**Recommendations on the data quality evaluation of S-100 products**

**Data Quality Working Group**

**Document Control**

|  |  |  |  |
| --- | --- | --- | --- |
| **Edition Number** | **Date** | **Author** | **Notes** |
| **0.1** | **March 2024** | **DQWG** | **Drafted based on ISO 19157-2015** |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Content**

[1. When to use data quality evaluation procedures 1](#_Toc27331)

[2. How to implement the data quality evaluation procedures 1](#_Toc31119)

[2.1 Ordering in data quality evaluation 1](#_Toc23700)

[2.2 Data quality evaluation methods 3](#_Toc4216)

[2.2.1 Classification of data quality evaluation methods 3](#_Toc31838)

[2.2.2 Direct evaluation 3](#_Toc1705)

[2.2.3 Indirect evaluation 3](#_Toc4036)

[2.3 Guidelines for the use of data quality elements 3](#_Toc1709)

[2.3.1 The relationships between the data quality elements 4](#_Toc3252)

[2.3.1.1 General 4](#_Toc5639)

[2.3.1.2 Data quality elements related to missing attribute values 4](#_Toc21705)

[2.3.1.3 Relationships between the different aspects of accuracy 4](#_Toc16766)

[2.3.1.4 Dependency between completeness and accuracy 4](#_Toc5345)

[2.3.2 Data quality elements-example of use 5](#_Toc20710)

[2.3.2.1 Completeness 5](#_Toc12863)

[2.3.2.1.1 General 5](#_Toc26153)

[2.3.2.1.2 Commission - excess data present in a data set 5](#_Toc14840)

[2.3.2.1.3 Omission - data absent from a data set 6](#_Toc18842)

[2.3.2.2 Logical consistency 6](#_Toc11379)

[2.3.2.2.1 General 6](#_Toc30169)

[2.3.2.2.2 Conceptual consistency - adherence to rules of the conceptual schema 6](#_Toc3655)

[2.3.2.2.3 Domain consistency - adherence of values to the value domains 6](#_Toc18379)

[2.3.2.2.4 Format consistency - degree to which data are stored in accordance with the physical structure of the data set 6](#_Toc12485)

[2.3.2.2.5 Topological consistency - correctness of the explicitly encoded topological characteristics of a data set 7](#_Toc7034)

[2.3.2.3 Positional accuracy 7](#_Toc30685)

[2.3.2.4 Temporal quality 7](#_Toc9437)

[2.3.2.4.1 Accuracy of a time measurement - closeness of reported time measurements to values accepted as or known to be true 7](#_Toc20561)

[2.3.2.4.2 Temporal consistency - correctness of the order of events 8](#_Toc3020)

[2.3.2.4.3 Temporal validity - validity of data with respect to time 8](#_Toc24650)

[2.3.2.5 Thematic accuracy 8](#_Toc19632)

[2.3.2.5.1 General 8](#_Toc3534)

[2.3.2.5.2 Classification correctness - comparison of the classes assigned to features or their at- tributes to a universe of discourse (e.g. ground truth or reference data set) 8](#_Toc11190)

[2.3.3 Discussions on difficult cases 8](#_Toc4518)

[2.3.3.1 Relation between misclassification and completeness at feature type level 8](#_Toc13962)

[2.3.3.2 Quality elements related to unique identifiers 9](#_Toc17077)

[2.4 Aggregation of data quality results 9](#_Toc28894)

[2.4.1 Introduction 9](#_Toc25813)

[2.4.2 100 % pass/fail 9](#_Toc27217)

[2.4.3 Weighted pass/fail 10](#_Toc24486)

[2.4.4 Maximum/minimum value 10](#_Toc12406)

[3 Data quality reporting 10](#_Toc9480)

[3.1 General 10](#_Toc26110)

[3.2 Why report data quality 10](#_Toc3081)

[3.3 When to report quality information 10](#_Toc20569)

[3.4 How to report quality information 11](#_Toc19591)

[3.4.1 General 11](#_Toc30542)

[3.4.2 Reporting quality information as metadata 12](#_Toc26865)

[3.4.3 Reporting quality information within a standalone quality report 12](#_Toc3670)

[3.4.4 Particular cases 12](#_Toc32431)

[3.4.4.1 Reporting aggregation (aggregated results) 12](#_Toc23994)

[3.4.4.2 Reporting derivation (derived results) 13](#_Toc31557)

[3.4.4.3 Reference to the original data quality result 13](#_Toc20939)

[Annex 14](#_Toc25398)

**Recommendations on the data quality evaluation of S-100 products**

**1. When to use data quality evaluation procedures**

Quality evaluation procedures may be used in different phases of a S-100 product's life cycle. The stages of a S-100 product's lifecycle during which data quality evaluation may be applied are as follows:

1. Development of a data product specification or user requirements: When developing a data product specification or defining user requirements, quality evaluation procedures may be used to facilitate the establishment of conformance quality levels that should be met by the final product. A data product specification or user requirements may include conformance quality levels for the data and quality evaluation procedures to be applied during production and updating.
2. Quality control during data set creation: At the production stage, the producer may apply quality evaluation procedures, either explicitly established or not contained in the data product specification, as part of the process of quality control. The description of the applied quality evaluation procedures, when used for production quality control, may be reported as lineage metadata including, but not necessarily limited to, the quality evaluation procedures applied, conformance quality levels established and the results.
3. Inspection for conformance to a data product specification: On completion of the production, a quality evaluation process may be used to produce and report data quality results. These results may be used to determine whether a data set conforms to its data product specification or not. If the data set passes inspection (composed of a set of quality evaluation procedures), the data set is considered to be ready for use. The results of the inspection operation should be reported in accordance with Chapter 3. See also the example in Annex describing evaluation and reporting of data quality. The outcome of the inspection will be either acceptance or rejection of the data set. If the data set is rejected, then, after the data have been corrected, a new inspection will be required before the product can be deemed to be in conformance with the data product specification.
4. Evaluation of data set conformance to user requirements: Data Quality evaluation procedures may be used to establish if a data set meets the conformance quality levels specified in user requirements. Indirect as well as direct methods may be used in analyses of data set conformance to user requirements.
5. Quality control during data set update: Data Quality evaluation procedures are applied to data set update operations, both to the items being used for update and to benchmark the quality of the data set after an update has occurred.

**2. How to implement the data quality evaluation procedures**

**2.1 Ordering in data quality evaluation**

When evaluating a S-100 product, one individual error may influence several data quality elements. For measurements resulting in rates (e.g. percentage rates of aspects of completeness) the use of proper denominators describing the total population is important, see Figure 1.

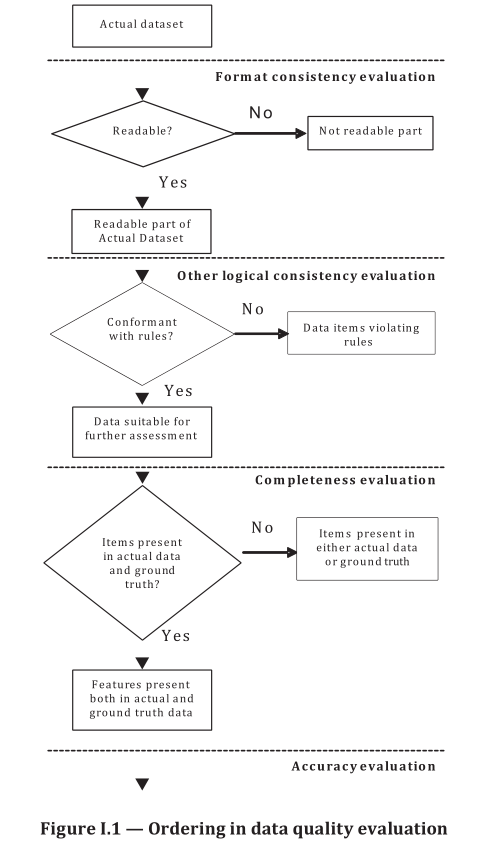


Figure 1 – Ordering in data quality evaluation

When evaluating data quality, the usual ordering is:

1. Logical consistency/Format consistency: The very first to be evaluated is the readability (or interpretability) of the data to decide whether it is possible to decode/read/understand the data or not. Not interpretable data should be reported and ignored in the further evaluation. The result of the format consistency should describe which parts of the data are not readable.
2. Logical consistency: Decide if the rules set up for the data set are followed. Parts of the data set not conforming to the rules should be ignored in the further evaluation.
3. Completeness: The next step in the evaluation is the feature existence aspect covered by completeness. To evaluate this, the features in the actual data set and the ground truth data are compared, and commissions and omissions reported.
4. Accuracy (positional, thematic and temporal aspects): The last step in the evaluation covers the accuracy aspect, measuring the deviation between actual and ground truth feature properties. These measurements can be based only on parts of the data set present in both the actual data set and the universe of discourse.

**2.2 Data quality evaluation methods**

**2.2.1 Classification of data quality evaluation methods**

A data quality evaluation procedure comprises one or more data quality evaluation methods. Data quality evaluation methods can be divided into two main classes: direct and indirect. Direct evaluation methods determine data quality through the comparison of the data with internal and/or external reference information. Indirect evaluation methods infer or estimate data quality using information on the data such as lineage. Direct evaluation methods should be used in preference to indirect evaluations. The direct evaluation methods are further sub classified by the source of the information needed to perform the evaluation, if internal or external.

**2.2.2 Direct evaluation**

A direct evaluation method is a method of evaluating the quality of a data set based on inspection of the items within the data set.

The direct evaluation methods can be classified as internal or external. Internal direct data quality evaluation uses only data that resides in the data set being evaluated. External direct quality evaluation requires reference data external to the data set being tested.

**NOTE1**: Reference data are data accepted as representing the universe of discourse.

For both external and internal evaluation methods, one of the following inspection methods may be used:

- Full inspection;

- Sampling.

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

**NOTE2**: Full inspection is most appropriate for small populations or for tests that can be accomplished by automated means. Sampling means that tests are performed on subsets of the S-100 data defined by the data quality scope.

**2.2.3 Indirect evaluation**

An indirect evaluation method is a method of evaluating the quality of a data set based on external knowledge or experience of the data product and can be subjective.

This external knowledge may include, but is not limited to, one or more non-quantitative quality information usage, lineage and purpose or other data quality reports on the data set or data used to produce the data set. Data quality may be estimated, for example, from knowledge about the source, tools and methods used for the capturing of the data and evaluated against procedures and specifications worked out for this product. Indirectly evaluated data quality may also be based on experience alone. If indirectly evaluated data quality has been reported, it should be accompanied by a description on how it was determined.

In some cases it might be misleading or not even possible to report indirectly evaluated data quality as quantitative results. In those cases the data quality may be described in textual form using a descriptive result.

**EXAMPLE:** The relative positional accuracy is higher between a geological feature and a nearby feature the absolute positional accuracy on the geological feature itself.

**2.3 Guidelines for the use of data quality elements**

In some cases, there may be several possible quality elements for one specific quality requirement and one detected error in a quality evaluation. This section provides guidelines for which quality element to use.

**2.3.1 The relationships between the data quality elements**

**2.3.1.1 General**

Many data quality elements are related to each other. In some cases this may lead to uncertainty about how identified deviations/errors in the data should be reported. This section discusses the relationship between the data quality elements.

**2.3.1.2 Data quality elements related to missing attribute values**

At least three different values should be considered to indicate "no value available". The way these three are used may influence the data quality element selected for reporting the missing value. The three values have different semantics:

-The empty value. In this case, the attribute has no value at all;

-The not applicable value. This indicates that for this specific feature the attribute is not valid, i.e. have no meaning;

-The unknown value. In this case, the attribute is valid i.e. there should have been a value, but the value is not known.

Mandatory attributes with empty values should be reported as logical consistency errors. Not applicable mandatory attributes should not be counted when evaluating attribute completeness. The amount of unknown occurrences should be reported as attribute completeness.

A way of increasing the attribute completeness is to add artificial values to a data set. By doing so, the data set will become better from an attribute consistency point-of-view, but the attribute accuracy will decrease.

**2.3.1.3 Relationships between the different aspects of accuracy**

Deviations of actual data from the universe of discourse can be measured using positional accuracy, time (temporal) accuracy and attribute (thematic) accuracy. Examples of alternative ways of expressing the deviation are:

-Attribute versus space: For attributes where the geographical distribution is known, a deviation can be expressed either by the theme or the positional component. The height value of a contour line can be considered as an attribute of the contour line. The deviation of the current position from the true position can be measured either by the attribute component ("half a meter too high") or by the space component ("the contour line has an offset of 10 m in north direction").

-Space versus time: If the movement of a feature is known, a difference between measured and real position can be expressed either by the time component or by the positional component.

-Attribute versus time: For attributes with known temporal variations, deviations can be represented by the theme or temporal component. The height value of the water level can be considered as an attribute of the water level. The deviation between measured data and actual data can be represented by attribute component ("water level is one meter higher") or time component ("this data is 10 minutes late in time").

**2.3.1.4 Dependency between completeness and accuracy**

Evaluation of completeness usually is based on comparison of the data set and the universe of discourse. The critical operation is the linking between features in the data set and the universe of discourse. When a unique identifier exists the linking is usually based on this.

When handling features without this kind of identification of the individuals, methods based on closeness of attributes and attribute values have to be used. When linking geographical features two aspects have to be considered:

1. the thematic closeness (usually expressed as feature type);
2. the geographical closeness of the features.

When two features (a pair with one in the data set and the other in the ground truth) are decided to be representations of the same real-world phenomenon, the deviations between the two are handled as accuracy. If the pair of features is decided to represent different phenomena, the deviation between the two is reported using completeness (omission and/or commission).

For example when evaluating completeness and accuracy for feature type 1, see Figure 2, there is no problem in positions A.B.C and D. Here the classification is identical (thematic deviation equal to zero) and the geographical deviations between actual and real position are within the accepted level. The features are linked, and the deviations are described by positional accuracy. In position E, the two instances have different thematic classifications but are located very close to each other. A decision has to be made whether the difference in classification is within the level of acceptance for linking. If yes, the two instances will contribute to the accuracy evaluation (positional and/or thematic). It not it is a question of completeness (one point missing and one in excess). In positions F and G, the two instances have the same classification, but differ in position. If this geographical deviation is considered to be within the level of acceptance for linking, the deviation will contribute to positional accuracy (probably an outlier), if not it is a question of completeness (omission and commission).

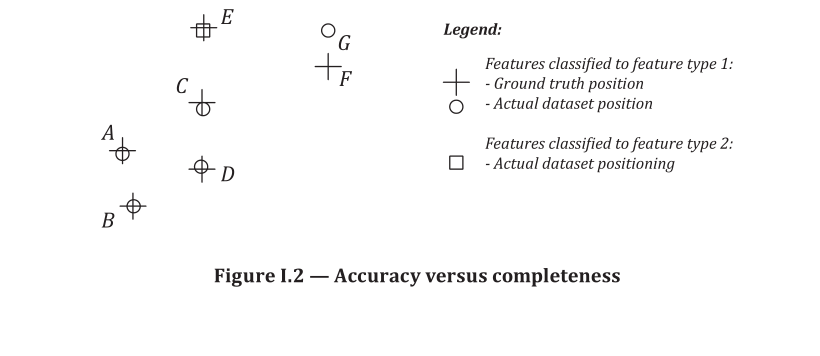


Figure 2 – Accuracy versus completeness

**2.3.2 Data quality elements-example of use**

**2.3.2.1 Completeness**

**2.3.2.1.1 General**

The presence and absence of features may be described by the data quality elements commission and omission. Completeness should mainly be used on the feature type level, describing whether the features in the universe of discourse are found in the data set or not.

Completeness may also be relevant for feature properties ("attribute completeness" and “relationship completeness"). Before using completeness for this, the logical consistency/conceptual consistency should be carefully considered.

**2.3.2.1.2 Commission - excess data present in a data set**

This may be applied at the feature instance level. This means that data are considered to be in "excess" if it is a whole feature instance. If there is non-required data within a feature instance or attribute of a feature instance then this is not considered commission.

This definition incorporates feature instances which are present in the data set but which are not within the scope (as defined in the specification).

**2.3.2.1.3 Omission - data absent from a data set**

Similarly to commission, this may be applied at the feature instance level. In practice this refers to the absence of feature instances whose inclusion is specified in the specification.

Omission should mainly be used when a "whole item", e.g. a feature instance is missing. If a mandatory part of an item, e.g. a mandatory attribute of a feature instance, is missing, this should be reported as a conceptual consistency error

**2.3.2.2 Logical consistency**

**2.3.2.2.1 General**

The degree of adherence to logical rules of data structure, attribution and relationships (data structure can be conceptual, logical or physical) may be described by the following data quality elements.

**2.3.2.2.2 Conceptual consistency - adherence to rules of the conceptual schema**

Applications usually have a conceptual schema describing the requirements to the data structure. This conceptual schema may include:

1. the name of all classes (feature types, data types, etc.),
2. the attribute names for all classes, and also the multiplicity limitations,
3. the domains for all attributes,
4. the relationships between the classes,
5. the topological relationships between feature types, e.g., the relationship between an area and the border lines.
6. the relationship between feature type attributes for different feature types, e.g. the relationship between the height-above-sea value from a contour line and the same from a road in the geographical crossing point for the two feature instances.

Conceptual consistency may cover all these aspects of data quality.

Others logical consistency elements(domain consistency, topological consistency) may also be considered for some of the aspects listed above if conceptual consistency is used only to ensure that the correct feature properties are present for each feature instance,

**2.3.2.2.3 Domain consistency - adherence of values to the value domains**

Domains of values are usually described by the conceptual schema of the application, and may be reported as part of the conceptual consistency or as domain consistency. If the domain definitions are not existing or not valid in the conceptual schema then only the quality element domain consistency can be used.

**EXAMPLE:** An organization defines the valid value domains for each field in terms of length, data type and content. Domain consistency is used to ensure compliance to these conditions with the following exceptions:

* Where the field contains position data, in which case it is considered as positional accuracy;
* Where the field contains date/time data, in which case it is considered as temporal quality;
* Where the field contains a primary key, in which case it is considered under logical consistency.

**2.3.2.2.4 Format consistency - degree to which data are stored in accordance with the physical structure of the data set**

Format consistency should mainly be used as the first quality evaluation testing whether the data set is in the correct format according to the (product) specification.

If certain rules are defined for defining the format of specific attributes, e.g. for generated IDs, format consistency can also be relevant for single attribute values. If attributes values are checked compared to a list of legal values (a domain), the domain consistency should be used.

**EXAMPLE:** The data product specification of a product specifies GML as the distribution format. If the data set is not a GML file, then this error should be reported as format consistency error. If one single item in the GML file is "in wrong format", e.g. text instead of number, this can be reported as conceptual consistency error or domain consistency error.

**2.3.2.2.5 Topological consistency - correctness of the explicitly encoded topological characteristics of a data set**

Topological characteristics of the data set describe the geometric relationships between data set items unchanged by "rubber-sheet transformations". The main parts of the topological constraints are supposed to be described in the conceptual schema, and may be reported as conceptual consistency or topological consistency. In the case when the relevant topological requirements are not part of the conceptual schema, only topological consistency could be used.

**EXAMPLE1**: For a data set with feature types defined to be located on the shoreline of water bodies (feature types like shore line, harbor, boathouse), and also feature types for water bodies (lakes, seas, etc.).The topological relationships between the feature types are well defined in the conceptual schema, and the quality element conceptual consistency is used to report whether shorelines (1dimension) geometry coincide with the water body (2 dimensions) geometry.

**EXAMPLE2:** In a network data set, with vague requirement in the conceptual schema for a "clean network", the "dirty parts" (undershoot, overshoot, overlapping, self-intersecting, etc.) should be reported as topological consistency errors.

**2.3.2.3 Positional accuracy**

Accuracy of the position of features may be described using the data quality elements in this section.

Measuring positional accuracy using ground truth implies establishing "correspondence pairs" with one feature instance from the data set and the corresponding one in the control (ground truth) data set. If the features have unique identifiers this correspondence can be set up using the identifiers, and gross errors, bias, standard deviation can be estimated and reported as positional accuracy.

With no available identifiers the correspondence has to be established using the positions. A “correspondence distance limit" shall be defined. This makes it impossible to compute gross errors. This “correspondence distance limit" shall be documented in the report. In this case:

* the feature instances in the data set with no corresponding control data set feature instance should be reported as completeness/commission,
* the control data set feature instances with no corresponding data set feature instance should be reported as completeness/omission.

**2.3.2.4 Temporal quality**

**2.3.2.4.1 Accuracy of a time measurement - closeness of reported time measurements to values accepted as or known to be true**

**EXAMPLE:** Within a certain organization accuracy of a time measurements is used to ensure that:

* the value does not contravene a specific condition imposed on the field (over and above the conditions imposed by the nature of date/time data).

Example rule: The START\_DATE field cannot contain a value in the future.

**2.3.2.4.2 Temporal consistency - correctness of the order of events**

The rules describing the "correctness of the order of events" may be part of the conceptual schema. It might be reported either as temporal consistency or as conceptual consistency if the rules are part of the conceptual schema.

**EXAMPLE:** Within a certain organization temporal consistency is used to:

* Confirm the consistency between date/time values relating to the lifecycle of the real-world object,
* Ensure the consistency of date/time values used in the management of the feature instances in the data set.

Example rule: The END\_DATE shall be the same as or after START\_DATE.

Temporal consistency error example: START\_DATE ="2010-02-02", END\_DATE="2000-01-01".

**2.3.2.4.3 Temporal validity - validity of data with respect to time**

The rules describing the "validity of data with respect to time" may be part of the conceptual schema. It might be reported either as temporal validity or as conceptual consistency if the rules are part of the conceptual schema.

**EXAMPLE:** Within a certain organization accuracy of a time measurements is used to:

- ensure that the content of a date or time field is in the correct format and uses the calendar defined in the specification.

Example rule: The date value shall be in ISO 8601 format - "CCYY-MM-DD".

Temporal validity error example:"01-01-2010"or“2010-51-15".

**2.3.2.5 Thematic accuracy**

**2.3.2.5.1 General**

The accuracy of quantitative attributes and the correctness of non-quantitative attributes and of the classifications of features and their relationships may be described using the following data quality elements.

**2.3.2.5.2 Classification correctness - comparison of the classes assigned to features or their at- tributes to a universe of discourse (e.g. ground truth or reference data set)**

**EXAMPLE:** Within a certain organization, this definition is used strictly. Classifications which are not defined within the data set specification are not considered as classification correctness (these are considered to be domain consistency).

**2.3.3 Discussions on difficult cases**

**2.3.3.1 Relation between misclassification and completeness at feature type level**

At feature type level, completeness and thematic accuracy/classification correctness are strongly related to each other. Indeed the misclassification of one feature instance to the wrong feature type will appear in the evaluation of completeness for both feature types (one commission and one omission).

Therefore it is recommended when evaluating completeness at feature level to be aware that some of commission or omission error may come from misclassification issues. It could then be useful to provide classification correctness information, but the error will then be reported twice.

To avoid reporting errors twice, it is possible to report completeness at one upper level (data set, grouping of feature type, etc.), and misclassification at feature level.

**2.3.3.2 Quality elements related to unique identifiers**

Some use cases are presented below associated with relevant data quality elements for describing issues with unique identifiers, see Table 1.

Table 1 – Quality elements related to unique identifiers

|  |  |
| --- | --- |
| **Use case** | **Data quality element to consider** |
| All the unique identifiers shall have a format that fits the rules for defining them. | format consistency domain consistency |
| All the unique identifiers used are valid according to a list of reserved unique identifiers. | domain consistency |
| The same feature instance is present twice with the same unique identifier. | completeness conceptual consistency (unique identifiers shall be unique) |
| The same feature instance is present twice with different unique identifiers.  NOTE The challenge here is to be sure that the two feature instances are really two representations of the same real world object. | commission |

**2.4 Aggregation of data quality results**

**2.4.1 Introduction**

An evaluation based on a single data quality element is usually not sufficient for a user to be satisfied. For a potential user, it will be of great advantage to find a statement telling that the product is evaluated based on a specification. Such a statement is an aggregated data quality result, and may be useful also in other situations than reporting conformance to a specification.

The quality of a data set may be represented by one or more aggregated data quality results (ADQR).The ADQR combines quality results from data quality evaluations based on different data quality elements or different data quality scopes.

Examples of methods that may be used for producing an ADQR are given in 2.4.2 to 2.4.4.A data set may be deemed to be of an acceptable aggregate quality even though one or more individual data quality results fails acceptance. Aggregation should therefore only be used when compelling reasons exist. The meaning of the aggregate data quality result should always be made clear.

As the ADQR may be difficult to fully understand, the meaning of the aggregate data quality result should be understood before drawing conclusions based on aggregate data quality results for the quality of the data set.

**2.4.2 100 % pass/fail**

Each data quality result involved in the computation is given a Boolean value of one (1) if it passed and zero (0) if it failed. The aggregate quality is determined by the equation,

ADQR=v1\*v2 \*v3 \*... \* vn, where n is the number of data quality measurement frames.

If ADQR=1, then the overall data set quality is deemed to be fully conformant, hence pass. If ADQR=0, then it is deemed non-conformant, hence fail. The technique does not provide a result that indicates location or magnitude of the non-conformance.

**2.4.3 Weighted pass/fail**

Each data quality result involved in the computation is given a Boolean value of one (1) if it passed and a zero (0) if it failed. Based on the significance for the purpose of the product, a weight value between 0 and 1, inclusive, is assigned to each data quality result. The total of all the weights should equal 1.The choice of weights is a subjective decision made by the data producer or user. The reason for the data producer's decision should be reported as part of the result. The aggregated quality is determined by the equation,

ADQR= v1\*w1 + v2\*w2 + v3\*w3+...+vn\*wn, where n is the number of data quality measurement frames.

This technique does provide a magnitude value indicating how close a data set is to full conformance as measured. It does not provide a quantitative value that indicates where conformance or non- conformance occurs.

**2.4.4 Maximum/minimum value**

Each data quality result is given a value v based on the significance of a data quality result for the purpose of the product. The reason for the data producer's decision should be reported as part of the data set's quality result. The aggregated quality is determined by either of the two equations,

ADQR = MAX (vi, i=1...n) or ADQR = MIN (vi, i = 1...n) where n is the number of data quality measurement frames measured.

This technique provides a magnitude value indicating how close a data set is to full conformance as measured, but only in terms of the data quality measurement frame represented by the maximum or minimum. It does provide a quantitative value that indicates where conformance or non-conformance occurs when the selected data quality measurement frame is reported along with the ADQR. However, this type of ADQR tells little about the magnitude of the other data quality results.

**3 Data quality reporting**

**3.1 General**

Data quality shall be reported as metadata. In order to provide more details than reported as metadata, a standalone quality report may additionally be created. Its structure is free. However, the standalone quality report shall not replace the metadata. The metadata should provide a reference to the standalone quality report when it exists.

**3.2 Why report data quality**

The need to report data quality exists for a number of reasons including the following:

* to aid discovery and encourage use of the data set;
* to demonstrate the compliance to a data product specification or to user requirements;
* as part of supplier management initiatives;
* to permit downstream judgments about the quality of information derived from the data set;
* to permit rational (optimal) decision-making when it is known that all data contains imperfections.

**3.3 When to report quality information**

Data sets are continually being created, updated and merged with the result that the quality or a component of the quality of a data set may change. The quality of a data set can be affected by three conditions:

* when any quantity of data are deleted from, modified or added to a data set,
* when a data set's data product specification is modified or new user specified data quality requirements are identified,
* when the real world has changed.

The first condition, a modification to a data set, may occur frequently. Many data sets are not static. There is an increase in the interchange of information, the use of data sets for multiple purposes and an accompanying update and refinement of data sets to meet multiple purposes. If the reported quality of a data set is likely to change with modifications of the data set, the quality of this data set should be reassessed and updated as required when changes occur.

Complete knowledge of all applicable data quality elements should be available when a data set is created. Only the data producer's usage (assuming the data producer actually uses the data set) of a data set can initially be reported. There is a reliance on data users to report uses of a data set that differ from its intended purpose so that continual updates to this particular data quality overview element can be made to reflect occurring, unforeseen uses.

The second condition, a modification to a data set's data product specification, is most likely to occur before initial data set construction and prior to the release of quality information. It is conceivable, however, that as a data set is used, its data product specification is updated so that future modifications to the data set will better meet the actual needs. As the data product specification changes, the quality of the current data set also changes. The quality information for a data set should always reflect the current data set given its current data product specification,

The third condition, a change of the real world, occurs continuously. Changes may be caused by natural phenomena. But it is most often a result of human activity. Changes are often very rapid and dramatic. For this reason, the date of data collection is equally important as the date of quality evaluation when judging the quality of a data set. In some cases, when known, even the rate of change is of interest. The update frequency of the data set may also be of interest in some cases. However, this international standard recognizes that it might not be possible to created new data quality report every time the real world changes.

**3.4 How to report quality information**

**3.4.1 General**

Quality information may be reported as metadata and as a standalone quality report. These two mechanisms complement each other by allowing the reporting of data quality evaluation with different levels of detail:

* The metadata aims at providing short, synthetic and generally-structured information to enable metadata interoperability and web services usage;
* The standalone quality report may be used to provide fully detailed information about the data quality evaluation. The standalone quality report is to be provided attached to the data set or product for direct human reading.

For example, in the case of aggregation of different quality results, the standalone quality report will provide full information on the original results (with evaluation procedures and measures applied), the aggregated result and the aggregation method whereas the metadata may describe only the aggregated result with a reference to the original results described in the standalone quality report.

**3.4.2 Reporting quality information as metadata**

The class MD\_Metadata, aggregates zero, one or several data quality units (instances of the class DQ\_DataQuality). See Figure 2.



Figure 2 – Data quality information (from S-100 part 4c)

**3.4.3 Reporting quality information within a standalone quality report**

The standardization of terminology (e.g. the data quality elements) and structure of the underlying data quality information will be of benefit to users familiar with the standard and facilitate better understanding and comparison. Further, a statement of compliance to the standard within the report may be of value to users.

A standalone quality report should contain a scope to easily identify the extent to which the report covers the data set under evaluation.

Each report should contain sufficient information to meaningfully describe the relevant aspects of data quality and their results. This may take the form of references to supporting documentation such as a data product specification or measure catalogue.

The full structure of this standalone quality report has intentionally not been standardized so that each particular organization is able to adapt it for its own needs, practices and evaluation procedures. It may be some free text. However, the amount of quality information may be important. It is then important to present it in a succinct, easily understood and easily retrievable way. It is for example possible to follow the organization described in this International Standard.

**3.4.4 Particular cases**

**3.4.4.1 Reporting aggregation (aggregated results)**

Where the result has been aggregated, a standalone quality report should be provided to complete the information provided in the metadata. Within this standalone quality report, fully detailed information on the original result [with measure(s) and evaluation procedure(s)], aggregated result and aggregation method should be provided.

Within the metadata:

1. When several quality results for the same data quality element are aggregated into a single result of this element, the result should be reported in metadata as a result for this data quality element.
2. When several quality results for different data quality elements are aggregated into a single result, this should be reported in metadata as a result for the usability element.

In both cases, in metadata, at least a reference to the original data quality results shall be provided for an aggregated result, and information on the aggregation measure and aggregation method may be provided.

**3.4.4.2 Reporting derivation (derived results)**

When derived results are only reported in metadata, a standalone quality report should also be generated to provide the original data quality results from which the derived result has been determined. The metadata should then provide the reference to the standalone quality report and the original data quality result.

**EXAMPLE**: Conformance result is often derived from a quantitative result. If only the conformance result is provided in metadata, then the quantitative results should be provided in a standalone quality report.

**3.4.4.3 Reference to the original data quality result**

When derived or aggregated result(s) are reported in metadata, the reference to the original data quality result may be provided using two attributes:

* The attribute derivedElement references a quality element [and its result(s)] described in the metadata;
* The attribute standaloneQualityReportDetails references the part of the standalone quality report where the original result(s) are described.

**Annex**

**An Example of Evaluating and Reporting Data Quality of a S-100 Product**

*Note: An example of a standalone quality report will be provided in this Annex.*