**DATA QUALITY CHECKLIST**

Drafted: 2018-05-09, version 0.3

Author: R. Broekman (DQWG-Chair)

**Introduction:**

The IHO Data Quality Working Group (DQWG) has been tasked by the IHO Hydrographic Services and Standards Committee (HSSC) to develop and maintain a Data Quality Checklist for Product Specification developers. This contributes to the assigned DQWG objective to ensure that data quality aspects are addressed in an appropriate and harmonized way for all S-100 based product specifications.

**Usage:**

This Data Quality Checklist can be used by HSSC Workgroups developing S-1xx based Product Specifications. It provides 10 recommendations of appropriate data quality measures as deemed necessary to be used within S-1xx based Product Specifications. A Data Product Specification is a detailed description of a dataset or dataset series together with additional information that will enable it to be created, supplied to, and used by another party. A Data Product Specification provides a description of Hydrographic Concepts and a specification for mapping the Universe of Discourse to a dataset. It may be used for production, sales, end-use or other purposes.

When drafting a Product Specification, the Data Quality Checklist will serve as a guidance document to verify if the appropriate Data Quality Elements have been included in the Product Specification. A Data Quality Element is a quantitative component documenting the quality of a dataset. The applicability of a data quality element to a dataset depends on both the dataset’s content and its Product Specification, the result being that all available data quality elements may not be applicable to all datasets.

**Definitions[[1]](#footnote-1):**

A data quality:

* subelement is a component of a data quality element describing a certain aspect of that data quality element;
* measure is an evaluation of a data quality subelement;
* evaluation procedure is the whole of operations used in applying and reporting quality evaluation methods and their results;
* result is a 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;
* scope is the extent or characteristic(s) of the data for which quality information is reported;
* 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;
* overview element is the non-quantitative component documenting the quality of a dataset. Information about the purpose, usage and lineage of a dataset is non-quantitative quality information.

The relations between these definitions are presented in the figure below:

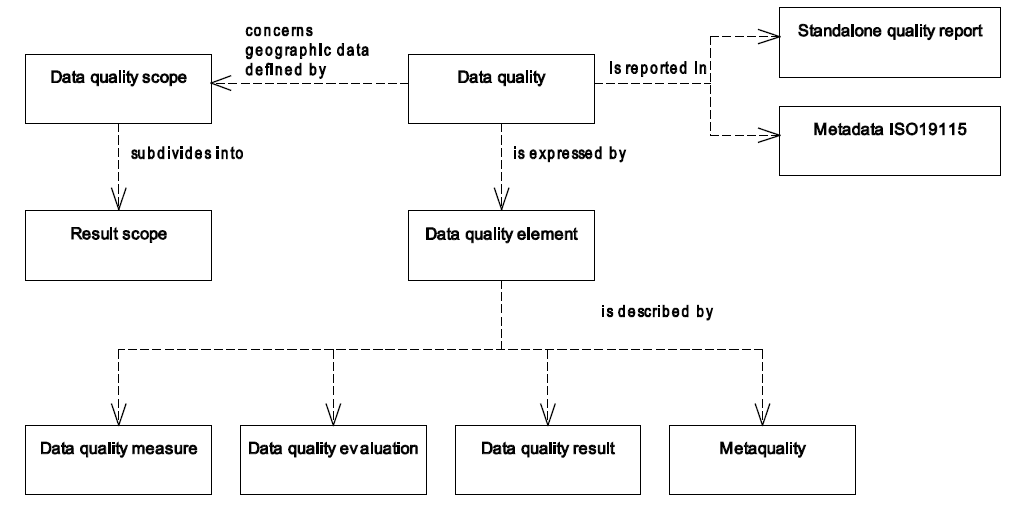


Figure 1 – Conceptual model of quality for geographic data[[2]](#footnote-2)

The components of Data Quality Measure can be divided into the following elements[[3]](#footnote-3):

1. Completeness
2. Logical Consistency
3. Positional Accuracy
4. Thematic Accuracy
5. Temporal Quality
6. Aggregation
7. Usability

Data Quality Evaluation can be divided into the following elements:

1. Data Quality Full Inspection
2. Data Quality Sample Based Inspection
3. Data Quality Indirect Evaluation
4. Data Quality Aggregation Derivation

Data Quality Result can be divided into the following elements:

1. Data Quality Conformance Result
2. Data Quality Quantitative Result
3. Data Quality Descriptive Result

**Description of Data Quality Measures[[4]](#footnote-4)**:

**Completeness** is defined as the presence and absence of features, their attributes and relationships. It consists of two data quality elements:

* commission – excess data present in a dataset;
* omission – data absent from a dataset.

**Logical Consistency** is defined as the degree of adherence to logical rules of data structure, attribution and relationships (data structure can be conceptual, logical or physical). If these logical rules are documented elsewhere (for example in a data product specification) then the source should be referenced (for example in the data quality evaluation). It consists of four data quality elements:

* conceptual consistency – adherence to rules of the conceptual schema;
* domain consistency – adherence of values to the value domains;
* format consistency – degree to which data is stored in accordance with the physical structure of the dataset;
* topological consistency – correctness of the explicitly encoded topological characteristics of a dataset.

**Positional Accuracy** is defined as the accuracy of the position of features within a spatial reference system. It consists of three data quality elements:

* absolute or external accuracy – closeness of reported coordinate values to values accepted as or being true;
* relative or internal accuracy – closeness of the relative positions of features in a dataset to their respective relative positions accepted as or being true;
* gridded data positional accuracy – closeness of gridded data spatial position values to values accepted as or being true.

**Thematic Accuracy** is defined as the accuracy of quantitative attributes and the correctness of non-quantitative attributes and of the classifications of features and their relationships. It consists of three data quality elements:

* classification correctness – comparison of the classes assigned to features or their attributes to a Universe of Discourse (e.g. ground truth or reference data);
* non-quantitative attribute correctness – measure of whether a non-quantitative attribute is correct or incorrect;
* quantitative attribute accuracy – closeness of the value of a quantitative attribute to a value accepted as or known to be true.

**Temporal Quality** is defined as the quality of the temporal attributes and temporal relationships of features. It consists of three data quality elements:

* accuracy of a time measurement – closeness of reported time measurements to values accepted as or known to be true;
* temporal consistency – correctness of the order of events;
* temporal validity – validity of data with respect to time.

**Usability** is based on user requirements. All quality elements may be used to evaluate usability. Usability evaluation may be based on specific user requirements that cannot be described using the quality elements described above. In this case, the usability element shall be used to describe specific quality information about a dataset’s suitability for a particular application or conformance to a set of requirements.

**Recommendations for S-1xx Product Specification developers**.

The Data Quality concept from ISO-19157 now has to be applied to S-1xx Product Specifications. This document provides recommendations where to find and apply Data Quality Measures as described in S-100 Universal Hydrographic Data Model – Edition 3.0.0 - … 2017.

**DQ\_CompletenessCommission**: Appendix 4c-C, Hydrographic Quality Metadata Attribute Definitions, page 10.

**DQ\_CompletenessOmission**: Appendix 4c-C, Hydrographic Quality Metadata Attribute Definitions, page 11

|  |
| --- |
| Recommendation 1: Data Quality Measure Completeness (Commission/Omission) to be included in the Product Specification |

**Logical Consistency > Conceptual Consistency**: the Conceptual Schema Language is described in S-100 -- Part 1, Conceptual Schema Language. It provides the description of:

* classes;
* attributes;
* basic data types;
* primitive types;
* complex types;
* predefined derived types;
* enumerated types;
* codelist types;
* relationships and associations;
* composition and aggregation;
* stereotypes and optional, conditional and mandatory – attributes and associations;
* naming and name spaces;
* notes;
* packages.

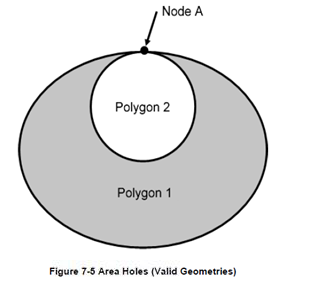
|  |
| --- |
| Recommendation 2: Data Quality Measure Conceptual Consistency to follow the guidelines from S-100 Part 1 and to be included in the Product Specification |

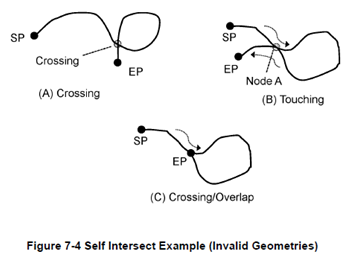
**Logical Consistency > Domain Consistency**: this is described in S-100 Part 5 – Feature Catalogue. This Part provides a standard framework for organizing and reporting the classification of real world phenomena in a set of geographic data. It defines the methodology for classification of the feature types and specifies how they are organized in a feature catalogue and presented to the users of a set of geographic data. This methodology is applicable to creating catalogues of feature types in previously uncatalogued domains and to revising existing feature catalogues to comply with standard practice. It applies to the cataloguing of feature types that are represented in digital form. Its principles can be extended to the cataloguing of other forms of geographic data.

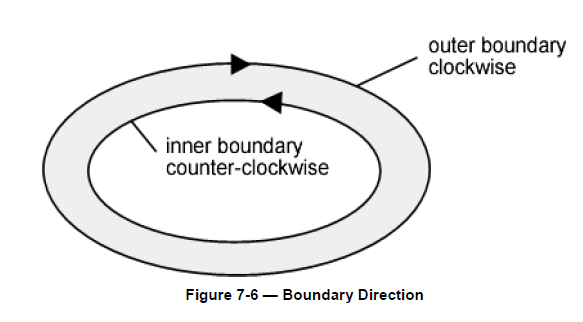
|  |
| --- |
| Recommendation 3: Data Quality Measure Domain Consistency to follow the guidelines from S-100 Part 5 and to be included in the Product Specification |

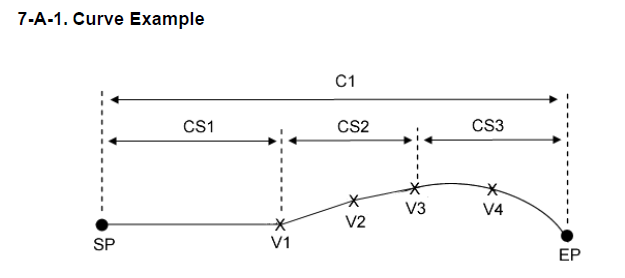
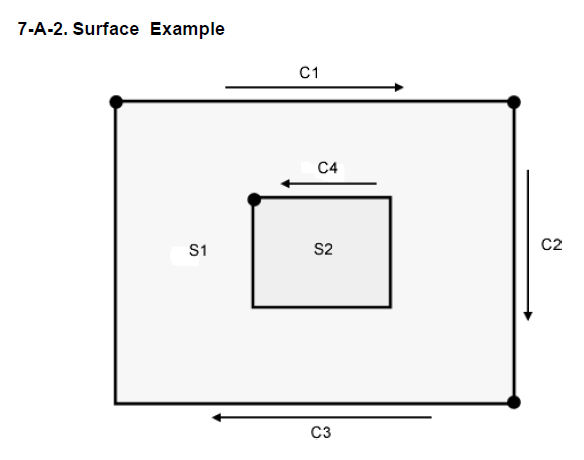
**Logical Consistency > Format Consistency** this is described in S-100 Part 10 – Encoding formats. S-100 does not mandate particular encoding formats so it is left to developers of product specifications to decide on suitable encoding standards and to document their chosen format. The issue of encoding information is complicated by the range of encoding standards that are available, which include but are not limited to: ISO/IEC8211, GML, XML, GeoTiff, HDF-5, JPEG2000.

|  |
| --- |
| Recommendation 4: Data Quality Measure Format Consistency to follow the guidelines from S-100 Part 10 and to be included in the Product Specification |

**Logical Consistency > Topological Consistency** this is described in S-100 Part 7 – Spatial Schema. It supports 0, 1, 2 and 2.5 dimensional spatial schemas and two levels of complexity – geometric primitives and geometric complexes. Some figures of this part are copied here as a quick reference:



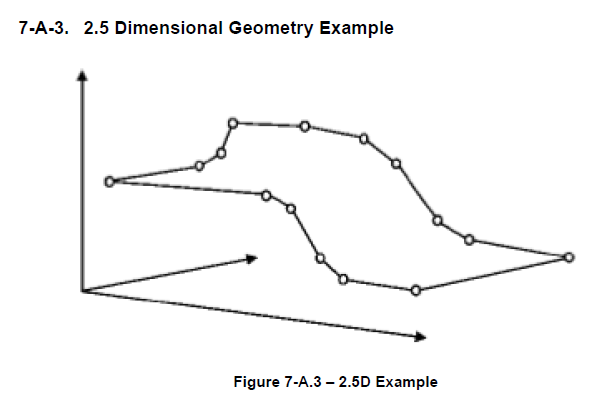


The following describes the geometrical elements of the curve example (Figure 7-A.1).

C1 (GM\_Curve) consists of CS1, CS2 and CS3 (GM\_CurveSegment). CS1 uses a geodetic interpolation, CS2 linear and CS3 circularArc3Points. SP (start point) and EP (end point) (GM\_Point) are the start and end points of C1 and can also be used indirectly as a 0-dimension position for a point feature. An array of control points for each segment consists of a combination of SP, EP and vertices as indicated in the above diagram. The orientation of C1 is + (forward) from SP to EP.

The following describes the geometrical elements of the surface example (Figure 7-A.2).

S1 (GM\_Surface) is represented by the surface patch P1 (GM\_Polygon) the boundary of which consists of exterior and interior rings. The exterior ring CC1 (GM\_CompositeCurve) is an aggregation of C1, C2, C3 (GM\_Curve), the interior ring C4 is a simple GM\_Curve.



In the depicted example, the curve which constitutes the exterior boundary of a GM\_Polygon consists of an array of 3D control points. Note that the surface interpolation must be “none”, which means that the position of interior points is not determined. The “planar” interpolation would only be acceptable if all points were lying on a plane.

The conditions for topological consistency are provided in S-100 Part 7 – Spatial Schema,

page 9-13.

|  |
| --- |
| Recommendation 5: Data Quality Measure Topological Consistency to follow the guidelines from S-100 Part 7 and to be included in the Product Specification |

**Positional Accuracy** is described by Part 4c - Metadata - Data Quality.

This is further subdivided into Absolute or External Accuracy, Vertical Position Accuracy, Horizontal Positional Accuracy , Gridded Data Position Accuracy.

One should take notice of the different ways how spatial data referencing is applied. Point set data includes a coordinate direct position for each point in the point set. (points/curves). Gridded data references the grid as a whole. The two spatial properties of gridded data describe how the spatial extent was tessellated into small units and spatial referencing to the earth. The ISO 19123 standard indicates that a grid may be defined in terms of a coordinate reference system. This requires additional information about the location of the grid’s origin within the coordinate reference system, the orientation of the grid axes, and a measure of the spacing between the grid lines. A grid defined in this way is called a rectified grid. If the coordinate reference system is related to the Earth by a datum, the grid is a georectified grid. The essential point is that the transformation of grid coordinates to coordinates of the external coordinate reference system is an affine transformation.

For Positional Accuracy, currently in the Hydrographic Community the 95% confidence level (Gaussian distribution) is commonly used. The Root Mean Square Error is commonly used in the scientific community. RMSE is the square root of the average of the set of squared differences between dataset coordinate values and coordinate values from an independent source of higher accuracy for identical points.

Other calculation methods are also possible, depending on the specific Product Specification.

|  |
| --- |
| Recommendation 6: Data Quality Measure Positional Accuracy to follow the guidelines from S-100 Part 4c and to be included in the Product Specification. The calculation of the Positional Accuracy is to be further harmonized within S-1xx Product Specifications. |

**Thematic Accuracy** is described in S-100 - Part 3: General Feature Model and Rules for Application Schema.

The data content of a geographic application is defined in accordance with a view of real world features and in the context of the requirements of a particular application. The content is structured in terms of objects. This document considers two types of object:

1. Features – features are defined together with their properties.
2. Information Types – information types are used to share information among features and other information types. Information types have only thematic attribute properties.

The assignment of an item to a certain class can either be correct or incorrect.

|  |
| --- |
| Recommendation 7: Data Quality Measure Thematic Accuracy to follow the guidelines from S-100 Part 3 and to be included in the Product Specification. |

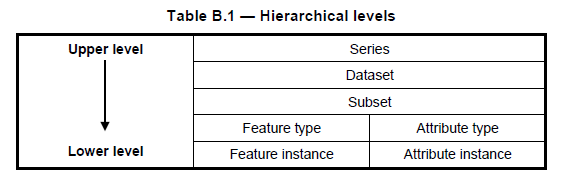
**Temporal Quality** is described by Part 4c - Metadata - Data Quality.

Temporal Consistency and Temporal Validity are recommended to be included as this provides the user with the guarantee that any information in the temporalspatio domain is registered correctly. For data elements with a very precise temporal attribute (e.g. remote sensing), the temporal accuracy may also be provided.

|  |
| --- |
| Recommendation 8: Data Quality Measure Temporal Quality to follow the guidelines from S-100 Part 4c and to be included in the Product Specification. Temporal Consistency and Temporal Validity should be included. |

**Aggregation**

Data Quality specified at upper level (e.g. series) is applicable at lower level (e.g. dataset), see Table below. If the Data Quality differs between upper and lower level, then supplemental information should be provided at lower level.



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.

The aggregated Data Quality result provides a result if the dataset has passed conformance to the Data Product Specification.

|  |
| --- |
| Recommendation 9: Data Quality Measure Aggregation results should be included to indicate if the dataset/dataset series have passed the Product Specifications. |

**Usability**

Usability is based on user requirements. All quality elements may be used to evaluate usability. Usability evaluation may be based on specific user requirements that cannot be described using the quality elements described above. In this case, the usability element shall be used to describe specific quality information about a dataset’s suitability for a particular application or conformance to a set of requirements.

All Product Specification have a paragraph describing Data Quality. To ensure harmonization across different Product Specifications, DQWG recommends that all Product Specifications share a common text explaining the concept of Data Quality -> Introduction to Data Quality.

The text below is a proposal for this common introduction:

**Introduction to data quality**

Data quality allows users and user systems to assess fitness for use of the provided data. Data quality measures and the associated evaluation are reported as metadata of a data product. This metadata improves interoperability with other data products and provides usage by user groups that the data product was not originally intended for. The secondary users can make assessments of the data product usefulness in their application based on the reported data quality measures.

For <this Product Specification> the following data quality elements have been included[[5]](#footnote-5);

* Conformance to this Product Specification;
* Intended purpose of the data product;
* Completeness of the data product in terms of coverage;
* Logical Consistency;
* Positional Uncertainty and Accuracy;
* Thematic Accuracy;
* Temporal Quality;
* Aggregation measures;
* Anything specifically required for the specified product;
* Validation checks or conformance checks including:
* General tests for dataset integrity;
* Specific tests for a specific data model;

|  |
| --- |
| Recommendation 10: Paragraph “Introduction to data quality” to be used by all S-1xx based Product Specifications. |

This list is derived from ISO 19157.

There are 20 different data quality measures defined that can be used for validation of S-1xx Product Specifications (PS). There is a recommendation for the target result of Positional Accuracy for depth contour lines and gridded bathymetry.

The full list of ISO 19157 applicable to S-1xx is available in S-100 – Appendix 4c-C, Hydrographic Quality Metadata Attribute Definitions.

| Data quality element and sub element | Definition | DQ measure / description | Evaluation scope | Applicable to spatial representation types |
| --- | --- | --- | --- | --- |
| Completeness / Commission | Excess data present in a dataset, as described by the scope. | numberOfExcessItems / This data quality measure indicates the number of items in the dataset, that should not have been present in the dataset. | dataset/dataset series | All S-1xx |
| Completeness / Commission | Excess data present in a dataset, as described by the scope. | numberOfDuplicateFeatureInstances / This data quality measure indicates the total number of exact duplications of feature instances within the data. | dataset/dataset series | All S-1xx |
| Completeness / Omission | Data absent from the dataset, as described by the scope. | numberOfMissingItems / This data quality measure is an indicator that shows that a specific item is missing in the data. | dataset/dataset series/spatial object type | All S-1xx |
| Logical Consistency / Conceptual Consistency | Adherence to the rules of a conceptual schema. | numberOfInvalidSurfaceOverlaps / This data quality measure is a count of the total number of erroneous overlaps within the data. Which surfaces may overlap and which must not is application dependent. Not all overlapping surfaces are necessarily erroneous. | spatial object / spatial object type | PS with geometric surfaces. |
| Logical Consistency / Domain Consistency | Adherence of the values to the value domains. | numberOfNonconformantItems / This data quality measure is a count of all items in the dataset that are not in conformance with their value domain. | spatial object / spatial object type | All S-1xx |
| Logical Consistency / Format Consistency | Degree to which data is stored in accordance with the physical structure of the data set, as described by the scope | physicalStructureConflictsNumber / This data quality measure is a count of all items in the dataset that are stored in conflict with the physical structure of the dataset. | dataset/dataset series | All S-1xx |
| Logical Consistency / Topological Consistency | Correctness of the explicitly encoded topological characteristics of the dataset, as described by the scope. | rateOfFaultyPointCurveConnections / This data quality measure indicates the number of faulty link-node connections in relation to the number of supposed link-node connections. This data quality measure gives the erroneous point-curve connections in relation to the total number of point-curve connections. | spatial object / spatial object type | PS with curves. |
| Logical Consistency / Topological Consistency | Correctness of the explicitly encoded topological characteristics of the dataset, as described by the scope. | numberOfMissingConnectionsUndershoots / This data quality measure is a count of items in the dataset within the parameter tolerance that are mismatched due to undershoots. | spatial object / spatial object type | PS with curves. |
| Logical Consistency / Topological Consistency | Correctness of the explicitly encoded topological characteristics of the dataset, as described by the scope. | numberOfMissingConnectionsOvershoots / This data quality measure is a count of items in the dataset within the parameter tolerance that are mismatched due to overshoots. | spatial object / spatial object type | PS with curves. |
| Logical Consistency / Topological Consistency | Correctness of the explicitly encoded topological characteristics of the dataset, as described by the scope. | numberOfInvalidSlivers / This data quality measure is a count of all items in the dataset that are invalid sliver surfaces. A sliver is an unintended area that occurs when adjacent surfaces are not digitized properly. The borders of the adjacent surfaces may unintentionally gap or overlap to cause a topological error. | dataset / dataset series | PS with geometric surfaces. |
| Logical Consistency / Topological Consistency | Correctness of the explicitly encoded topological characteristics of the dataset, as described by the scope. | numberOfInvalidSelfIntersects / This data quality measure is a count of all items in the dataset that illegally intersect with themselves. | spatial object / spatial object type | PS with curves / geometric surfaces. |
| Logical Consistency / Topological Consistency | Correctness of the explicitly encoded topological characteristics of the dataset, as described by the scope. | numberOfInvalidSelfOverlap / This data quality measure is a count of all items in the dataset that illegally self-overlap. | spatial object / spatial object type | PS with curves / geometric surfaces. |
| Positional Accuracy / Absolute or External Accuracy | Closeness of reported coordinative values to values accepted as or being true. | Root Mean Square Error / Standard deviation, where the true value is not estimated from the observations but known a priori. | spatial object / spatial object type | PS with objects that have coordinative values associated. |
| Positional Accuracy / Vertical Position Accuracy | Closeness of reported coordinative values to values accepted as or being true. | linearMapAccuracy2Sigma / Half length of the interval defined by an upper and lower limit in which the true value lies with probability 95%. | spatial object / spatial object type | PS with objects that have a vertical coordinative values associated. |
| Positional Accuracy / Horizontal Position Accuracy | Closeness of reported coordinative values to values accepted as or being true. | linearMapAccuracy2Sigma / Half length of the interval defined by an upper and lower limit in which the true value lies with probability 95%. | spatial object / spatial object type | PS with objects that have a horizontal coordinative values associated. |
| Positional Accuracy / Gridded Data Position Accuracy | Closeness of reported coordinative values to values accepted as or being true. | Root mean square error of planimetry / Radius of a circle around the given point, in which the true value lies with probability P. | spatial object / spatial object type | PS with objects that have a gridded coordinative values associated. |
| Temporal Quality / Temporal Consistency | Consistency with time. | Correctness of ordered events or sequences, if reported. | dataset/dataset series/spatial object type | PS with objects that have a time value associated. |
| Thematic Accuracy / ThematicClassificationCorrectness | Comparison of the classes assigned to features or their attributes to a universe of discourse. | miscalculationRate / This data quality measure indicates the number of incorrectly classified features in relation to the number of features that are supposed to be there. [Adapted from ISO 19157]  This is a RATE which is a ratio, and is expressed as a REAL number representing the rational fraction corresponding to the numerator and denominator of the ratio.  For example, if there are 1 items that are classified incorrectly and there are 100 of the items in the dataset then the ratio is 1/100 and the reported rate = 0.01. | dataset/dataset series/spatial object type | All S-1xx PS. |
| Aggregation Measures / AggregationMeasures | In a data product specification, several requirements are set up for a product to conform to the specification. | DataProductSpecificationPassed / This data quality measure is a boolean indicating that all requirements in the referred data product specification are fulfilled. | dataset/dataset series/spatial object type | PS that a require a complete pass of all elements of a dataset/dataset series/spatial object types |
| Aggregation Measures / AggregationMeasures | In a data product specification, several requirements are set up for a product to conform to the specification. | DataProductSpecificationFailRate / This data quality measure is a number indicating the number of data product specification requirements that are not fulfilled by the current product/dataset in relation to the total number of data product specification requirements. | dataset/dataset series/spatial object type | PS that a require a complete pass of all elements of a dataset/dataset series/spatial object types |

Recommendations for Positional Accuracy / Absolute or External Accuracy[[6]](#footnote-6):

Maximum RMSE (horizontal) = E / 10000

Maximum RMSE (vertical) = Vint / 6

Recommendation for Positional Accuracy / Gridded Data Position Accuracy:

Maximum RMSE (horizontal) = GSD / 6

Maximum RMSE (vertical) = GSD / 3

Where:

E = Denominator of intended scale of mapping

Vint = normal contour line interval

GSD = Ground Sampling Distance

1. S-100 version 3.0.0 page A6-7 Terms and Definitions [↑](#footnote-ref-1)
2. ISO 19157 – Geographic Information, Data Quality page 6. [↑](#footnote-ref-2)
3. ISO 19157 – Geographic Information, Data Quality page 7 [↑](#footnote-ref-3)
4. ISO 19157 – Geographic Information, Data Quality page 9 [↑](#footnote-ref-4)
5. As deemed necessary by the IHO – Hydrographic Standards and Services Committee [↑](#footnote-ref-5)
6. INSPIRE D2.8.II.1 Data Specification on Elevation – Technical Guidelines, page 95 [↑](#footnote-ref-6)