Waterbase – Water Quantity
Version 7

Quality control documentation

24 April 2012

Waterbase – Water Quantity

Data on water quantity 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 all time series related to water quantity data, reported by member and collaborating countries. Reported data have been assessed and processed by the ETC-ICM and the EEA. Results of quality assessment have been incorporated into the individual data tables.

QA/QC activities

This document briefly presents the ETC-ICM (former ETC Water) and the EEA activities focused on quality of the Waterbase – Water quantity dataset and the results of these activities.
The Quality control tests have been performed on the Waterbase – Water Quantity database provided in 22 March 2012 by ETC-ICM. 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-water-quantity-6

Waterbase – Water Quantity dataset contains nine data tables:
- GENTITIES
- RAIN_GAUGE_STATIONS
- RESERVOIRS
- STREAM_FLOW_STATIONS
- WELLS
- LARGE_ITEMS
- AREAS
- WATER_ABSTRACTION_SOURCES
- TIME_SERIES

For reporting the data under the WISE-SoE Water Quantity data collection process the data reporters use a dedicated reporting tools developed by the ETC-ICM. The tool automatically control data on the entry level and assures that the errors, which often haunts the other WISE-SoE datasets, are not entered into the dataset. The following QA tests are performed automatically:
- Mandatory values tests
- Data definition compliance tests
- Outlier tests
- Logical rule tests
- Areal rules tests
- Cross-table relation tests

The tests that are not performed automatically are:
- Primary key/Duplicates tests
- Station coordinates tests

Summary

Summary of deliveries and dataset is available in the Waterbase_Water_uantity_v7_QAdocument_Summary.xls file (a part of the archive that was containing also this file)

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 or the records are excluded from the dataset.

List of data tables primary keys:

- GENTITIES: CountryCode, GEntityCode, GEntityType
- RAIN_GAUGE_STATIONS: CountryCode, GEntityCode
- RESERVOIRS: CountryCode, GEntityCode
- STREAM_FLOW_STATIONS: CountryCode, GEntityCode
- WELLS: CountryCode, GEntityCode
- LARGE_ITEMS: CountryCode, GEntityCode
- AREAS: CountryCode, GEntityCode
- WATER_ABSTRACTION_SOURCES: CountryCode, GEntityCode, SectorCode, WaterSource
- TIME_SERIES: CountryCode, GEntityCode, Year, TimeStep, Index, VariableCode

2. Logical rules violation tests

The following logical rule was tested in “TimeSeries” table:

271 - Σ (MonthlyValues) = AnnualValue (The monthly values do not sum up to the annual value. This rule applies to time series if annual and monthly data are given. Doesn't apply to Stream flow and Groundwater level time series.)

The following logical rule was tested in “Stream flow stations” data table:

272 - MinDischarge <= MaxDischarge (Min. stream flow value should be lower or equal than Max. stream flow value)

3. Areal rule tests

The areal rules are specific logical rules that the data provided have to follow to be considered as valid. The records violating any of the rules are marked in a special QA field (QA_ALRviolations) with code of the rule.

The list of the rules together with their respective codes is provided separately in the Code list section of the dataset web page on the EEA dataservice (http://www.eea.europa.eu/data-and-maps/data/waterbase-water-quantity-6/code-lists).

4. Station coordinates tests

Point data records (Rain_Gauge_Stations, Reservoirs, Stream_Flow_Stations, Wells) that are missing coordinates values are marked in a special QA field (QA_station_issues) with flag 502.

Point data coordinates that were not reported in the requested coordinate system were excluded from the dataset.

5. Outlier detection tests

Detection of outliers was performed on the Mean values in the “TimeSeries” 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

409 – The value can’t be confirmed by data provider (the original source data are unavailable)

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 401-409 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.