Waterbase – Rivers
Version 14

Quality control documentation

13 June 2014

Waterbase – Rivers

Data on quality of water in rivers are collected annually through the WISE-SoE data collection process. Data and information obtained through the WISE-SoE data collection process are primarily used to compile indicator factsheets, associated with the EEA's Core Set Indicators, upon which EEA assessment reports are based. Data collected through the WISE-SoE data collection process are also published in Waterbase, a series of water topic-specific databases and web pages, publicly accessible via the EEA Data Service's web site.

Dataset contains data on nutrients, organic matter, hazardous substances and other chemical determinands in water, proxy pressure data on the upstream catchments and physical characteristics of the WISE-SoE river monitoring stations and since 2013 also biological data.

QA/QC activities

This document briefly presents the ETC-ICM (former ETC Water) and the EEA activities focused on quality of the Waterbase – Rivers dataset and the results of these activities. The Quality control tests have been performed on the Waterbase – Rivers database provided in 13 June 2014 by ETC-ICM. This database is included in the EEA data service as version 14, and is publicly available. The database and metadata are available at the following URL: http://www.eea.europa.eu/data-and-maps/data/waterbase-rivers-10

Waterbase – Emisions to water dataset contains five data tables:

- STATIONS
- PRESSURES
- NUTRIENTS
- HAZSUBS
- SUPPDET
- BIOLOGY

Following main types of the tests have been performed on the data tables. Mandatory value and Measurement value tests, Primary key/Duplicate tests, Logical rules violation test, Outlier detection tests, Chemical rules violation tests, Stations tests and Data definition compliance test.

Summary

Summary of deliveries and dataset is available in the Waterbase_Rivers_v14_QAdocument_Summary.xls file (a part of the archive that was containing also this file)
Note: The summary is not fully finished yet. The document will be updated once it is done.

1. Mandatory values tests

Mandatory values have to be present in each of the records. Records where any of these values is missing are excluded from the dataset:

- STATIONS: Country Code, NationalStationID
- PRESSURES: Country Code, NationalStationID
- NUTRIENTS: Country Code, NationalStationID, Year, Aggregation Period, Determinand Nutrients
- HAZSUBS: Country Code, NationalStationID, Year, Determinand HazSubs
- SUPPDET: Country Code, NationalStationID, Year, Determinand Suportive
- BIOLOGY: Country Code, NationalStationID, Year, Aggregation Period, Determinand Biology

1.1 Measurement value tests

Mean values in all three tables containing the determinand concentrations are subject of this test. Detected issues are then stored as a code in a special QA field (QA_MVissues) as follows:

101 – the Mean value is missing

102 – the Mean value is negative and negative values are not allowed or possible

103 – the Mean value is equal 0 and 0 values are not allowed or possible

180 – the value was not requested (e.g. wrong water category)

181 – MeanValueEQR was reported in wrong scale

182 – MeanValueNormEQR cannot be caluculated from the reported data

Records flagged with any of these flags either can’t be used (101) or it is recommended that they are excluded from further use or analysis (102, 103).

2. Primary key tests

Primary key is a field or combination of fields with values which have to be unique in the data table. If primary key is duplicated it is an error which has to be solved or the records are excluded from the dataset.

List of data tables primary keys:

- STATIONS: Country Code, NationalStationID
- PRESSURES: Country Code, NationalStationID
- NUTRIENTS: Country Code, NationalStationID, Year, Aggregation Period, Determinand Nutrients
- HAZSUBS: Country Code, NationalStationID, Year, Determinand HazSubs
- SUPPDET: Country Code, NationalStationID, Year, Determinand Suportive
- BIOLOGY: Country Code, NationalStationID, Year, Aggregation Period, Determinand Biology

3. Logical rules violation tests

The following logical rules were tested in “NUTRIENTS” and “HAZSUBS” data tables:

201 – Mean >= Minimum

202 – Mean <= Maximum

203 – Median >= Minimum

204 – Median <= Maximum

205 – Minimum <= Maximum

206 – If Minimum > 0 Then StandardDeviation < Maximum

207 – If Minimum < Maximum Then StandardDeviation > 0

210 – All measurement values >= 0 (exceptions: Alkalinity, Temperature)

211 – If NumberOfSamples = 1 Then (Mean = Minimum = Maximum = Median)

212 – If NumberOfSamples = 1 Then StandardDeviation = 0

213 – If NumberOfSamples = 0 Then all measurement values are Null

217 – NumberOfSamplesBelowLOQ <= NumberOfSamples

221 – If(Maximum < LimitOfQuantification) Then NumberOfSamples = NumberOfSamplesBelowLOQ

222 – If(Maximum < LimitOfQuantification) Then Mean ~ LimitOfQuantification /2

A special QA field (QA_LRviolations) has been added to the data tables. Information on the rules violated in the respective record is kept in this field as a coma separated list of those rules codes (the codes are the same as the numbers of the rules above above). It is recommended that the records where QA_LRviolation field is not empty should not be used in a further analysis or only after a careful consideration. The detected data quality inconsistencies will be tried to be solved in the near future.

4. Chemical rules violation tests

Chemical rules were tested in the “NUTRIENTS” data table. Following chemical rules were defined between Mean concentrations of certain related determinands from the same monitoring station, year and aggregation period:

310 – COD * 1.01 > BOD5

320 – Total Phosphorus * 1.01 >= Orthophosphate

330 – Total Nitrogen * 1.05 >= Kjeldahl Nitrogen + Nitrate + Nitrite

340 – Kjeldahl Nitrogen * 1.05 >= Organic Nitrogen + Total Ammonium

350 – Total Oxidised Nitrogen * 1.05 >= Nitrate + Nitrite

The 1.01 and 1.05 tolerance was applied in order to avoid situations when records are detected only due to value differences caused by rounding.

Due to the fact that the concentration values in the tested records are results of aggregation and it is possible that the individual measurements participating in the aggregation could be taken for different determinands in different times and/or by different methods, thus the discrepancies in concentrations are justified because the particular records are not comparable, following additional testing steps were taken to filter out incomparable records.

1 – comparison of Number of Samples

2 – comparison of Aggregation Length and Aggregation Months values

3 – comparison of Limit of Detection and Limit of Quantification values

4 – considering clarifications and approvals of chemical rules violations that were provided by data reporters

The details about the individual tests can be found at https://taskman.eionet.europa.eu/ETCW/ticket/86#comment:7.

A special QA field (QA_CRviolations) has been added to the data table. Information on the rules violated in the respective records (all records of each of the determinands from both sides of formula) is kept in this field as a coma separated list of those rules numbers (the numbers are the same as in the table above).

Records that were detected by the first step of the test but have been found incomparable in any of the further steps are also flagged with the code of the test but “0” is replaced by the number of the step as used above (e.g., 311, 324, 332). Such records are then treated as not violating the chemical rules.

It is recommended that the records flagged in the QA_CRviolations with the any of the codes ending with “0” are not used in further analysis or only after a careful consideration. The detected data quality inconsistencies will be tried to be solved in the near future.

5. Outlier detection tests

Detection of outliers was performed on the Mean values in the “Nutrients” and “SuppDet” data table.

Different methods of outlier detection were used, from simple comparison of measurement value with the defined limit value for particular determinand, to more complex statistical tests.

Sometime the whole time series where the measurement values are naturally very high (e.g. because of the positioning of the monitoring station close to the source of the pollution) have been also detected.

Some of previously detected errors have been already corrected by countries or were approved as natural high/low values.

A special QA field (QA_outlier) has been added to the tables and records, where the any of the situations mentioned above has been detected, have been flagged in this field as follows:

401 – Standard potential outlier - value is either higher/lower than limit value or is suspiciously high/low comparing to the rest of the time series or value change between two consecutive values is suspiciously abrupt or is marked as an outlier by a content expert

402 – Measurements are probably taken from a highly polluted locations but information was not confirmed

403 – The whole country delivery is considered as problematic because it contains too many quality issues

410 – The record previously detected as an outlier has been corrected (is not an outlier anymore)

491 – Outlier has been confirmed by country as correct value

492 – Outlier has been confirmed by (ETC) content expert as correct value

493 – Measurement has been confirmed by country to be taken from a highly polluted area

It is recommended that the records where QA_outlier field contains codes 401-403 and eventually also code 493, should not be used in a further analysis or only after a careful consideration. The detected data quality inconsistencies will be tried to be solved in the near future.

6. Stations tests

A number issues in stations records and in related records in other tables are checked in these tests. A special QA_field (QA_station_issues) was added to all data tables where these issues, if detected, are indicated by appropriate flag as follows:

500 station coordinates fall slightly outside the respective country boundary, but were confirmed as correct by country

501 station coordinates fall outside the respective country boundary

502 one or both station coordinates are missing

503 more stations with the same coordinates (if it might indicate an error)

505 Altitude is suspicious or illogical

511 Water Category value is incompatible with this particular dataset

512 station coordinates fall outside the respective River Basin District

513 Catchment Area is suspicious or illogical

599 station is not defined in the station table

These issues should be taken into the account in further use and analysis of the data. The detected data quality inconsistencies will be tried to be solved in the near future.

7. Data definition compliance tests

All dataset values have to follow specifications defined in the respective Data dictionary. Values, which are of a different data type than requested (e.g. string instead of numeric) or which are not available in a set of allowable values, have been either removed or, if possible, replaced by a correct value. The original, incorrect value has been stored in a special QA field (QA_DDviolations) in the following format:

Name_of_field: Erroneous_Value; [Name_of_field: Erroneous_Value; …]

This field serves as an indication why some of the values are missing, as a reference for solving similar problems in the future or in certain cases as background information for future update of Data dictionary.