time of change of the characteristic or dataset. It also shows when the data were last updated.

The change period is usually specified in the metadata. In the case of continuously or frequently updated datasets, the indicator is not meaningful, but even in these cases, the change period should be provided for reasons related to data lifecycle management.

4.5 Quality criterion: Completeness

Synonym: *coverage*

Description: *Completeness* describes the temporal and regional target coverage of the data, as well as the target units and characteristics data. It also indicates the degree to which the dataset contains the desired data.

Examples: 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. Over-coverage indicates that the dataset includes units that do not belong to the dataset. Under-coverage indicates that units belonging to the dataset are missing. Non-response is also included in under-coverage. On the other hand, completeness also indicates whether the dataset contains all the characteristics specified for the target units in the dataset, for example, the details of the population and area of the Finnish municipalities in the dataset, or whether address or turnover data have been provided for all enterprises in the dataset.

4.5.1 Temporal target coverage

Temporal target coverage indicates that the intended temporal coverage and frequency of the dataset have been described. Temporal target coverage refers to the period that the dataset is intended to cover, and the frequency with which the characteristic values have been measured.

- Assessment level: dataset
- Format: yes/no; additional data/specifications can also be included

The value of this indicator tells if the temporal target coverage of the dataset is specified, e.g. in the dataset description.

Example: The population of the national FinHealth 2017 study is the population aged 18 and over living in Mainland Finland in 2017.

4.5.2 Regional target coverage

Regional target coverage indicates that the intended regional coverage and density of the dataset have been described. Regional target coverage refers to the geographical area the dataset is intended to cover. Density refers to the accuracy with which the regional level is described in the dataset.

- Assessment level: dataset
- Format: yes/no; additional information/specifications can also be included

The value of this indicator tells if the regional target coverage of the dataset is specified, e.g. in the dataset description.

Example: The regional coverage of the FinHealth 2017 study is Mainland Finland.

4.5.3 Target units

Target units indicates that the dataset description also clearly specifies the other boundaries of the dataset besides the temporal and regional target coverage.

- Assessment level: dataset
- Format: yes/no; additional information/specifications can also be included
- Background: ISO 19157

The description of the target units should specify the relevant boundaries of the target units. The aspects of temporal and regional coverage are measured with separate indicators in the set.

Examples: A dataset only covers enterprises in specific industries. A dataset only includes data on buildings larger than 10 m^2.

4.5.4 Shortcomings in characteristics

Shortcomings in characteristics indicates if the dataset under review lacks characteristics that are relevant to the phenomenon the dataset describes. It is important to describe the shortcomings in more detail in the dataset description, for example.

- Assessment level: dataset
- Format: yes/no

It should be noted that to assess the quality of the dataset, it is important to identify the phenomena described by the dataset and the characteristics needed to describe or measure the phenomena. If the dataset is missing a certain aspect relevant for the analysis, i.e. some characteristics, this must be clearly stated to the data user.

4.5.5 Missing units

Missing units describes the under-coverage in the dataset, i.e. the percentage of target units missing from the target population.

- Assessment level: dataset
- Format: percentage
- Background: ISO 19157

The value of the indicator is expressed as the proportion of target units missing from the dataset (the target population) to the total number of target units in the dataset.

In the case of pre-compiled data, e.g. statistics, the missing values are usually corrected using statistical methods. In such cases, it can be assumed that the dataset does not have any missing target units. If necessary, missing target units may also be reviewed at the reporting level, e.g. at the industry level rather than at the enterprise level.

If it proves challenging to calculate the exact value, the percentage can be estimated based on experience. It is important to address the issue of missing units in the dataset description if there is a sufficient understanding of which target units are missing from the dataset.

4.5.6 Additional units

Additional units describes the over-coverage in the dataset, i.e. the percentage of target units present in the dataset that do not belong in the target population and are therefore considered to be additional units.

- Assessment level: dataset
- Format: percentage
- Background: ISO 19157

The value of the indicator is expressed as the proportion of additional target units to the total number of target units in the dataset.

In the case of pre-compiled data, e.g. statistics, the over-coverage is usually already corrected. If it proves challenging to calculate the exact value, the percentage can be estimated based on experience. If the dataset contains certain types of additional units, and it is possible to describe them, it is important to address this issue in the dataset description.

Examples: The same target units appear more than once in a dataset because the data have been imported from several sources, and not all of them use the same target unit identifiers. A sample of target units in a sample survey also includes individuals who have left the country because the address data in the database on which the sample is based have not been updated.

4.5.7 Incomplete units

Incomplete units describes the proportion of target units with missing characteristics to the total number of target units in the dataset.

- Assessment level: characteristic and dataset
- Format: percentage

The value of the indicator is expressed as the proportion of target units with even one missing characteristic to the total number of target units in the dataset. Structurally missing data, i.e. situations where a target unit should not have a value for a certain characteristic in the first place, are excluded from this indicator. The focus is on units where the missing value is relevant for the target unit concerned. For example, children should not have a value for “occupation”.

This indicator may not necessarily produce meaningful values in the context of pre-compiled data when evaluated at the dataset level. Similarly, in the case of larger unit-level datasets or longer time series, the proportion of incomplete target units may become very high, as individual data may be missing for almost every target unit. In the case of a characteristic-level review, it is also useful to describe the structurally missing data if the issue concerns a particular target population.

4.5.8 Incomplete characteristics

Incomplete characteristics describes the proportion of target units that have missing characteristic values. The indicator describes the degree to which the dataset contains values for a given characteristic.

- Assessment level: dataset
- Format: percentage

The value of the indicator is expressed as the proportion of target units that have missing characteristic values to the total number of target units with the characteristic in question. Structurally missing data are not taken into account for this indicator – they are only considered when the missing value is relevant for the target unit concerned. Structurally missing data refers to situations where the target unit should not have a value for a certain characteristic, for example, income should not exist for the types of income that the individual in question does not receive.

5 Category 2: "How has the information been described?"

Without a data description, even high-quality data can at worst be unusable. Data usability requires the description of the data and characteristics to ensure that these can be interpreted meaningfully and understandably. Compliance with different recommendations, standards, practices and regulations is important for data interoperability.

Traceability indicates whether the origin of the data and the changes made are known and can be examined retrospectively.

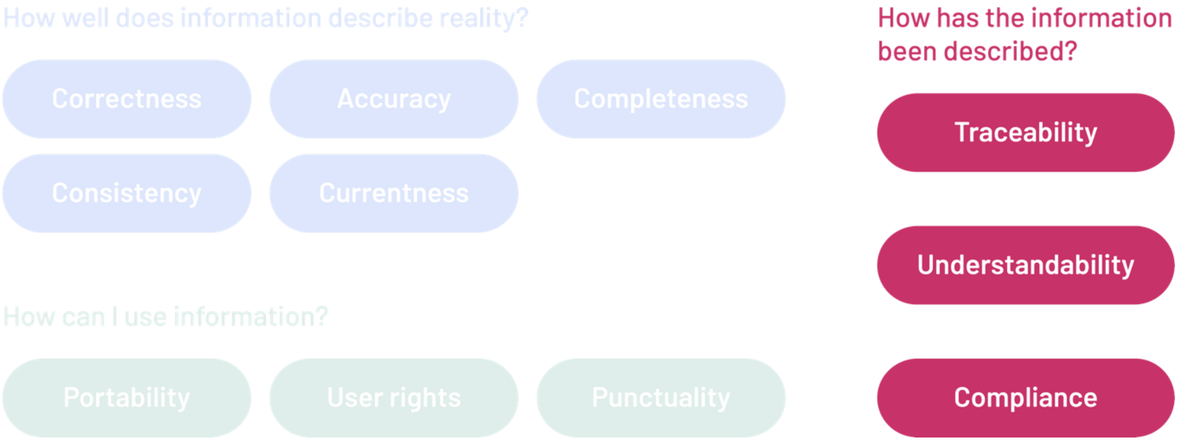

Figure 4. The quality criteria for the "How has the information been described?" category. Understandability and compliance cover very similar aspects and are partly overlapping. The third criterion in this category is traceability.

5.1 Quality criterion: Traceability

Synonym: *non-repudiation*

Description: *Traceability* indicates that changes made to the dataset and its data can be traced. The origin of the data is known.

Examples: The origin of the data and the change history are described, and time stamps of the changes are available. The data can be shown to be indisputable, and the data in the dataset can be verified.

5.1.1 Data source

The *data source* indicator shows the proportion of target units or characteristics for which source data are available.

- Assessment level: characteristic and dataset
- Format: percentage

The value of the indicator is expressed as the proportion of data with source data available to all the data in the dataset. The focus is on the data source from which the data have been directly obtained for the party assessing the data quality. This makes it possible to trace possible previous data history using the quality reports of the previous data supplier.

The data user will be interested to know the data sources, and if any source data are missing. Regarding the data source, it is important to specify the previous data supplier so that possible data quality aspects related to the different actors in a chain can also be assessed if necessary. For example, if the data source is missing for a certain target unit population, it would be useful to describe such shortcomings in the dataset description. Besides the data source, any edits to the data can also be described. This provides the data user with information on the data processing history in addition to the data source.

5.1.2 Data lifecycle

Data lifecycle indicates whether the data lifecycle is determined and described.

- Assessment level: dataset
- Format: yes/no

The value of this indicator shows whether the data lifecycle is determined and described. The data lifecycle describes the changes made to the data from creation to archiving or deletion. It also covers changes made to the data sources or to the production process, e.g. changes made to the calculation method.

It should also be noted that issues related to the creation and production of the data are essential information when assessing data usability.

5.1.3 Change management

Change management indicates that changes in the structural or source data of the characteristics are monitored.

- Level of assessment: characteristic
- Format: yes/no

The value of this indicator shows whether changes to the characteristics are recorded so that they can be viewed at a later stage.

It is also useful to specify how the change history can be accessed. The changes should be described at general level, e.g. in the metadata.

Examples: A record is kept of changes to the data characteristics, or the change history can be accessed through metadata.

5.2 Quality criterion: Understandability

Synonyms: *interpretability, comprehensibility*

Description: *Understandability* describes the degree to which a dataset contains metadata that help users understand the data being used.

Examples: The dataset and data characteristics are described in the metadata descriptions at a sufficient level to facilitate understanding of the data content and its significance. The code lists used for the data characteristics have been recorded and are consistent with the data. The descriptions of the code lists are available e.g. via links. Essential concepts are described, and links to the necessary glossaries are included in the metadata descriptions.

5.2.1 Dataset descriptions

Dataset descriptions indicates whether a dataset description is available, and in which languages.

- Assessment level: dataset
- Format: language versions

The value of the indicator is expressed by specifying the languages in which the dataset description is available.

The different language versions can be specified even if not all language versions are equally comprehensive. It is also important to add definitions and use consistent terminology in the dataset description. Users can thus check whether different datasets are comparable.

5.2.2 Definitions of concepts

The *definitions of concepts* indicator tells whether the concepts in the dataset are clearly defined and available, and in which language versions. Here, “concept” refers to those aspects of the dataset that cannot necessarily be measured directly.

- Assessment level: dataset
- Format: language versions

The value of the indicator is expressed by specifying the languages in which the concept descriptions are available. Clearly defined concepts are an important element of the data description. It ensures a common understanding of the terminology.

The different language versions can be specified even if not all language versions are equally comprehensive.

Examples: The data description includes a definition of the concept of “wellbeing”. This means describing the perspectives (for example economical, physical or life quality) through which wellbeing is considered in the dataset

5.2.3 Data descriptions of characteristics

Data descriptions of characteristics indicates whether the data descriptions and characteristic code lists are available, and in which language versions. Here, “characteristics” refers to the characteristics of the dataset that can be measured and that receive values.

- Level of assessment: characteristic
- Format: language versions

The value of the indicator is expressed by specifying the languages in which the descriptions are available.

The different language versions can be specified even if not all language versions are equally comprehensive. It is also important to add definitions and use consistent terminology in the descriptions. Users can thus check whether different datasets are comparable.

5.2.4 Customer feedback on comprehensibility

Customer feedback on comprehensibility indicates that it is possible to provide feedback on the comprehensibility of the dataset through an existing feedback channel or a targeted customer survey. Feedback is also monitored and followed up.

- Assessment level: dataset
- Format: yes/no

The purpose of the indicator is to describe whether the data user has an opportunity to provide feedback, and whether the feedback will be followed up. The feedback received can be used to improve the dataset.

In addition, a summary of the feedback or changes made to the dataset as a result of the feedback would provide useful additional information for the data users.

5.3 Quality criterion: Compliance

Synonyms: *compatibility, semantic conformity, conformity*

Description: Compliance indicates that the dataset and its characteristics comply with known standards, practices and regulations, and that they are specified in the dataset description.

Examples: For example, national conformity can be supported by using uniform national terminology and code lists when planning datasets. International conformity can be supported by using standard classifications adopted by the EU, as well as ISO language codes, for example.

5.3.1 Regulations and standards to be complied with

Regulations and standards to be complied with indicates whether the regulations, standards, good practices and recommendations followed in the dataset are listed in the dataset description.

- Assessment level: characteristic and dataset
- Format: yes/partly/no
- Background: INSPIRE/FAIR principles

There are many levels of regulations and standards (e.g. general, sector-specific, national and international standards). At least the most relevant standards should be listed, for example, in the dataset description. For characteristics, common code lists should be used. Deviations from the general standard should be highlighted.

Examples: A characteristic follows an international coding standard, but the code list has been supplemented with additional classes, depending on the organisation. The additional classes are defined separately from the coding standard to ensure that the new classes complement the standard but do not contradict it.

6 Category 3: "How can I use information?"

There are typically restrictions on the use of datasets in terms of who can access the data, for which purposes the data may be used, and the format in which the data are available. Access may also be restricted based on the data given to the data supplier at the time of collection or for data protection considerations. It is also essential to ensure that the data are available when promised. Data accessibility is also examined, especially from the perspective of portability.

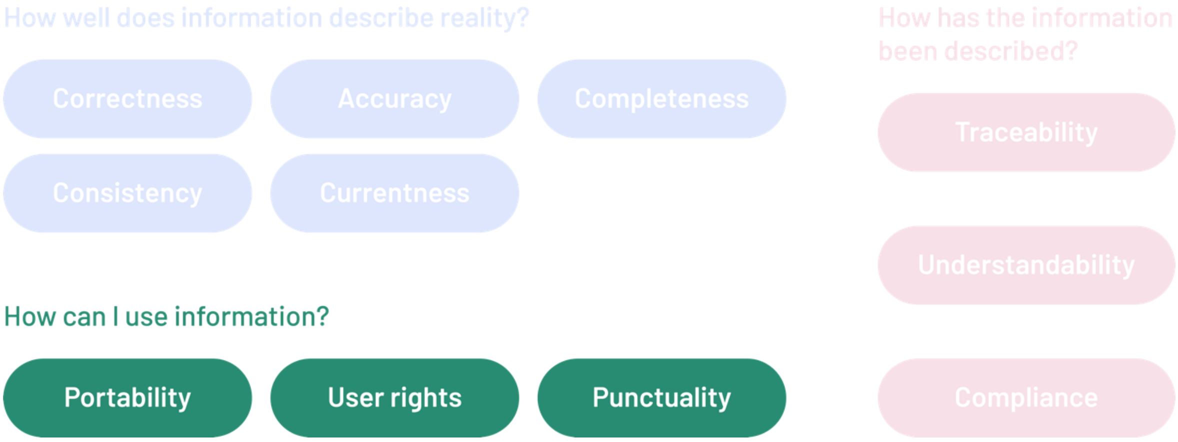

Figure 5. The quality criteria for the "How can I use information?" category are portability, user rights and punctuality.

6.1 Quality criterion: Portability

Description: *Portability* describes whether the dataset is structured so that the data can be processed in an automated manner and in different information systems.

Examples: The dataset is in a structured format (e.g. .csv, .json or .xml). The structure of the dataset is described by using a schema, for example.

6.1.1 The dataset data model

The *dataset data model* indicator specifies whether the dataset is described in a structured manner.

- Assessment level: dataset
- Format: yes/no

The value of this indicator shows whether the structure of the dataset is described using a data model/schema or equivalent standard. If the dataset is described using a data model, it is considered to be portable.

It is also useful to indicate the data model or standard according to which the dataset is described.

Examples: The structure of the building data in the national topographic database is described in the JHS 210 standard. The structure of a dataset is described by a schema (e.g. .xml, .json).

6.1.2 Permanent identifier of the target unit

Permanent identifier of the target unit indicates that the target units in the dataset have at least a dataset-specific permanent identifier that distinguishes the target units from each other.

- Assessment level: dataset
- Format: yes/no

When applying this indicator, it is sufficient for the assessment to have at least dataset-specific permanent identifiers. Of course, in terms of data usability, it would be more useful to use uniform national or even international permanent identifiers. In addition to the permanence of the identifier, attention should be paid to the uniqueness of the identifiers, i.e. the aim should be to avoid including the same target unit in the dataset more than once.

Example: The INSPIRE Implementing Rules require the publication of spatial object identifiers in the http URI format. (JHS 193)

6.1.3 Customer feedback on portability

Customer feedback on portability indicates that it is possible to provide feedback on the portability of the dataset through an existing feedback channel or a targeted customer survey. Feedback is also monitored and followed up.

- Assessment level: dataset
- Format: yes/no

The purpose of the indicator is to describe whether the data user has an opportunity to provide feedback, and whether the feedback will be followed up. The feedback received can be used to improve the dataset’s portability. In addition, a summary of the feedback or changes made as a result of the feedback would provide useful additional information for the data users.

6.2 Quality criterion: User rights

Description: *User rights* describes the user rights to the data, and how the data can be used (i.e. for what purposes).

Examples: For example, a dataset is available for scientific research, subject to certain restrictions. Open data are licensed.

6.2.1 Access rights

Access rights describes how access to the dataset is restricted, i.e. who can use the data.

- Assessment level: dataset
- Format: access restrictions

The value of the indicator indicates who has access to the data. For example, the dataset may be open data or public data, or only available under a time-limited licence, contract or for official use. Access may also be restricted for data protection considerations or based on the information given to the data supplier at the time of collection. In particular, the use of personal data is limited by the General Data Protection Regulation (Regulation (EU) 2016/679).

Examples: Statistical data are public data. Statistics Finland’s unit-level data are available under licence, e.g. for research purposes.

6.2.2 Restrictions on use

Restrictions on use describes the purposes for which the data in the dataset may be used.

- Assessment level: dataset
- Format: restrictions on use

The value of the indicator specifies the purposes for which the data may be used. For example, the dataset may only be available for official use (e.g. in decision making or as reference material), under a licence or contract, or as permitted by an open licence.

It is also useful to describe briefly any restrictions on the use of the data. In direct data collections, the use purposes notified to the data supplier during the data collection restrict how the data can be used or combined with other possible datasets.

Examples: Statistics Finland’s unit-level data may be used under licence to produce anonymous and aggregated survey results. Published data may be used freely, as long as the source is acknowledged.

6.3 Quality criterion: Punctuality

Synonym: *timeliness*

Description: *Punctuality* means that the dataset is released at the indicated time and updated with sufficient frequency to reflect changes in the dataset.

Examples: The time and frequency of publication have been specified. Changes to the release schedule are announced in advance.

6.3.1 Compliance with due dates

Compliance with due dates describes the monitoring of the planned delivery schedule against the actual delivery schedule.

- Assessment level: dataset
- Format: delay/not relevant
- Background: European Statistics Code of Practice

The indicator is used to describe the delay of the delivered material in relation to the agreed due date (actual delivery date versus agreed delivery date).

In addition, in the case of a delay, the reason for the delay can be described to the data user as additional information.

6.3.2 Frequency of updates

Frequency of updates describes the regular update frequency of the dataset.

- Assessment level: characteristic and dataset
- Format: written description
- Background: ISO 19139 MDMaintenanceFrequencyCode and maintenanceNote

The value of the metric should describe the regular update frequency with expressions such as “real-time”, “continuously”, “weekly”, “monthly”, “once a year” or another update frequency.

Examples: The road network and street names are **continuously** updated. Administrative boundaries and buildings are updated **annually**. Other sites are updated on a map sheet basis as part of a periodic updating process **every five to ten years**.

6.3.3 Values changed in the update

Values changed in the update describes the proportion of values changed in the update to all the values in the dataset when comparing the updated dataset to the previous version.

The purpose of this indicator is to describe the magnitude of the change in the dataset caused by the updates.

- Level of assessment: characteristic
- Format: percentage

When applying this indicator at the characteristic level, the proportion of changed values can be separately determined for each characteristic. However, at the dataset level, it is probably easier to describe the magnitude of the change in words. The purpose of this indicator is to describe the impact that waiting for an update may have on the content of the dataset already in use, for example. It should also be noted that even in the case of continuously updated data, providing a description of the magnitude of the change can be useful for the data user. For example, the proportion of values that have changed in the update can be viewed over a period that is meaningful in terms of the update.

Examples: After a dataset delivery, the Business Register Information Service makes a new delivery which includes new units (i.e. new businesses) and their characteristics, as well as any changes made to the characteristics of existing units and possible information on the closure of a business.

7 Appendices

Appendix 1. Descriptions of the terms used in the document

This appendix provides descriptions of the terms used in this recommendation.

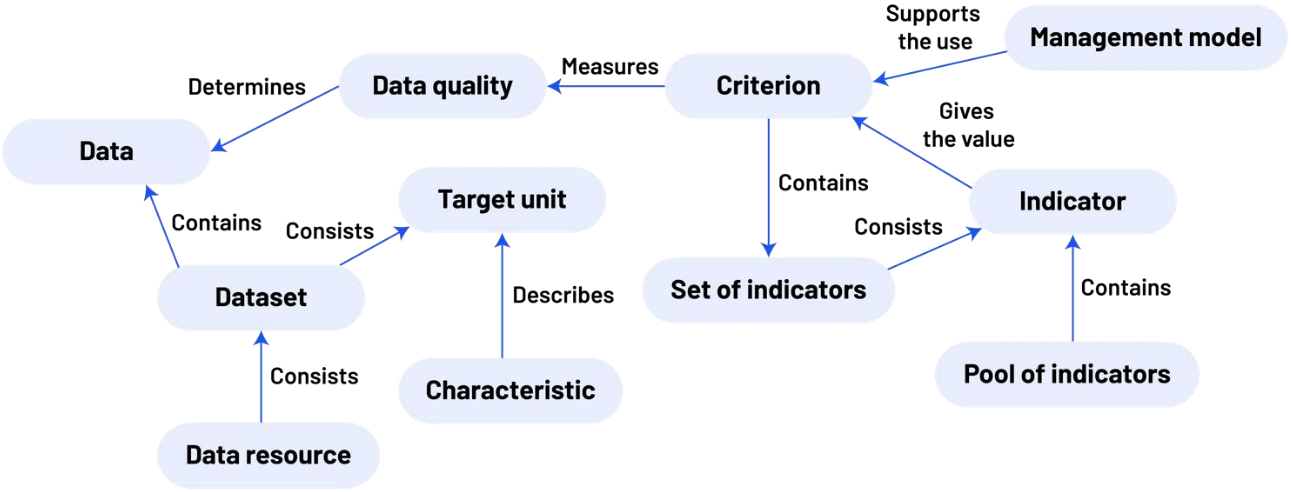

Figure 6. Illustration of the terms used to describe data quality in this document.

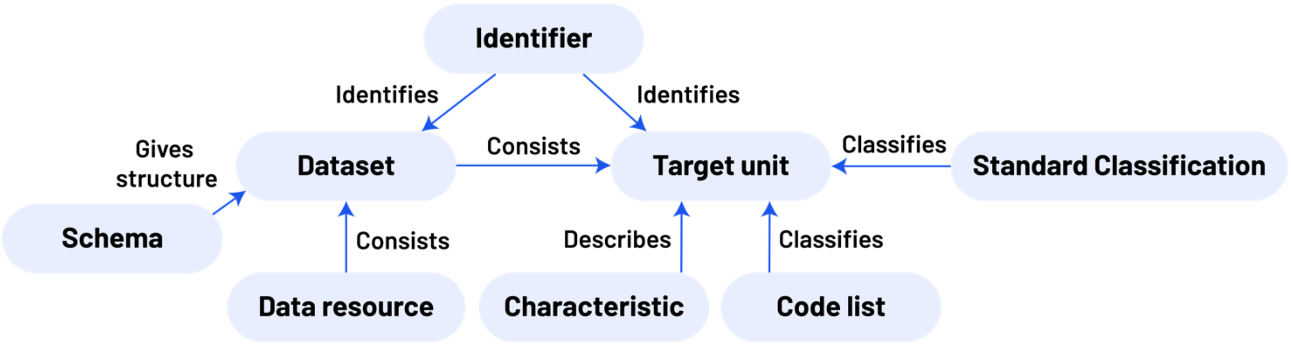

Figure 7. Illustration of the terms used to describe structured data in this document.

Table 1. Descriptions of the terms used in the document

Term (synonyms)	Description
Target unit (class, statistical unit, unit, target)	Unit of observation in the dataset.
Code list	A collection of codes for units that differ in certain characteristics.
Quality criterion	Describes the quality from a particular perspective.
Indicator	Gives practical expression to the quality criterion and measures the quality of the dataset. Each indicator is associated with a specific quality criterion. However, each quality criterion can have several indicators.
Pool of indicators	A collection of potential and proposed indicators, from which the most suitable were selected during the TiHA project as a set of indicators for the quality criteria. In the future, the pool will also serve as a collection of potential indicators to complement the current set.
Set of indicators	A collection of indicators produced to support the application of the data quality criteria.
Characteristic (feature, attribute, target value, variable)	Data describing the target unit.
Base register	A lower-level data resource or a data resource used as background data in data processing.
Structured data	Portable data defined by metadata that can be used to structure the data.
Schema	Defined representation of the data structure.
Standard classification (classification standard)	A classification recommendation based on international standards such as those laid down in EU directives.
Data user	The person using the data. From the perspective of the data handling process, the last person in the process, the person who actually uses the data. The data user can also further process the data, but this is a new process compared to describing the data quality.
Data quality	The degree to which the characteristics of a dataset meet requirements or objectives. The suitability of the data for the purpose the data user intends to use the data, and for which the data producer provides the data.
Data	Data can refer to a variety of things such as a string, a message, a fact, an observation, an interpretation or a perception. Here, data refer to the lowest-level information about a phenomenon or information produced by processing such data.
Dataset	A set of data stored on a data medium.

Data resource	A dataset or collection of datasets formed for a specific purpose, consisting of logically or physically related data.
Identifier	String of characters used for identification.

Table 2. Descriptions of the statistical concepts used in the document

Term (synonyms)	Description
Under-coverage, statistical concept	Under-coverage is related to the study population. The population must have a frame, i.e. a list of units for which data are to be collected by the sample survey. Under-coverage means that the frame is missing some of the units belonging to the population, i.e. the target population of the survey. For example, persons without a telephone are missing from the frame of units of a telephone survey.
Imputation, statistical method	Imputation is the process of replacing a missing or anomalous value in a dataset by using an imputation method. Possible imputation methods include logical imputation (correcting for logically impossible errors, e.g. a child cannot be older than parents), hot-deck imputation (the missing value is obtained from another respondent), cold-deck imputation (the missing value is obtained from a previous response by the same respondent), and regression and other model-based methods (a statistical model is used to predict the missing value).
Population, statistical concept	A population is the group of units that is the subject of the survey, and for which data are to be collected, e.g. citizens of voting age. It is also more precisely referred to as the target population. A frame population refers to a target population covered by the register or another list of units used in the survey, but it does not always fully match the survey target population (see over-coverage, under-coverage).
Outlier, statistical concept	An outlier is a value that differs significantly from the vast majority of other observations. An outlier can be a true or false outlier. Outlier values can have a significantly distorting effect on the statistical indicators used, such as the mean, standard deviation or regression line.
Over-coverage, statistical concept	Over-coverage refers to target units in the sample frame that are no longer part of the target population, e.g. people who have been institutionalised, have died or have emigrated. There will always be some of these cases among the sample units because the registers from which the samples are drawn are not always completely up to date.

Appendix 2. Principles and standards referenced in the document

Table 3. Principles and standards referenced in the document

Name	Description
EIF	European Interoperability Framework
European Statistics Code of Practice	European Statistics Code of Practice (CoP)
FAIR	The FAIR principles: The data should be Findable, Accessible, Interoperable and Re-usable
ISO 19139	Geographic information – Metadata – XML schema implementation
ISO 19157	Geographic information – Data quality
ISO 25012	Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Data quality model
ISO 25024	Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Measurement of data quality
ISO language codes	ISO 639 – Language codes

Appendix 3. Indicators

Table 4. “How well does information describe reality?” – The quality criteria and related indicators

Name	Description	Format	Assessment level	Quality criterion
Methodically produced values	The proportion of values produced methodically or using surrogate data to the total number of characteristic values in the dataset	Percentage	Characteristic	Correctness
Incorrect values	The proportion of target units with incorrect characteristic values to the total number of target units in the dataset	Percentage	Characteristic	Correctness
Misclassification	The proportion of target units with misclassified characteristic values to the total number of target units in the dataset	Percentage	Characteristic	Correctness
Standard deviation	Describes how spread out the characteristic values are relative to the mean	Standard deviation	Characteristic	Accuracy
Outliers	The proportion of outliers to the total number of target units in the dataset	Percentage	Characteristic	Accuracy
Logic of data reviewed	Logical conditions have been used in the collection, compilation or processing of the data	Yes/no	Characteristic and dataset	Consistency
Baseline period	The point in time to which the data apply	Period	Characteristic and dataset	Currentness
Creation period	The time of creation of the target unit or characteristic	Period	Characteristic and dataset	Currentness
Review period	The time of revision of the target unit or characteristic	Period	Characteristic and dataset	Currentness
Change period	The time of change of the target unit or characteristic	Period	Characteristic and dataset	Currentness

Temporal target coverage	**The intended temporal coverage and frequency of the dataset have been described**	**Yes/no**	**Dataset**	**Completeness**
Regional target coverage	The intended regional coverage and density of the dataset have been described	Yes/no	Dataset	Completeness
Target units	The dataset description clearly also specifies the other boundaries of the dataset in addition to the temporal and regional target coverage	Yes/no	Dataset	Completeness
Shortcomings in characteristics	The dataset under review are missing characteristics that are relevant to the phenomenon described by the dataset	Yes/no	Dataset	Completeness
Missing units	Describes the under-coverage in the dataset, i.e. the percentage of target units missing from the target population	Percentage	Dataset	Completeness
Additional units	Describes the over-coverage in the dataset, i.e. the percentage of target units present in the dataset that do not belong in the target population	Percentage	Dataset	Completeness
Incomplete units	The proportion of target units with even one missing characteristic to the total number of target units in the dataset	Percentage	Characteristic and dataset	Completeness
Incomplete characteristics	Proportion of target units that have missing characteristic values to the total number of target units with the characteristic in question	Percentage	Dataset	Completeness

Table 5. “How has the information been described?” – The quality criteria and related indicators

Name	Description	Format	Assessment level	Quality criterion
Data source	The source data for the dataset, target unit or characteristic are available	Percentage	Characteristic and dataset	Traceability
Data lifecycle	The data lifecycle is defined and described	Yes/no	Dataset	Traceability
Change management	Changes in the structural or source data of the characteristics are monitored	Yes/no	Characteristic	Traceability
Dataset descriptions	What language versions are available for the dataset description	Language versions	Dataset	Understandability
Definitions of concepts	What language versions are available for the definitions of concepts	Language versions	Dataset	Understandability
Data descriptions of characteristics	What language versions are available of the definitions of concepts	Language versions	Characteristic	Understandability
Customer feedback on comprehensibility	It is possible to provide feedback on understandability, and the feedback is monitored and followed up	Yes/no	Dataset	Understandability
Regulations and standards to be complied with	The regulations, standards, good practices and recommendations followed in the dataset are listed in the dataset description	Yes/partly/no	Characteristic and dataset	Compliance

Table 6. How can I use information? – The quality criteria and related indicators

Name	Description	Format	Assessment level	Quality criterion
The dataset data model	The dataset is in a structured format	Yes/no	Dataset	Portability
Permanent identifier of the target unit	The target units of the dataset have at least one dataset-specific permanent identifier	Yes/no	Dataset	Portability
Customer feedback on portability	It is possible to provide feedback on portability, and the feedback is monitored and followed up	Yes/no	Dataset	Portability
Access rights	The restrictions on access are described (e.g. processor, processing environment)	Access restrictions	Dataset	User rights
Restrictions on use	The restrictions on use are described (e.g. open data licences or terms of use)	Restrictions on use	Dataset	User rights
Compliance with due dates	The data delivery schedule is monitored against the actual delivery schedule	Delay	Dataset	Punctuality
Frequency of updates	The regular update frequency of the dataset	Written description	Characteristic and dataset	Punctuality
Values changed in the update	The proportion of values changed in the update to all the values in the dataset	Percentage	Characteristic	Punctuality

Statistics Finland

Figures for tomorrow