

Data Quality-Flags

Background & Motivation

Scientific data may comprise data-points of very different quality for different reasons (e.g. sensor malfunction, failed calibration, high measurement error). For their analysis it is essential to classify data-points, for example, as reliable, questionable, or even wrong value. This can be done by assigning numerical codes known as *data quality-flags*.

The quality of an individual data-point is classified according to *quality-flags* — integer numbers from value 0 upwards, starting with the 6 quality-flags most commonly needed:

| quality flag | |
|--------------|---|
| 0 | unknown no information on data-quality available |
| 1 | good data certainly good data-point |
| 2 | probably good data likely good data-point - few tests pending |
| 3 | questionable data data-point probably needs to be corrected |
| 4 | bad data do not use this data-point |
| 5 | value changed data-point has been corrected - now good |

Each value within a data-set should be accompanied by one of the data quality-flags, even if no sanity checks have been performed so far. In this case use the placeholder quality-flag 0 to indicate unchecked data.

The meta-data describing the data-set should inform about how the data have been tested (e.g. automatic outlier detection, manual comparison with control measurements, published procedures). If a data-point is valid according to all testing procedures, it is flagged as good data (quality-flag 1). In cases where a data-point passed some tests (e.g. automatic procedures), but some are still to be conducted (e.g. manual control), the data-point is probably good (quality-flag 2). If, however, a data-point passes all tests except non-critical ones, it should be considered questionable (quality-flag 3). In cases where at least one critical test fails, the data-point is bad (quality-flag 4). When possible, questionable data-points (quality-flag 3) should be corrected. While retaining the original value in the data (quality-flag 3) or meta-data, the corrected value is added and flagged accordingly (quality-flag 5).

Preliminary analyses might use all data-points except those flagged bad (quality-flag 4). In contrast, final analyses should exclusively rely on good and corrected data (quality-flags 1 and 5, respectively).

Other quality-flags encode less common conditions like exceeding the measurement range (quality-flags 6 and 7) or measurements for which a value is missing for various reasons (quality-flag 9). Please refer to the following table for more details and further quality-flags.

| quality flag | definition | description |
|--------------|-----------------------|---|
| 0 | unknown | Information on data quality unavailable. No quality test performed on this data-point. |
| 1 | good data | Passed all documented quality tests. This data-point is of good quality. |
| 2 | probably good data | Data-point has not been checked by all documented quality tests. All quality tests applied indicate that this point has good quality. |
| 3 | questionable data | Data-point probably bad: failed non-critical documented test or subjective validation. This data-point must not to be used without further checking and e.g. scientific correction. |
| 4 | bad data | Failed critical documented quality test(s) or as assigned by the data-producer. This data-point is not correctable. Do not use this data-point. |
| 5 | value changed | Data-point needed correction. Value adjusted during quality control — now good. Original value flagged questionable or reported in meta-data. |
| 6 | value smaller | Signal below detection limit of sensor / method used. True value equal or less than reported value. |
| 7 | value larger | Signal above detection limit of sensor / method used. True value equal or larger than reported value. |
| 8 | interpolated value | Data-point was missing, but has been interpolated from neighbouring data-points (in space or time). |
| 9 | missing value | The data-point is missing. The reported value has no meaning. |
| 10+ | individual definition | Refer to meta-data for assigned meaning. |

If none of the ten predefined quality-flags is suitable, any integer from 10 onwards can be used to define individual data-flags. The interpretation of individual quality-flags must be documented in the accompanying meta-data.

Technical Information The quality-flags in this document are in line with international guidelines like the Ocean Data Standards recommendations and they relate to those used in projects like SeaDataNet, PANGAEA, and COSYNA. Definitions were merged and missing descriptions have been meaningfully filled.