Project for Establishment of National Spatial Data Infrastructure (NSDI) for Bangladesh

Guidelines for Data Quality (DRAFT)

10 August 2020

Survey of Bangladesh, Ministry of Defence, People's Republic of Bangladesh
Working Group on NSDI for Bangladesh
Japan International Cooperation Agency

This document should be reviewed on a regular basis and updated by the NSDI-WG when necessary.

Revision history

Version	Date	Comments
1.0	xx August 2020	The first version

Foreword

National Spatial Data Infrastructure (NSDI) is a framework of efficient exchange and utilization of geographic information including legislation, rules, management and ICT systems.

Through the utilization of NSDI, we can expect the following three merits.

1. Removing duplication of data, investment, labour, etc.,
2. Raising efficiency of data interoperability contributing to the productivity of current work,
3. Incubating new businesses and projects.

For evaluation of the geographic information, we need to conduct the data quality evaluations. Then, we should share the result of the evaluation among the stakeholders, in order to discover the geographic data that are suitable for each user's objectives.

This document, the Guidelines for Data Quality, helps the data producers to prepare rules and instructions for data quality. The data users can also commonly understand the criteria and definitions on data quality.

For developing the guidelines, we used "Rules for Quality Requirements, Evaluation and Reporting" (by Geospatial Information Authority of Japan (GSI), <https://www.gsi.go.jp/common/000187337.pdf>, in Japanese), "Guidelines on NWRD Spatial Data Quality" (by WARPO, <http://old.warpo.gov.bd/guidlines/SpatiaDataQualityGuideline.pdf>), INSPIRE Data Specifications Template <https://inspire.ec.europa.eu/documents/data-specifications-template> and ISO 19157 (Data quality) as references. The first one is a part of "Japan Profile for Geographic Information Standards" (JPGIS) developed by GSI.

The Guidelines for Data Quality relate with the Guidelines for Data Product Specifications and the Guidelines for Metadata. These three guidelines have been developed for the "Project for Establishment of National Spatial Data Infrastructure (NSDI) for Bangladesh".

Table of Contents

Acronyms and Abbreviations

No.	Abbreviation	Official name
	DPS	Data Product Specifications
	GIS	Geographic Information System
	ICT	Information and Communication Technology
	ISO	International Organization for Standardization
	ISO/TC 211	International Organization for Standardization/Technical Committee 211 (Geographic information/Geomatics)
	JISC	Japanese Industrial Standards Committee
	JMP 2.0	Japan Metadata Profile 2.0
	JPGIS	Japan Profile for Geographic Information Standards
	MoD	Ministry of Defence
	NDC	National Datacenter
	NSDI	National Spatial Data Infrastructure
	NSDI-PF	National Spatial Data Infrastructure Platform
	NSDI-PP	National Spatial Data Infrastructure Pilot Project
	NSDI-PPWG	National Spatial Data Infrastructure Pilot Project Working Group
	NSDI-PS	National Spatial Data Infrastructure Prototype System
	NSDI-WG	National Spatial Data Infrastructure Working Group
	SoB	Survey of Bangladesh
	SoB BM	Survey of Bangladesh's Base Map
	TC	Technical Committee

CHAPTER 1. Basic Information of Data Quality

1.1. Overview

This document, the guidelines for data quality, provides principles for describing the quality for geographic data and concepts for handling quality information for geographic data, and a consistent and standard manner to determine and report the data quality information. It aims also to provide instructions for evaluating procedures of quality information for geographic data.

The quality of geospatial information is not the differences between the real world and the dataset. It represent a difference between the quality requirements described in the data product specifications and the actual dataset.

There are various kind of geographic data in the world. Each user may choose the most suitable data that match their objectives. In such occasions, the comparison of each data quality is essential for decision-making to select the data.

For this reason, it is important that the results of the data quality reports are expressed in a comparable way and that there is a common understanding of the data quality measures that have been used. These data quality measures provide descriptors of the quality of geographic data through comparison with the requirements of the user. The use of incompatible measures makes data quality comparisons impossible to perform.

1.2. Scope of the Guidelines

This document describes the principles for describing the quality of geographic data.

- components for describing data quality;
- components and content structure of a register for data quality measures;
- general procedures for evaluating the quality of geographic data;
- principles for reporting data quality.

This document also explains a set of data quality measures for use in evaluating and reporting

data quality. It is applicable to data producers providing quality information to describe and assess how well a data set conforms to its product specification and to data users attempting to determine whether specific geographic data are of sufficient quality for their particular application.

This document does not attempt to define minimum acceptable levels of quality for geographic data.

1.3. Terms and Definitions

1.3.1. accuracy

Closeness of agreement between a data quality result and the accepted reference value

1.3.2. conformance

Satisfying specified requirements

1.3.3. conformance quality level

Acceptable limit of data quality result

1.3.4. correctness

Correspondence with the data quality requirements

1.3.5. data product specification

Detailed description of a dataset together with additional information that will enable it to be created, supplied to and used by another party

1.3.6. data quality

Totality of characteristics of a product that bear on its ability to satisfy stated and implied needs

1.3.7. data quality date

The date or period for which a data quality measure is applied

1.3.8. data quality elements

Component describing a certain aspect of the data quality.

1.3.9. data quality evaluation procedure

Operations used in applying and reporting quality evaluation methods and their results

1.3.10. data quality measure

Evaluation of a data quality element

EXAMPLE: The percentage of the values of an attribute that are correct.

1.3.11. data quality result

Value or set of values resulting from applying a data quality measure or the outcome of evaluating the obtained value or set of values against a specified conformance quality level.

EXAMPLE: A data quality result of "90" with a data quality value type of "percentage" reported for the data quality element "commission" (excess item) is an example of a value resulting from applying a data quality measure to the data specified by a data quality scope. A data quality result of "true" with a data quality value type of "boolean variable" is an example of comparing the value (90) against a specified acceptable conformance quality level (85) and reporting an evaluation of a kind, pass or fail.

1.3.12. data quality scope

Extent or characteristic(s) of the data for which quality information is reported

NOTE: A data quality scope for a dataset can comprise a dataset series to which the dataset belongs, the dataset itself, or a smaller grouping of data located physically within the dataset sharing common characteristics. Common characteristics can be an identified feature type, feature attribute, or feature relationship; data collection criteria; original source; or a specified geographic or temporal extent.

1.3.13. data quality value type

Value type for reporting a data quality result

EXAMPLE: integer, real, boolean variable (true or false), string, date, percentage, ratio

NOTE: A data quality value type is always provided for a data quality result.

1.3.14. data quality value unit

Value unit for reporting a data quality result

EXAMPLE: percent

NOTE: A data quality value unit is provided only when applicable for a data quality result.

1.3.15. direct evaluation method

Method of evaluating the data quality of a dataset based on inspection of the items within the dataset

1.3.16. feature

Abstraction of real world geographic phenomena.

EXAMPLE: building, road, administrative boundary, district

1.3.17. feature attribute

Property of a feature. It has an attribute name, a value type and a value domain.

1.3.18. full inspection

Full inspection tests every item in the population specified by the data quality scope.

1.3.19. indirect evaluation method

Method of evaluating the data quality of a dataset based on external knowledge

NOTE: Examples of external knowledge are dataset lineage, such as production method or source data.

1.3.20. item

Any part of the data or dataset

1.3.21. lot

A collection of items produced under equal conditions for sampling

1.3.22. metadata

Information about a data

1.3.23. population

All of the concerned items in the data.

EXAMPLE: all points in a dataset, all of the street names in the district

1.3.24. reference data

Data adapted as true values

1.3.25. sampling

Taking a sample from the population

1.3.26. sample

One or more sampling units taken from the population to obtain information about the population

CHAPTER 2. Principles of Data Quality

2.1. Basic Concept of Geospatial Data Quality

The quality of geospatial data is the difference between the data and the quality requirements described in the data product specifications, rather than the one between the data and the real world.

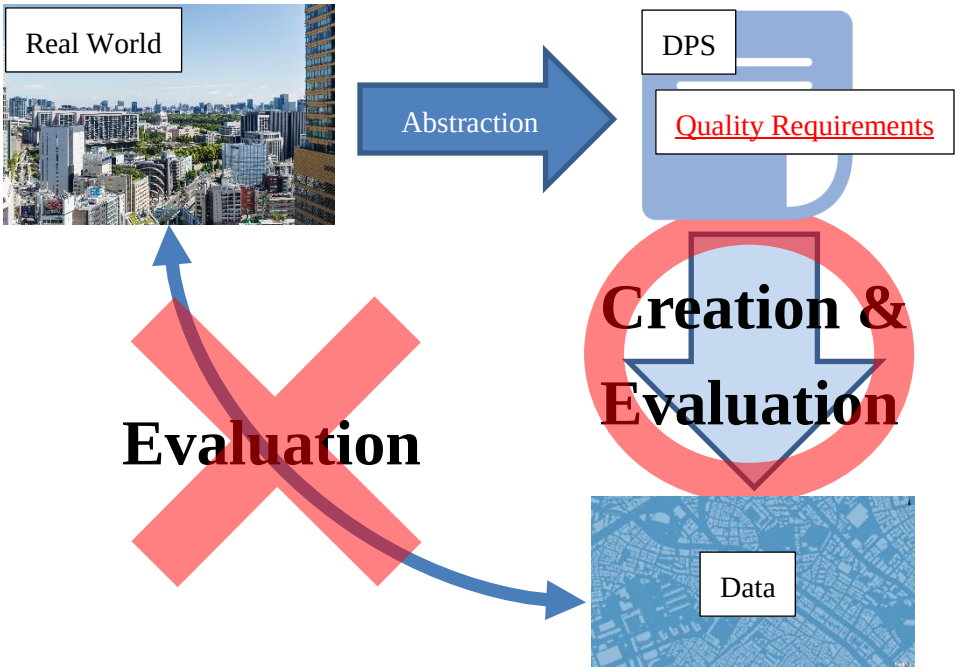

Figure 2-1 Relationship between the real world, DPS and data for quality evaluation

NOTE: Data quality information should be shared among all stakeholders for interoperability.

2.2. Data Quality Elements

2.2.1. Quantitative data quality elements

There are fifteen quantitative elements belonging to five categories for data quality description. Only necessary elements are selected to use for data quality descriptions.

Table 2-1 Quantitative data quality elements

#	Elements	Categories	Definitions
1	commission	Completeness	excess items present in the dataset
2	omission		missing items from the dataset
3	conceptual consistency	Logical consistency	adherence to rules of the conceptual schema
4	domain consistency		adherence of values to the value domains
5	format consistency		degree to which data are stored in accordance with the physical structure of the data set
6	topological consistency		correctness of the explicitly encoded topological characteristics of a data set
7	absolute or external accuracy	Positional accuracy	closeness of reported coordinate values to values accepted as or being true
8	relative or internal accuracy		closeness of the relative positions of features in a data set to their respective relative positions accepted as or being true
9	gridded data positional accuracy		closeness of gridded data spatial position values to values accepted as or being true
10	classification correctness	Thematic accuracy	comparison of the classes assigned to features or their attributes to the requirements in the data product specifications
11	non-quantitative attribute correctness		measure of whether a non-quantitative attribute is correct or incorrect
12	quantitative attribute accuracy		closeness of the value of a quantitative attribute to a value accepted as or known to be true
13	accuracy of a time measurement	Temporal quality	closeness of reported time measurements to values accepted as or known to be true

#	Elements	Categories	Definitions
14	temporal consistency		correctness of the order of events
15	temporal validity		validity of data with respect to time

2.2.2. Non-quantitative data quality elements

There are three non-quantitative elements for data quality. Their descriptions are in free text.

Table 2-2 Non-quantitative data quality elements

#	Elements	Definitions
1	purpose	the rationale for creating a data set and contains information about its intended use, which may not be the same as the actual use of the data set
2	usage	the application for which a data set has been used, either by the data producer or by other data users.
3	lineage	the history of a data set and recounts the life cycle of a data set from collection and acquisition through compilation and derivation to its current form

NOTE: The lineage is important to show the relationship with other data clearly.

EXAMPLE: Lineage “The geographical name has appended from BBS and verified through the field survey in 2011-2012.”

2.3. Descriptions of Data Quality Eements

2.3.1. Description of data quality requirements

For data quality requirements, the data quality scope and data quality measures should be described in the DPS and the metadata.

2.3.2. Description of data quality evaluations

For data quality evaluations, the data quality evaluation procedures should be described in the DPS and the metadata.

2.3.3. Description of data quality reporting

For reporting the evaluation results of data quality, the data quality results should be described in the metadata. The description may be quantitatively (number, rate, etc.), conformance to the criteria (true / false), or free text.

2.4. Procedures of data quality requirements and evaluations

2.4.1. Steps of data quality requirements and evaluation

1) create the specifications including the data quality requirements
2) create the evaluation procedures
3) conduct the quality evaluation
4) report the data quality results

2.4.2. Classification of evaluation methods

The method of data quality evaluation is generally separated into the direct evaluation method and indirect evaluation method. Then, the direct evaluation method is also separated into full inspection methods and sampling methods. See the figure below.

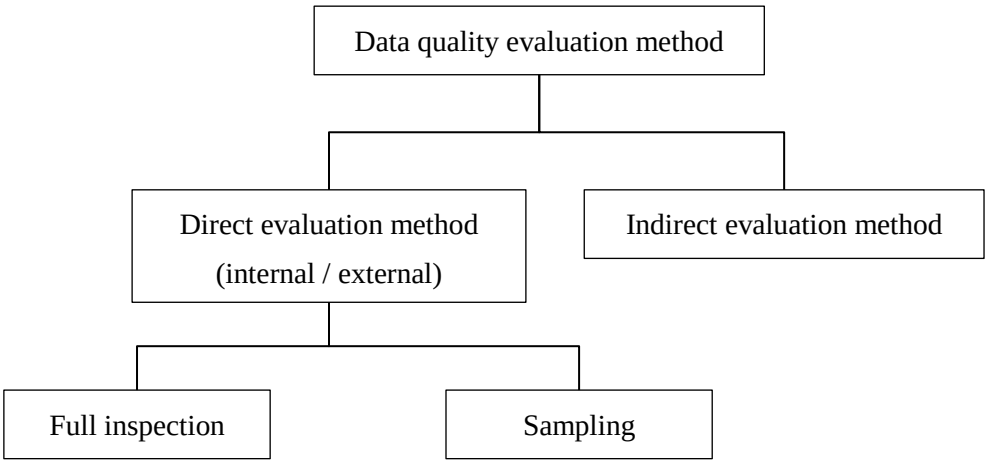

Figure 2-2 Classification of data quality evaluation method

2.4.3. Direct evaluation method

The direct evaluation method may refer internal data only or external reference data for compare. Both internal / external ways may apply to the full inspection and sampling.

2.4.3.1. Full inspection

Full inspection requires testing of all items in the population as defined by the data quality scope.

The table below shows the procedures of full inspection.

Table 2-3 Procedures of full inspection

#	procedures	descriptions
1	Definition of items	An item is a minimum target of the inspection. An item may be a feature, a feature attribute, and a feature relationship.
2	Inspection of items in the data quality scope	Inspect all items in the data quality scope

NOTE: Full inspection is most appropriate for tests on small populations or for tests that can be completed by automatic means.

2.4.3.2. Sampling

Sampling tests a subset of items (sample) in the population defined by the data quality scope. Each sample should be sufficient to obtain the suitable data quality results. The table below shows the procedures of sampling.

Table 2-4 Procedures of sampling

#	procedures	descriptions
1	Definition of sampling method	Definite how to extract items from the populations to obtain the proper data quality results.
2	Definition of items	An item is a minimum target of the inspection. An item may be a feature, a feature attribute, and a feature relationship.

#	procedures	descriptions
3	Division of the population into lots	A lot shall be a subset of items in the data quality scope. Each lot should consist of items created under the same conditions and at the same time as far as possible.
4	Division of lots into sampling units	The sampling unit shall be the area of the lot to be inspected.
5	Definition of sample rate or sample size	The sample rate gives information on how many items are extracted from each lot for inspection on average.
6	Selection of sampling unit	Select the number of sampling units required to meet the sample rate or sample size of the item.
7	Inspection of items in the sampling units	Inspect all items in the sampling unit.

Next table is an example of sampling and visual inspection.

Table 2-5 EXAMPLE: sampling and visual inspection of road

contents	descriptions
data quality scope	road
data quality measure	Definition of error: 1) the place which exists in the data but not in the reference data, 2) the place which does not exist in the data but exists in the reference data, Definition of the error submesh: submesh including one or more errors. Definition of the error rate: error rate (%) = number of error submesh / total number of submesh * 100
data quality evaluation method	Sampling and visual inspection 1) identify the inspection units 2) divide the mesh in the inspection unit into 10 by 10 submeshes 3) display all the items in the inspection unit

contents	descriptions
	4) check the items in each submesh visually comparing with reference data (aerial photo, field survey results, etc.) to identify the error submesh which include any mismatch. 5) Calculate the error rate based on the results above.
conformance quality level	pass: the error rate of all inspection unit is less than or equal to 5% fail: at least one inspection unit has the error rate beyond 5%.

2.4.4. Indirect evaluation method

The indirect evaluation method is a method of quality evaluation of a dataset based on external knowledge or experience. External knowledge may include one or more non-quantitative quality information or other quality evaluation reports on the resource dataset used in creating the dataset.

When reporting the indirectly evaluated data quality, it should include an explanation of how it was determined.

The indirectly evaluated data quality should not be reported as quantitative results. The indirectly evaluated data quality may be described in text using the descriptive results.

NOTE: This method should be used when the direct evaluation method is not available.

NOTE: non-quantitative quality information are purpose, usage and lineage.

CHAPTER 3. General Rules of Data Quality Requirements

3.1. Overview

The data quality requirements of the data product specifications provide criteria for determining the suitability of data for its intended use.

3.2. Description of data quality requirements

The data quality requirements for the data product specifications shall include the following

- data quality scope

- data quality measures (including conformance quality levels)

3.2.1. Data quality scope

The data quality scope defines the spatial extent, temporal extent, or common characteristics that identify the geospatial data for data quality evaluation.

Each data quality element needs at least one data quality scope.

The data quality scope shall be the dataset series to which this dataset belongs, this dataset, or the collection of items (lots) with common characteristics that are contained within this dataset.

A data quality report in metadata may include more than one data quality element and its result.

3.2.2. Data quality measures

Each data quality element needs one or more data quality measures. The data quality measures shall briefly describe the name and type of inspection. A data quality measure shall also include a conformance quality level.

CHAPTER 4. General Rules of Data Quality Evaluations

4.1. Overview

The data quality evaluation procedures shall specify a data quality evaluation method for each data quality element. The data quality evaluation method describes a procedure for applying the data quality measure to the items in the data quality scope.

4.2. Evaluation of Data Quality Elements

4.2.1. Evaluation of commission (excess item)

The quality evaluation of commission shall fulfill the following requirements.

1) The items to be inspected are clearly defined.
2) The number of the items in the reference data are clearly known.
3) If the items have identifiers, a one-to-one comparison is made between the items in the data and the reference data, and the number of correspondences or difference is counted. In the case where there is no identifier, calculate the difference between the number of items in the data set and the reference data.
4) The data quality result is the ratio, or the number of excess items and the number of reference data.

4.2.2. Evaluation of omission (missing item)

The quality evaluation of omission shall fulfill the following requirements.

1) The items to be inspected are clearly defined.
2) The number of the items in the reference data are clearly known.
3) If the items have identifiers, a one-to-one comparison is made between the items in the data and the reference data, and the number of correspondences or difference is counted. In the case where there is no identifier, calculate the difference between the number of items in the data set and the reference data.
4) The data quality result is the ratio, or the number of missing items and the number of reference data.

4.2.3. Evaluation of conceptual consistency

This evaluation is omitted because any conceptual schema is not used in this document.

NOTE: This is a future challenge to compliance with international standards.

4.2.4. Evaluation of domain consistency

The quality evaluation of domain consistency shall fulfill the following requirements.

1) The value domain of the feature attribute is specified in the data product specifications.
2) Inspect the values of the feature attribute are conforming to the value domain.
3) The data quality result is the ratio, or the number of errors and the number of inspected items.

4.2.5. Evaluation of format consistency

The quality evaluation of format consistency shall fulfill the following requirements.

1) Inspect the conformance of the format of data to the format specified in the data product specifications.
2) The data quality result is the ratio, or the number of errors and the number of inspected items.

4.2.6. Evaluation of topological consistency

The quality evaluation of topological consistency shall fulfill the following requirements.

1) The consistency of the topological attributes and the geometric attribute including topology specified in the data product specifications shall be inspected.
2) To check the consistency of relationships between features (inclusion, connection, overlap, edge matching, etc.) which specified in the data product specifications.
3) The data quality result is the ratio, or the number of errors and the number of inspected items.

4.2.7. Evaluation of absolute or external accuracy

The quality evaluation of absolute or external accuracy shall fulfill the following requirements.

1) The values adapted as true or true values of the absolute position of the target feature (reference data) shall be known.
2) The coordinate values reported as absolute positions are compared with the values of the reference data. Then the error is calculated.
3) The result of the data quality evaluation shall be the degree of error and/or uncertainty in measured values, etc.

NOTE: Absolute position is the position on the geocentric coordinate system or the position that can be converted to it.

4.2.8. Evaluation of relative or internal accuracy

The quality evaluation of relative or internal accuracy shall fulfill the following requirements.

1) The values adopted as true or true values of the relative position (relative distance or coordinate values) between the arbitrarily determined origin and the point to be evaluated

(reference data) shall be known.

2) The error is determined by comparing the relative position of the point and the reference data.
3) The data quality result shall be the degree of uncertainty of the error.

4.2.9. Evaluation of gridded data positional accuracy

The quality evaluation of gridded data positional accuracy shall fulfill the following requirements.

1) The adopted or true coordinates of the grid point (reference data) shall be known.
2) The error between the reported coordinates of the grid point and the reference data are determined through comparing the both values.
3) The data quality result shall be the degree of uncertainty of the error.

4.2.10. Evaluation of classification correctness

The quality evaluation of classification correctness shall fulfill the following requirements.

1) There shall be the reference data indicating the feature type of the item.
2) Inspection whether the feature type of the item is correctly classified through comparing it with the reference data.

   EXAMPLE: When an item of the road type in the data is identified as the river type in the reference data, this item is incorrect. The item shall be a river type.
3) The data quality result is the ratio, or the number of errors and the number of inspected items.

4.2.11. Evaluation of non-quantitative attribute correctness

The quality evaluation of non-quantitative attribute correctness shall fulfill the following requirements.

1) There shall be the reference data indicating the values adapted or considered as true for the non-quantitative attributes.
2) Inspection through comparing the values of the non-quantitative attributes in the data with the reference data.

   EXAMPLE: If the name of a road in the data differs from the name of the road in the road register, this item is incorrect.
3) The data quality result is the ratio, or the number of errors and the number of inspected items.

4.2.12. Evaluation of quantitative attribute accuracy

The quality evaluation of quantitative attribute accuracy shall fulfill the following requirements.

1) There shall be the reference data indicating the values adapted or considered as true for the

quantitative attributes.

2) The error is determined through comparing the values of the quantitative attributes in the data with the reference data.

   EXAMPLE: The difference between the area of the land in the data and the area registered in the cadaster system.

3) The data quality result shall be the degree of uncertainty of the error.

4.2.13. Evaluation of accuracy of a time measurement

The quality evaluation of accuracy of a time measurement shall fulfill the following requirements.

1) The time adopted or considered as true (reference data) shall be known.
2) The error is determined thorough comparing the measured time with the reference data.
3) The data quality result shall be the degree of uncertainty of the error.

4.2.14. Evaluation of temporal consistency

The quality evaluation of temporal consistency shall fulfill the following requirements.

1) The time sequence of the events for evaluation shall be known.
2) Inspect the inconsistencies in the time sequence of the events.
3) The data quality result is the ratio, or the number of errors and the number of inspected items.

4.2.15. Evaluation of temporal validity

The quality evaluation of temporal validity shall fulfill the following requirements.

1) The transaction time of the item for evaluation is known.

   EXAMPLE: A time point when the item is registered in the database.

2) The error is determined through comparing the measured transaction time with the transaction time adapted as true.

   EXAMPLE: When an observatory is required to register the observed results into a database, the error is the difference between the observatory's clock (the time stamp of the data) and the reference clock.

3) The data quality result shall be the degree of uncertainty of the error.

CHAPTER 5. General Rules of Data Quality Reporting

The data quality evaluation results should be shared through the metadata. The reporting contents of data quality results are shown in the table below.

Table 5-1 Reporting contents of data quality results as metadata

#	Reporting contents	Example
1	name of data quality element	omission (absent item)
2	data quality scope	hospitals in Bangladesh, as of 1st June 2020
3	data quality measure	Number of missing items compared to the original source (hospital list, as of 1st June 2020, by Ministry of Health)
4	conformance quality level	0 = pass, 1 or more = fail (no missing)
5	data quality evaluation method	Count the number of missing items using a software function
6	data quality result	0 (pass)
7	data quality value type	Integer
8	data quality value unit	-
9	data quality date	2020/06//01

Annex

Example of the data quality requirements, evaluations and reporting for NSDI-PF