MSDIWG14

14th Meeting of the International Hydrographic Organization Marine Spatial Data Infrastructures Working Group

Data Quality

Portugal – LCDR Telmo Dias

Genoa, January 30th to February 3rd, 2022

IHO

International Hydrographic Organization

Agenda

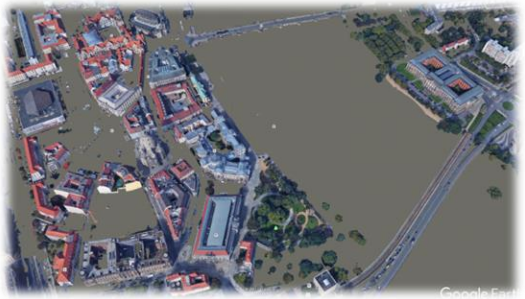
1. Background
2. Quality standards
3. Data quality elements
4. Data quality evaluation
5. Data quality reporting

1. Background

1.1. Data quality importance

➥ **Traditional purpose of hydrographic data:**

- Safety of navigation (bathymetry):
  - ✓ Minimum soundings;
  - ✓ IHO S-44 Survey Order;
  - ✓ CATZOC.



➥ **Other purposes than the producers intended ones:**

- e.g.: flood simulation (bathymetry, altimetry).

➥ **Increase the use of hydrographic data:**

- Data quality is crucial for a successful usage of geospatial data;
- Allows users to compare and select data best suited to their needs and requirements.
- Standards ensure a common understanding for quality, namely, what it is and how it can be expressed and measured.

1. Background

1.2. Data quality definition

➥ **Quality:**

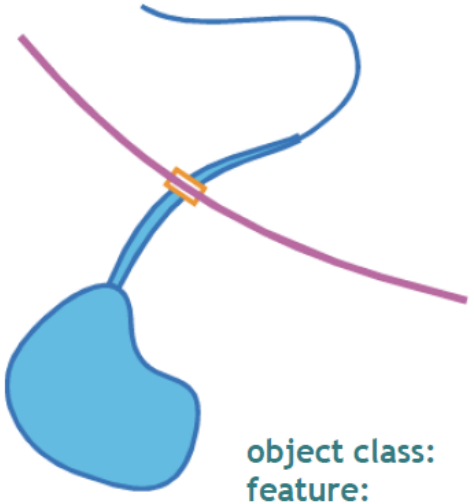
- Definition (ISO 9000:2005): degree to which a set of inherent characteristics fulfils requirements;
- Common understanding: fitness for use.

➥ **Comments:**

- The quality of something can be determined by **comparing** a set of inherent characteristics with a set of requirements.
- Quality is a question of **degree**: how well does this set of inherent characteristics comply with this set of requirements?
- Quality is a **relative concept** and it cannot be established in a vacuum. Quality is always relative to a set of requirements.

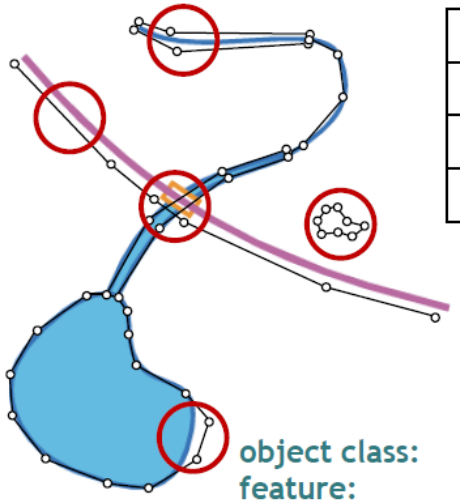
1. Background

1.3. Error types



object class: lake
feature:
name: Harry
max. depth: 12,6 m basic data
area: 10,17 ha

water quality: - thematic data



object class: lake
feature:
name: Harry
max. depth: 12,6 m basic data
area: 10,17 ha

water quality: - thematic data

lake_has_attributes	
lake	Environment
#1302	#08093
#1302	#08095

lake			
name	area	max. depth	ID
Harry	10,2	12,4	#1302

incorrect attributes

incorrect relation

environmental data for storage water bodies		
ID	water quality	time
#08093	✓	14.02.1997
#08094		15.02.1997

The attributes of the object belonging to the class "lake" are introduced in the relational model, while certain errors occur.

2. Quality standards

2.1. ISO publications

➥ **Standards related to quality:**

- **ISO 19131**
  Geographic information – Data product specifications
- **ISO 19157**
  Geographic information – Data quality
- **ISO 19115**
  Geographic information – Metadata

Geospatial Data

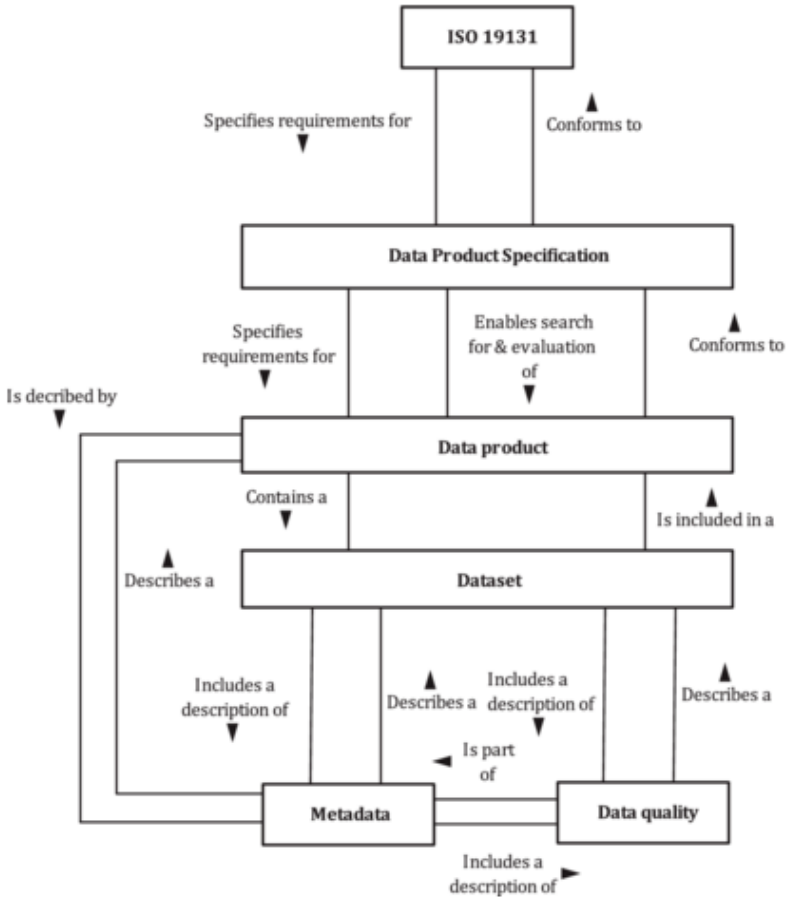
- **ISO 9000**
  Quality management systems – Fundamentals and vocabulary

2. Quality standards
2.2. ISO 19131

➥ Data product specification:

- A detailed description of a dataset, together with additional information that will enable it to be created, supplied to and used by another party.
- It is a precise technical description of the data product in terms of the requirements that it will or may fulfill.
- The data product specification only defines how the dataset should be, but for various reasons, compromises may need to be made in the implementation.
- The metadata associated with the product dataset should reflect how the product dataset actually is.
- It may be created by producers to specify their product, or by users to state their requirements.

Relations between this document (ISO 19131), the data product specification and the datasets



2. Quality standards

2.3. ISO 19157

➥ **Geographic information – data quality:**

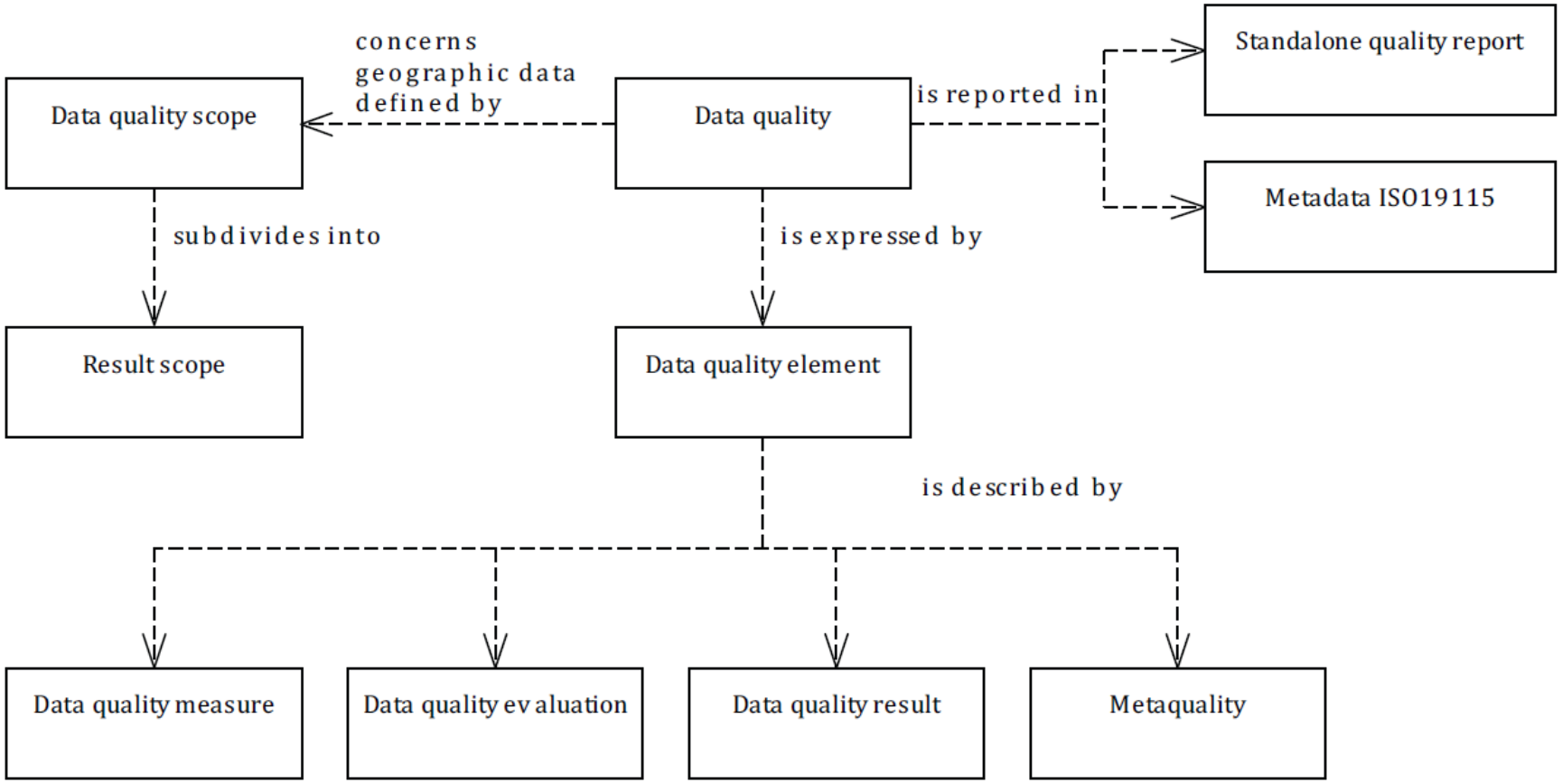


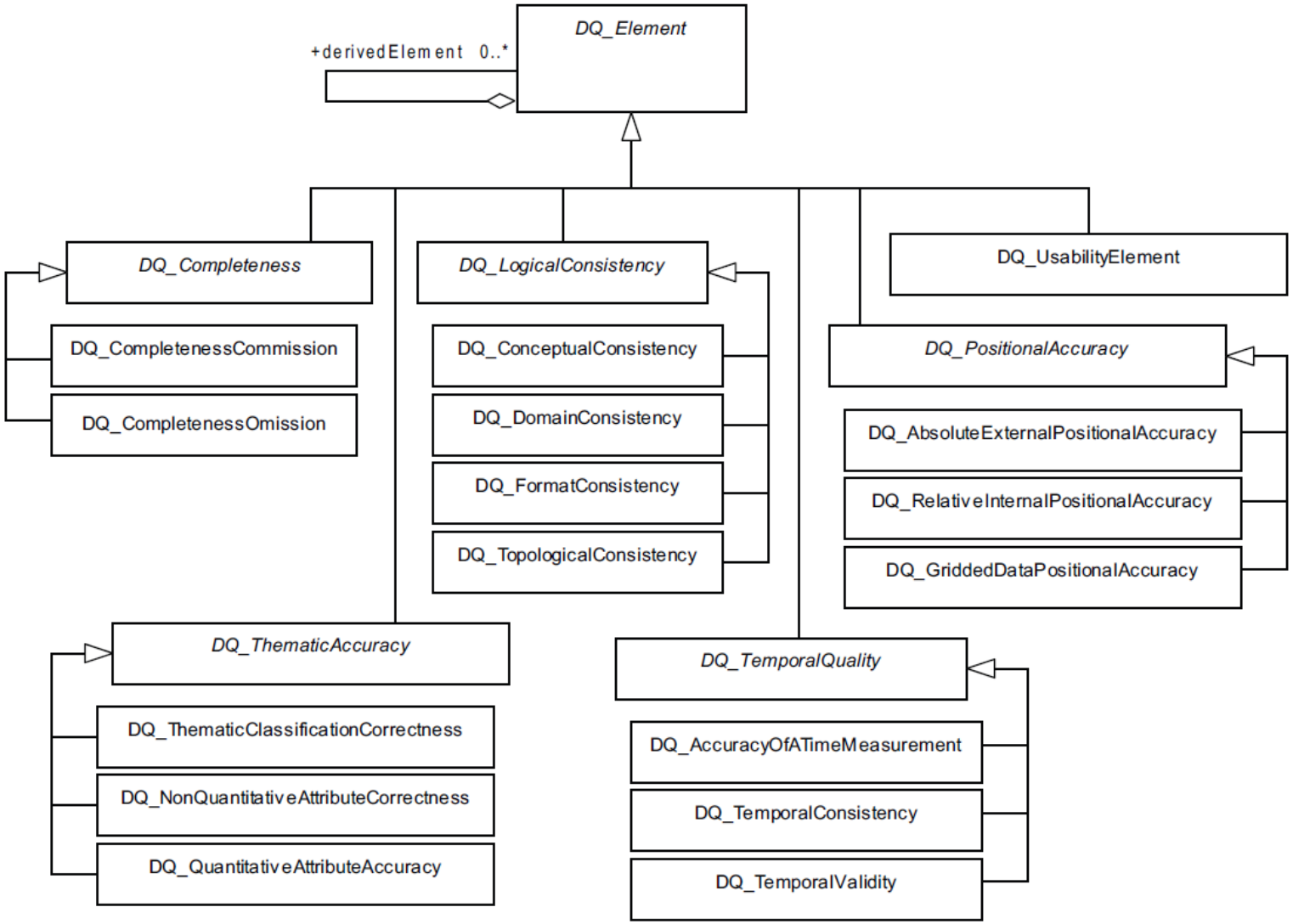
Figure 1 — Conceptual model of quality for geographic data

International Hydrographic Organization

3. Data quality elements

3.1. Overview

- Completness
- Logical consistency
- Positional accuracy
- Temporal accuracy
- Thematic accuracy
- Usability



3. Data quality elements

3.1. Overview

- **Completness**
  - Presence and absence of features, their attributes and relationships.
    - ✓ **Commission**: excess data present in a dataset.
    - ✓ **Omission**: data absent from a dataset.
- Logical consistency
- Positional accuracy
- Thematic accuracy
- Temporal quality
- Usability

3. Data quality elements

3.1. Overview

- Completeness
- **Logical consistency**
  - Conformity to logical rules of data structure (conceptual, logical, physical), attribution and relationships.
    - ✓ **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
- Thematic accuracy
- Temporal quality
- Usability

3. Data quality elements

3.1. Overview

- Completeness
- Logical consistency
- **Positional accuracy**
  - Accuracy of the position of features within a spatial reference system.
    - ✓ **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 position values to values accepted as or being true.
- Thematic accuracy
- Temporal quality
- Usability

3. Data quality elements

3.1. Overview

- Completeness
- Logical consistency
- Positional accuracy
- **Thematic accuracy**
  - Accuracy of quantitative attributes and the correctness of non-quantitative attributes and of the classifications of features and their relationships.
    - ✓ **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 correctness**: closeness of the value of a quantitative attribute to a value accepted as or known to be true.
- Temporal quality
- Usability

3. Data quality elements

3.1. Overview

- Completeness
- Logical consistency
- Positional accuracy
- Thematic accuracy
- **Temporal quality**
  - Quality of the temporal attributes and temporal relationships of features.
    - ✓ **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

3. Data quality elements

3.1. Overview

- Completeness
- Logical consistency
- Positional accuracy
- Thematic accuracy
- Temporal quality
- **Usability**
  - Specific quality information about a dataset suitability for a particular application or conformance to a set of requirements.
  - Based on specific user requirements that cannot be described using the other quality elements.

3. Data quality elements

3.2. Descriptors

➥ **Measure reference:**

- The type of evaluation (standardized, user defined, catalogue);
- Standardized: number of missing items, mean value of positional uncertainties, root mean square error.

➥ **Evaluation method:**

- The procedure used to evaluate the measure.

➥ **Result:**

- The output of the evaluation:
  - ✓ Quantitative result;
  - ✓ Conformance result;
  - ✓ Descriptive result.

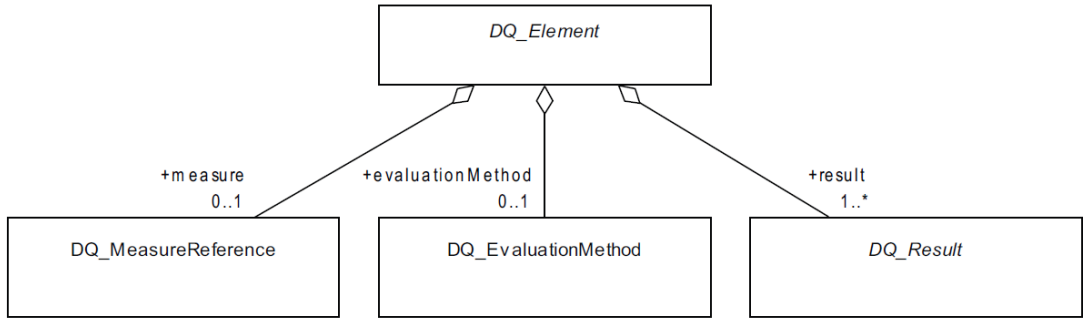


Figure 5 — Data quality element descriptors

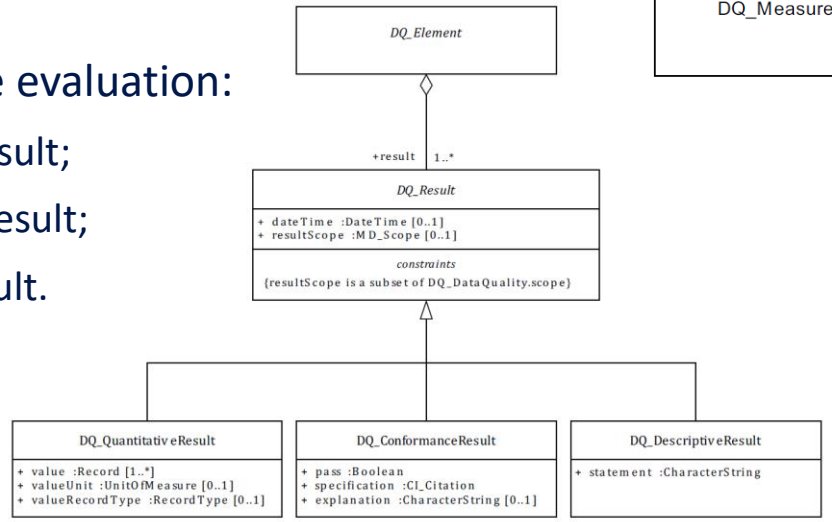


Figure 8 — Data quality result

3. Data quality elements

3.3. Metaquality

➥ Metaquality elements:

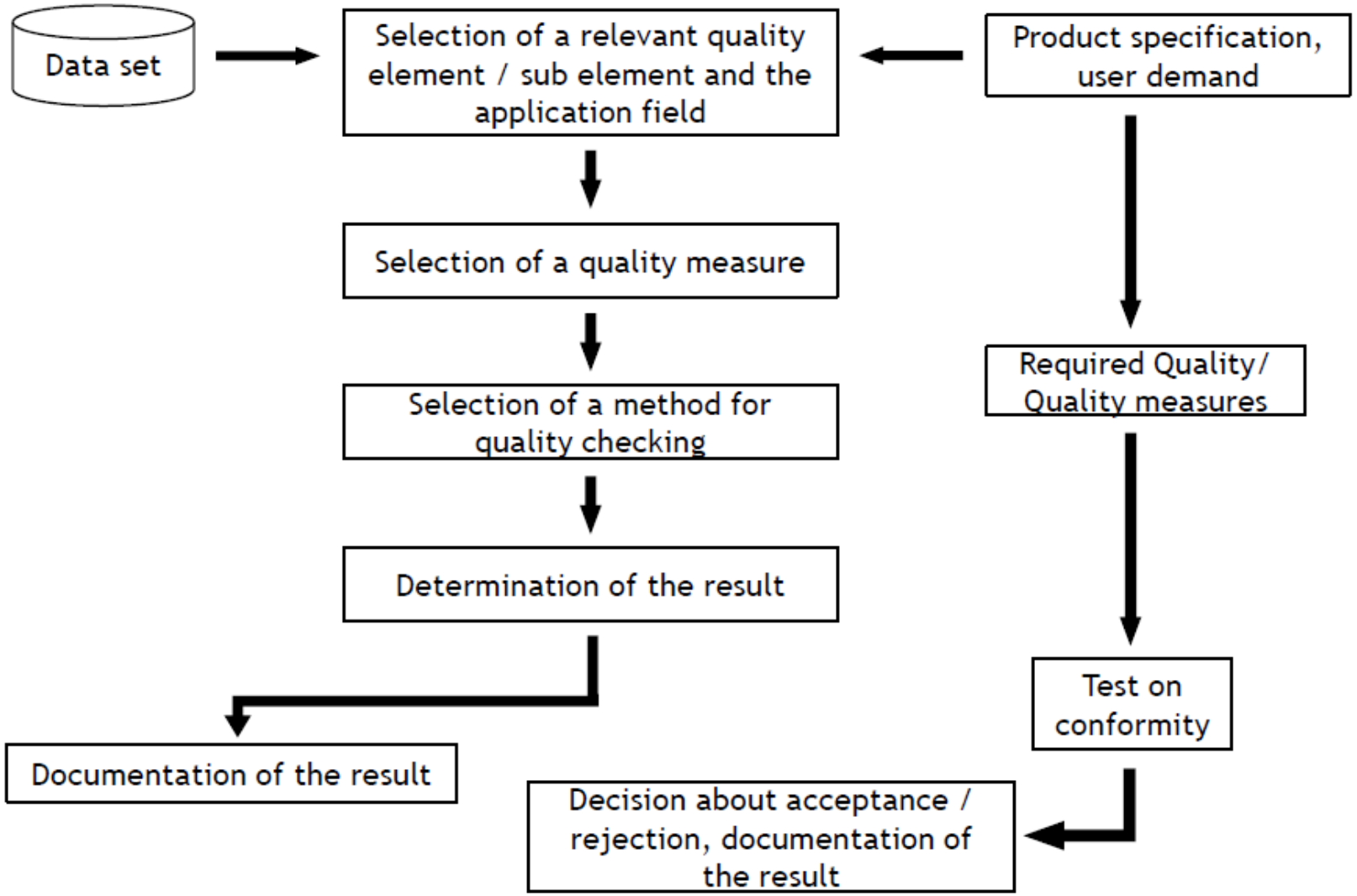
- Set of quantitative and qualitative statements about a quality evaluation and its result.
- The knowledge about the quality and the suitability of the evaluation method, the measure applied, and the given result may be of the same importance as the result itself.
  - ✓ Confidence: trustworthiness of a data quality result;
  - ✓ Representativity: degree to which the sample used has produced a result which is representative of the data within the data quality scope;
  - ✓ Homogeneity: expected or tested uniformity of the results obtained for a data quality evaluation.

➥ Comments:

- Statistical methods for the quality control of geospatial data are only acceptable under certain conditions;
- Checks based on samples are applied if this data is not used for decisions where the health or life of human beings could be endangered. In such cases the complete dataset shall always be checked.
- Checks based on samples require data with a homogenous error distribution.

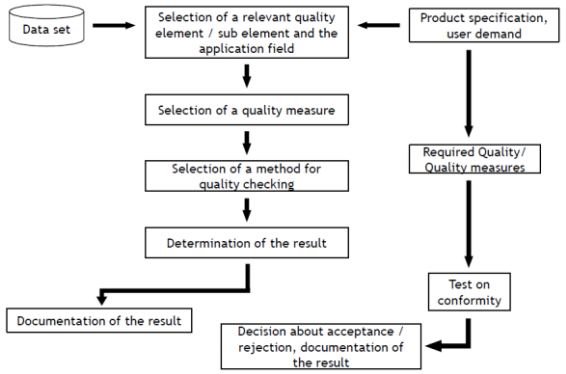
4. Data quality evaluation

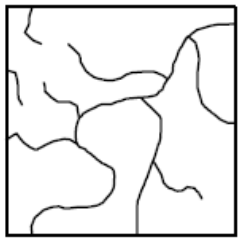
4.1. Process flow



4. Data quality evaluation

4.1. Process flow – example



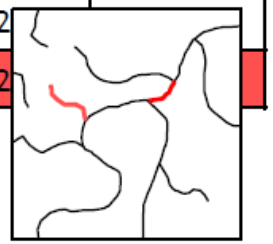


Selection of a quality element, e.g. Completeness (attribute "Width"),

Feature type:	River / water
Attribute:	Width
...	Real value (meter)
...	[-55555] unknown

check of the attribute of all objects (river segments)

#ID	width
223	33
2	
2	



"width" >/= 0, count no. of objects not fulfilling this criteria

Two river segments without attribute value

DQ_QuantitativeResult	
value	2
...	...

In this case: value=2
→ Decision: rejection

Goal: error rate = 0

Test on conformity
→ requirement: number of river segments without attribute value=0?

4. Data quality evaluation

4.2. Evaluation methods

➥ Direct evaluation:

- Evaluate the quality of a dataset based on **inspection of the items within the data set**.
  - ✓ Full inspection: tests every item in the population;
  - ✓ Sampling: testes are performed on subsets of the population.

➥ Indirect evaluation:

- Evaluate the quality of a dataset based on external knowledge or experience of the data product (can be subjective, e.g.: estimated from knowledge of the sources, tools and methods used).

➥ Aggregation and derivation:

- Aggregating or deriving existing results without carrying out a new data quality evaluation.
- Aggregation combines quality results from data quality evaluations based on different data quality elements or different data quality scopes.

5. Data quality reporting

5.1. General

➥ **Reporting:**

- Data quality shall be reported as metadata.
- A standalone quality report may be additionally created, to **provide more details** than reported as metadata.
- The standalone quality report shall not replace the metadata and the metadata should provide a reference to the standalone quality report when it exists.