Waterbase – TCM
Version 10

Quality control documentation

18 July 2013

Waterbase – Transitional, coastal and marine waters

Data on quality of water in TCM waters are collected annually through the WISE-SoE data collection process and through various marine conventions. Data and information obtained through the 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. Collected data 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 physical characteristics of the transitional, coastal and marine water monitoring and flux stations, proxy pressures on the upstream catchment, basin and River Basin District associated with transitional and coastal waters, chemical quality data on nutrients in seawater and hazardous substances in biota, sediment and seawater, as well as data on direct discharges and riverine input loads.

QA/QC activities

This document briefly presents the ETC-ICM (former ETC Water) and the EEA activities focused on quality of the Waterbase – TCM dataset and the results of these activities.
The Quality control tests have been performed on the Waterbase – TCM database provided in 11 July 2013 by ETC-ICM. This database is included in the EEA data service as version 10, and is publicly available. The database and metadata are available at the following URL: http://www.eea.europa.eu/data-and-maps/data/waterbase-transitional-coastal-and-marine-waters-9

Waterbase – TCM dataset contains 10 data tables:

- STATIONS_EIONET
- STATIONS_CONVENTIONS
- STATIONS_FLUX
- PRESSURES
- NUTRIENTS_DISAGG
- HAZSUBS_SEAWATER_DISAGG
- HAZSUBS_SEDIMENT_DISAGG
- HAZSUBS_BIOTA_DISAGG
- INPUTS
- DISCHARGES

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, Stations tests and Data definition compliance test.

Summary

Summary of deliveries and dataset is available in the Waterbase_TCM_v10_QAdocument_Summary.xls file (a part of the archive that was containing also this file)

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_EIONET: Country Code, NationalStationID
- STATIONS_CONVENTIONS: Country Code, NationalStationID
- STATIONS_FLUX: Country Code, NationalStationID
- PRESSURES: Country Code, NationalStationID
- NUTRIENTS_DISAGG: Country Code, NationalStationID, Year, Determinand Nutrients
- HAZSUBS_SEAWATER_DISAGG: Country Code, NationalStationID, Year, Determinand HazSubs
- HAZSUBS_SEDIMENT_DISAGG: Country Code, NationalStationID, Year, Determinand HazSubs
- HAZSUBS_BIOTA_DISAGG: Country Code, NationalStationID, Year, Determinand HazSubs
- INPUTS: Country Code, NationalStationID, Year, Determinand InputDischarge
- DISCHARGES: Country Code, NationalStationID, Year, Determinand InputDischarge

1.1 Measurement value tests

Load values (Input, Discharges), Concentration values (HazSubs_Seawater_disagg, HazSubs_Sediment_disagg, HazSubs_Bioat_Disagg) and Mean values (HazSubs_Bioata_Agg) are subject of this test. Detected issues are then stored as a code in a special QA field (QA_MVissues) as follows:

101 – the measurement value is missing

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

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

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_EIONET: Country Code, NationalStationID, Data Source
- STATIONS_CONVENTIONS: Country Code, NationalStationID, Data Source
- STATIONS_FLUX: Country Code, NationalStationID, Data Source
- PRESSURES: Country Code, NationalStationID
- NUTRIENTS_DISAGG: Country Code, NationalStationID, Year, Month, Day, SampleID, Determinand HazSubs, SampleDepth
- HAZSUBS_SEAWATER_DISAGG: Country Code, NationalStationID, Year, Month, Day, SampleID, Determinand HazSubs, SampleDepth
- HAZSUBS_SEDIMENT_DISAGG: Country Code, NationalStationID, Year, Month, Day, SampleID, Determinand HazSubs, Fraction, SedimentTop, SedimentBottom, Basis
- HAZSUBS_BIOTA_DISAGG: Country Code, NationalStationID, Year, Month, Day, SampleID, Determinand HazSubs, Species, Tissue, Basis
- HAZSUBS_BIOTA_AGG: Country Code, NationalStationID, Year, Aggregation Period, Determinand HazSubs, Species, Tissue, Basis
- INPUTS: Country Code, NationalStationID, Year, Determinand InputDischarge, Estimate, Method
- DISCHARGES: Country Code, NationalStationID, Year, Discharge type, Determinand InputDischarge, Estimate, Method

Result:

1206 NUTRIENTS_DISAGG records (PL) that has been detected in the duplicates test have been left in the dataset. The duplicated (or multiplicated) records have same primary key values except for SampleID which is not provided. They also have different concentration values which indicate that they are not duplicates but different samples.

3. Logical rules violation tests

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

251 – If Basis = D then DryWetRatio% is not NULL

252 – Dry WetRatio% >=1

253 – Fat <= DryWetRatio%

A special QA field (QA_LRviolations) has been added to the data tables. Information about the rules violated in the respective record is kept there as a coma separated list of the violated rule codes (the codes are the same as the rule numbers 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. Outlier detection tests

Detection of outliers was performed on the measurement values in the “NUTRIENTS” and “HAZSUBS” tables.

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) can be also detected.

Some of previously detected errors could already be 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

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.

5. 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)

599 station is not defined in the stations 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.

6. 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.