Waterbase – TCM
Version 7

Quality control documentation

25 June 2010

Waterbase – TCM

Data on transitional, costal and marine waters 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 WISE map viewer, Waterbase, a series of water topic-specific databases and web pages, publicly accessible via the EEA Data Service's web site.

The dataset contains data on nutrients and organic matter, proxy pressure data on the upstream catchments and physical characteristics of the WISE-SoE TCM monitoring stations.

QA/QC activities

This document briefly presents the 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 June 2010 by ETC/WTR. This database is included in the EEA data service as version 7, 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-6

Waterbase – TCM dataset contains 10 data tables:

- DISCHARGES
- FLUX STATIONS
- HAZSUBS BIOTA
- HAZSUBS SEDIMENT
- HAZSUBS WATER
- INPUTS
- PROXY PRESSURES
- QUALITY
- STATIONS EIONET
- STATION ICES MEDPOL

Five types of the tests have been performed on the data tables. Basic tests, Logical rules violation test, Outlier detection, Stations tests, and Valid data type and codes tests.

Chemical rules tests, that were first time introduced previous year, were also tested by the ETC/Water but the result were not incorporated due to critical comments obtained from countries as a reaction to number of false positive errors detected by the tests. The rules are presently being revised.

1. Basic tests

1.1 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.

List of data tables primary keys:

- DISCHARGES: CountryCode, WaterBodyID, WaterBodyName, Year, Determinand, Method
- FLUX STATIONS: CountryCode, WaterbaseID
- HAZSUBS BIOTA: CountryCode, WaterbaseID, Year, Month, Day, SampleID, Determinand, Species, Tissues
- HAZSUBS SEDIMENT: CountryCode, WaterbaseID, Year, Month, Day, SampleID, Determinand, Fraction, SedimentTop, SedimentBottom
- HAZSUBS WATER: CountryCode, WaterbaseID, Year, Month, Day, SampleID, Determinand, SampleDepth
- INPUTS: CountryCode, WaterbaseID, Year, Determinand, Estimate
- PROXY PRESSURES: CountryCode, WaterbaseID
- QUALITY: CountryCode, WaterbaseID, Year, Month, Day, SampleID, Determinand, SampleDepth
- STATIONS EIONET: CountryCode, WaterbaseID
- STATION ICES MEDPOL: CountryCode, WaterbaseID

Result:
Many duplicates were detected during the QA process in most of the tables. Three methods were used to handle them:

- The true duplicates (e.g. data entered/reported twice) have simply been removed. Checking was made that these data are true by looking at e.g. the concentrations as well - the number of duplicates with and without concentration should be the same
- The non-true duplicates have been marked in the working database in order to identify them for further questions to the submitting countries. These duplicates have not been extracted for Waterbase. The cause of these duplicates is most probably erroneous use of the combination of sample number and subsample number in the ICES data submissions.
- Duplicates detected in the Discharges and Input table have been left in the dataset. A special QA field – QA_duplicates (boolean type) – has been however added to the tables and all duplicates have been marked here

1.2 Mandatory values tests

Each table has defined a set of fields where values has to be provided be provided otherwise records can not be properly used. A special QA field – QA_MVmissing – was added to the most of the data tables. If a record is missing a mandatory value in some of the fields, name of this field was added to the QA field for fast inentification of the problem.

1.3 Measurement value errors tests

Validity and availability of measurement values were tested in the relevant tables:

Concentration – HazSubs Biota, HazSubs Sediment, HazSubs Water, Quality
Load – Discharges, Inputs

Two special QA fields – QA_concentration_error and QA_load_error – were added to the tables. Records where measurement value is not valid or is missing are flagged in the respective QA field as follows:

10 – measurement value is 0

11 – measurement value is negative

99 – measurement value is missing

1.4 Table relations tests

The unique Waterbase identifier (WaterbaseID) is present in each of the data tables. It can be used to link data from one table to another. The table relations tests detect identifiers which are not present in some of the tables. Records that do not have station counterpart in the respective stations tables have been flagged in the QA_station_problem field (see section 3.)

2. Outlier detection

Detection of outliers was performed on the “HAZSUBS” and “QUALITY” data.

Concentration values were tested against limiting values individually defined by an expert for each of the determinands. Records where values are higher then the limit are flagged in a special QA field (QA_outlier) added to tables. Following QA flag have been used:

1 – standard potential outlier - value is higher than limit value

3. Stations tests

Positions of all reported monitoring stations have been tested using the coordinates provided as well as stations availability. The cases when the station coordinates fall outside the respective country borders, when coordinates are missing or when the monitoring station is not available in the respective stations table, are documented in a special QA field (QA_station_problem). In addition some other station related issues were tested. Following QA flags have been used:

1 – monitoring station is located outside the respective country borders

99 – station is not available in the Stations table

These data quality inconsistencies will be tried to be solved in the near future.

4. Other specific tests

In some of the TCM data tables a set of specific rules can be defined. Records where these rules have been violated are flagged in a special QA field (QA_other_errors) as follows:

11 – violated rule - Basis = D and DryWetRatio% null (tables HazSubs Biota, HazSubs Sediment)

12 – violated rule - DryWetRatio% < 1 (table HazSubs Biota)

13 – violated rule - Fat > DryWetRatio% (table HazSubs Biota)

5. Data type and codes tests

All TCM dataset values have to follow specifications defined in the respective Data dictionary (DD) definitions. Values, which are of a different data type as 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.

There is one exception from this rule. Some of these “errors” are only formally wrong. The value is still valid but was not foreseen as possible and was therefore omitted to be included in the current DD definitions of the respective table field. In this case the original code has been left in the field untouched. It is planed that these codes will be added into the code list during the next DD update.

In all the cases the original, incorrect value or value missing in the DD code list, has been stored in a special QA field (QA_datatype_error) in the following format:

Name_of_field: Erroneous_Value; [Name_of_field: Erroneous_Value; …]

The cases where the errors couldn’t be corrected will be tried to be solved in the near future in cooperation with the data providers.