



# **Waterbase – Emissions to water**

## **Version 3**

---

**Quality control documentation**

**19 June 2012**

## Waterbase – Emissions to water

Data on emissions to water 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 selected from reporting of member and collaborating countries on emissions of nutrients and hazardous substances to water, aggregated within River Basin Districts.

## QA/QC activities

This document briefly presents the ETC-ICM (former ETC Water) and the EEA activities focused on quality of the Waterbase – Emissions to water dataset and the results of these activities.

The Quality control tests have been performed on the Waterbase – Emissions to water database provided in 19 March 2011 by ETC-ICM. This database is included in the EEA data service as version 2, and is publicly available. The database and metadata are available at the following URL:

<http://www.eea.europa.eu/data-and-maps/data/waterbase-emissions-2>

Waterbase – Emissions to water dataset contains five data tables:

- SPATIAL\_UNITS
- NUTRIENTS\_POINTS
- NUTRIENTS\_DIFFUSE
- HAZSUBS\_POINTS
- HAZSUBS\_DIFFUSE

Five 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, Stations tests and Data definition compliance test.

## **Summary**

Summary of deliveries and dataset is available in the Waterbase\_Emissions\_v3\_QAdocument\_Summary.xls file (a part of the zip 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:

- SPATIAL\_UNITS: Country Code, Spatial Unit Code, Spatial Unit Name, Spatial Unit Category, River Basin District Code or River Basin District Name
- NUTRIENTS\_POINTS: Country Code, Spatial Unit Code, Spatial Unit Name, Spatial Unit Category, Period - Years, Nutrient Determinand ID, Source of Point Emissions
- NUTRIENTS\_DIFFUSE: Country Code, Spatial Unit Code, Spatial Unit Name, Spatial Unit Category, Period - Years, Nutrient Determinand ID, Source of Diffuse Emissions
- HAZSUBS\_POINTS: Country Code, Spatial Unit Code, Spatial Unit Name, Spatial Unit Category, Period - Years, Hazardous Substance Determinand ID or Hazardous Substance Determinand Name, Source of Point Emissions
- HAZSUBS\_DIFFUSE: Country Code, Spatial Unit Code, Spatial Unit Name, Spatial Unit Category, Period - Years, Hazardous Substance Determinand ID or Hazardous Substance Determinand Name, Source of Diffuse Emissions

### 1.1 Measurement value tests

Emission values in all four tables are subject of this test. Detected issues are then stored as a code in a special QA field (QA\_MVissues) as follows:

- 101 – the Emission value is missing
- 102 – the Emission value is negative and negative values are not allowed or possible
- 103 – the Emission value is equal 0 and 0 values are not allowed or possible

In a case that Emission values have been provided under WFD, UWWTD, E-PRTR or SoE (Riverine Load or Direct Discharges) reporting, the reporters can use value -1, -2, -3 or -5 respectively instead of the real Emission value. If the real Emissions values could not be provided by checking the dataset of the above mentioned reportings, the reason is indicated in the QA\_MVissues field as follows:

- 161 – the real value of Emissions was checked out, but it is not available in relevant dataset
- 162 – the relevant dataset containing the real value of Emissions does not exist at all or it is not available
- 163 – the relevant dataset containing the real value of Emissions was not finished or not provided yet => there is a chance to get the real value in the future
- 164 – the real value of Emissions was not checked out yet

In addition if a) the value of emissions is not known or not available, or b) emissions are not relevant or not significant, then reporters use value (a) -8 or (b) -9 instead of the real Emission values.

## 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:

- SPATIAL\_UNITS: Country Code, Spatial Unit Code
- NUTRIENTS\_POINTS: Country Code, Spatial Unit Code, Period - Years, Nutrient Determinand ID, Source of Point Emissions
- NUTRIENTS\_DIFFUSE: Country Code, Spatial Unit Code, Period - Years, Nutrient Determinand ID, Source of Diffuse Emissions
- HAZSUBS\_POINTS: Country Code, Spatial Unit Code, Period - Years, Hazardous Substance Determinand ID or Hazardous Substance Determinand Name, Source of Point Emissions, EPRTR Facility
- HAZSUBS\_DIFFUSE: Country Code, Spatial Unit Code, Period - Years, Hazardous Substance Determinand ID or Hazardous Substance Determinand Name, Source of Diffuse Emissions

### 3. Logical rule violation tests

The following logical rules were tested in all emissions data tables:

261 –  $NP\_calculated = NP1\_reported + NP2\_reported + NP3\_reported + NP4\_reported + NP5\_reported + NP6\_reported + NP7\_reported + NP8\_reported$

262 –  $U1\_calculated = U11\_reported + U12\_reported + U13\_reported + U14\_reported$

263 –  $U2\_calculated = U21\_reported + U22\_reported + U23\_reported + U24\_reported$

264 –  $U\_calculated = U1\_reported + U2\_reported$

265 –  $I\_calculated = I3\_reported + I4\_reported$

266 –  $O\_calculated = O5\_reported + O6\_reported$

267 –  $PT\_calculated = U\_reported + I\_reported + O\_reported$

Detailed breakdown of the rules with additional description is available in the Waterbase\_Emissions\_v3\_QAdocument\_LRdescription.pdf (a part of the zip archive that was containing also this file).

A special QA field (QA\_LRviolations) has been added to the data tables. Information of the rules violated in the respective record are kept there 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. Stations tests**

Spatial unit code of each of the records in any of the emission data tables should be present in the Spatial\_units table. If it is not true the emission record is flagged in a special QA\_field (QA\_station\_issues) witch code 599.

No such issues were detected in this dataset.

## 5. 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 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. 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; ...]*

No such issues were detected in this dataset.