IHO International Hydrographic Organization

# 19th Meeting of the Data Quality Working Group

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

## Agenda Item 7.1A

DQWG-19,  VTC Event, 25- 26 March 2024

# BACKGROUND

- Metadata-Data Quality has been described in S-100 Part 4C.
- DQ Measures and Recommendations for Product Specification developers  are in S-97.
- A recommended template of DQ chapter of S-100 based product specifications has been developed by DQWG.
- Lack of guidance and recommendations on how to evaluation DQ of a S-100 products and how to report the DQ evaluation result.

# BACKGROUND

- Validation is only a part of and a kind of DQ evaluation of a S-100 products.

- DQWG will help the S-100 validation by participating in the developing, reviewing and revising the validation checks.

- DQWG will also help S-100 validation by developing the DQ evaluation document to describe how to evaluate whether a dataset comply with related PS and how to report the evaluation result.

# 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 as follows:

- Development of a data product specification or user requirements.
- Quality control during data set creation.
- Inspection for conformance to a data product specification (e.g. S-100 Validation).
- Evaluation of data set conformance to user requirements.
- Quality control during data set update .

IHO

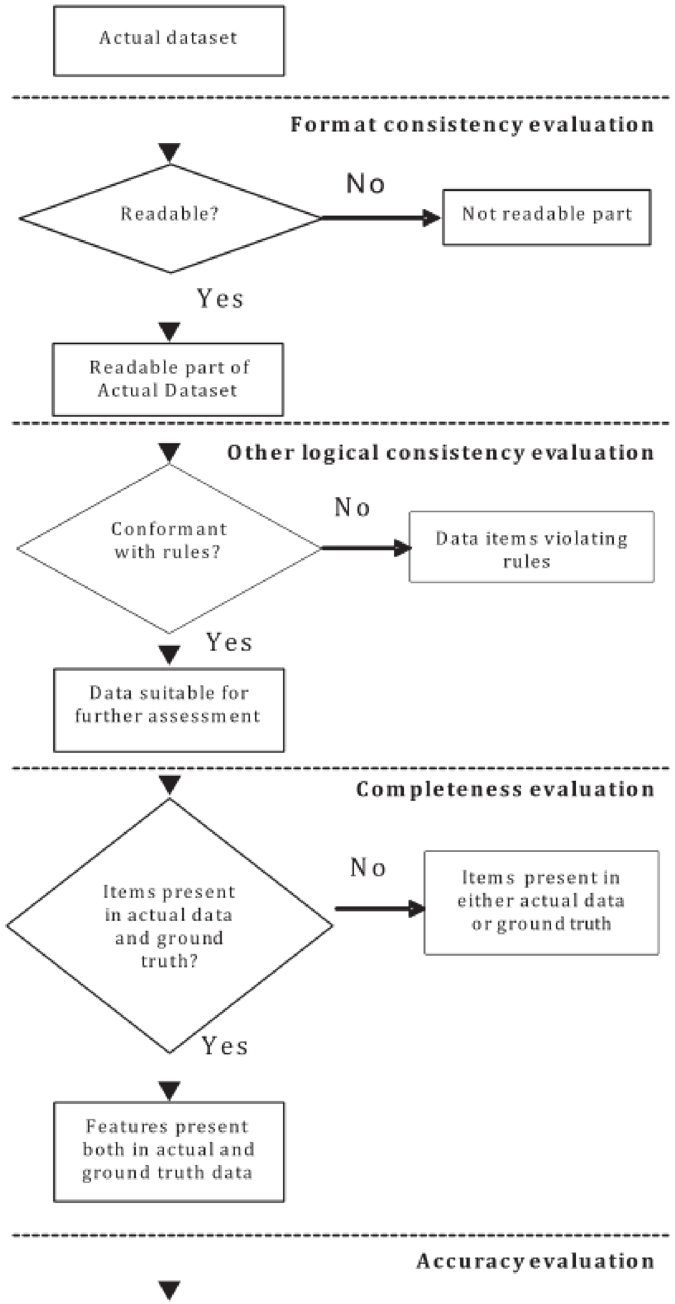
International Hydrographic Organization

# 2. HOW TO IMPLEMENT THE DATA QUALITY EVALUATION PROCEDURES

## 2.1 Ordering in data quality evaluation

When evaluating data quality, the usual ordering is:

1. Logical consistency/Format consistency;
2. Logical consistency/others;
3. Completeness;
4. Accuracy (positional, thematic and temporal aspects).



Actual dataset
Format consistency evaluation
Readable?
No
Not readable part
Yes
Readable part of Actual Dataset
Other logical consistency evaluation
Conformant with rules?
No
Data items violating rules
Yes
Data suitable for further assessment
Completeness evaluation
Items present in actual data and ground truth?
No
Items present in either actual data or ground truth
Yes
Features present both in actual and ground truth data
Accuracy evaluation

# 2. HOW TO IMPLEMENT THE DATA QUALITY EVALUATION PROCEDURES

## 2.2 Data quality evaluation methods

- **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.
  - For both external and internal evaluation methods, one of the following inspection methods may be used:- Full inspection; - Sampling.

- **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.

# 2. HOW TO IMPLEMENT THE DATA QUALITY EVALUATION PROCEDURES

## 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.

- 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.

# 2. HOW TO IMPLEMENT THE DATA QUALITY EVALUATION PROCEDURES

## 2.3 Guidelines for the use of data quality elements

### 2.3.1 The relationships between the data quality elements

### 2.3.1.1 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 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.

IHO

International Hydrographic Organization

# 2. HOW TO IMPLEMENT THE DATA QUALITY EVALUATION PROCEDURES

## 2.3.1.2 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: 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").

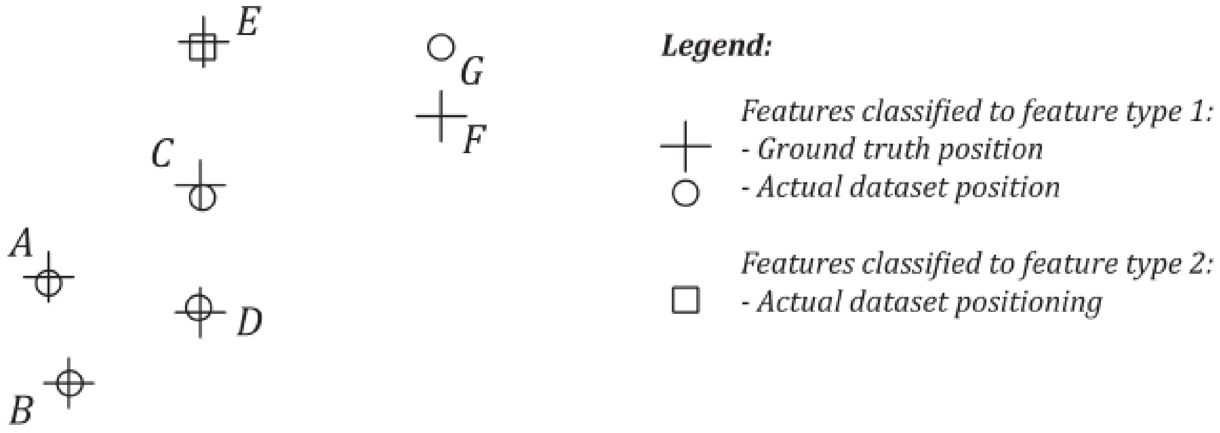
# 2. HOW TO IMPLEMENT THE DATA QUALITY EVALUATION PROCEDURES

**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:
- the thematic closeness (usually expressed as feature type);
- the geographical closeness of the features.

# 2. HOW TO IMPLEMENT THE DATA QUALITY EVALUATION PROCEDURES

- 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).


E
G
F
C
A
D
B
Legend:
Features classified to feature type 1:
- Ground truth position
- Actual dataset position
Features classified to feature type 2:
- Actual dataset positioning

# 2. HOW TO IMPLEMENT THE DATA QUALITY EVALUATION PROCEDURES

## 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.

#### 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. HOW TO IMPLEMENT THE DATA QUALITY EVALUATION PROCEDURES

## 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

- This conceptual schema may include:

a) the name of all classes (feature types, data types, etc.),
b) the attribute names for all classes, and also the multiplicity limitations,
c) the domains for all attributes,
d) the relationships between the classes,
e) the topological relationships between feature types, e.g., the relationship between an area and the border lines.
f) the relationship between feature type attributes for different feature types.

### 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.

# 2. HOW TO IMPLEMENT THE DATA QUALITY EVALUATION PROCEDURES

**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, 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.

**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.

# 2. HOW TO IMPLEMENT THE DATA QUALITY EVALUATION PROCEDURES

## 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. HOW TO IMPLEMENT THE DATA QUALITY EVALUATION PROCEDURES

## 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

- 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).

### 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.

- 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.

# 2. HOW TO IMPLEMENT THE DATA QUALITY EVALUATION PROCEDURES

**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.

- 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.

## 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 attributes to a universe of discourse**

- 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. HOW TO IMPLEMENT THE DATA QUALITY EVALUATION PROCEDURES

## 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. HOW TO IMPLEMENT THE DATA QUALITY EVALUATION PROCEDURES

## 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.

| Use case | Data quality element to consider |
|---|---|
| All the unique identifiers shall have a format that fits the rules for defining them. | Format consistency;<br>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;<br>Conceptual consistency (unique identifiers shall be unique) |
| The same feature instance is present twice with different unique identifiers.NOTE | Commission |

# 2. HOW TO IMPLEMENT THE DATA QUALITY EVALUATION PROCEDURES

## 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.

- 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.

- 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.

# 2. HOW TO IMPLEMENT THE DATA QUALITY EVALUATION PROCEDURES

## 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. HOW TO IMPLEMENT THE DATA QUALITY EVALUATION PROCEDURES

## 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. HOW TO IMPLEMENT THE DATA QUALITY EVALUATION PROCEDURES

## 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.

# 3 DATA QUALITY REPORTING

## 3.1 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 DATA QUALITY REPORTING

## 3.2 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.

# 3 DATA QUALITY REPORTING

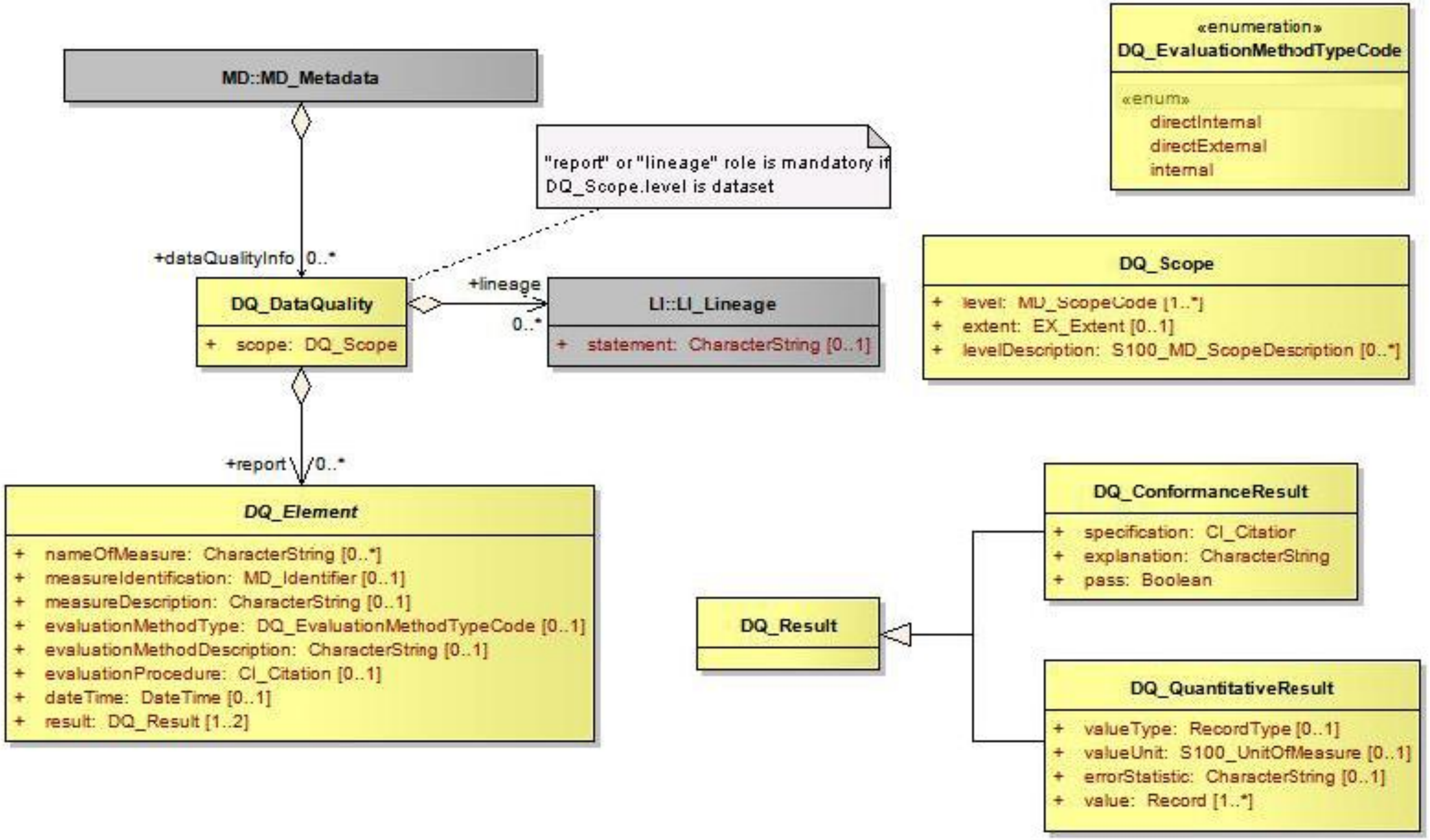
## 3.3 How to report quality information

### 3.3.1 General

- Data quality shall be reported as metadata.
- The metadata aims at providing short, synthetic and generally-structured information to enable metadata interoperability and web services usage;
- 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.
- The standalone quality report is to be provided attached to the data set or product for direct human reading.

## 3.3.2 Reporting quality information as metadata

The class MD_Metadata, aggregates zero, one or several data quality units (instances of the class DQ_DataQuality).


MD::MD_Metadata
+dataQualityInfo
0..*
"report" or "lineage" role is mandatory if DQ_Scope.level is dataset
DQ_DataQuality
+ scope: DQ_Scope
+lineage
0..*
LI::LI_Lineage
+ statement: CharacterString [0..1]
«enumeration»
DQ_EvaluationMethodTypeCode
«enum»
directInternal
directExternal
internal
DQ_Scope
+ level: MD_ScopeCode [1..*]
+ extent: EX_Extent [0..1]
+ levelDescription: S100_MD_ScopeDescription [0..*]
+report
0..*
DQ_Element
+ nameOfMeasure: CharacterString [0..*]
+ measureIdentification: MD_Identifier [0..1]
+ measureDescription: CharacterString [0..1]
+ evaluationMethodType: DQ_EvaluationMethodTypeCode [0..1]
+ evaluationMethodDescription: CharacterString [0..1]
+ evaluationProcedure: CI_Citation [0..1]
+ dateTime: DateTime [0..1]
+ result: DQ_Result [1..2]
DQ_Result
DQ_ConformanceResult
+ specification: CI_Citation
+ explanation: CharacterString
+ pass: Boolean
DQ_QuantitativeResult
+ valueType: RecordType [0..1]
+ valueUnit: S100_UnitOfMeasure [0..1]
+ errorStatistic: CharacterString [0..1]
+ value: Record [1..*]

# 3 DATA QUALITY REPORTING

## 3.3.3 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.

# 3 DATA QUALITY REPORTING

## 3.3.4 Particular cases

### 3.3.4.1 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:
- 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.
- 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.

IHO

International
Hydrographic
Organization

# 4 AN EXAMPLE OF EVALUATING AND REPORTING DATA QUALITY OF A S-100 PRODUCT

- This Annex will be an essential part of the document, though it is empty now.
- The Annex will present a sample real world, a sample universe and a sample dataset, quality requirements that sample dataset must comply with, quality evaluation process, reporting data quality, etc..
- An example of a standalone quality report will also be provided in this Annex.
- People will be able to do the DQ evaluation of S-100 products by imitating this example.

# ACTIONS

The DQWG is requested to:

a.**Note** the information provided;

b.**Establish** a new subWG to develop the recommendations on the data quality evaluation of S-100 products.