Waterbase – Groundwater
Version 12

Quality control documentation

23 April 2012

Waterbase – Groundwater

Data on quality of water in groundwater 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 chemical quality of groundwater, characteristics of groundwater bodies and sampling sites. Reported data have been assessed and processed by the ETC-Water and the EEA. Disaggregated records were annually aggregated by groundwater body, substance and year, and statistic value calculated. 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 – Groundwater dataset and the results of these activities.
The Quality control tests have been performed on the Waterbase – Groundwater database provided in 28 March 2012 by ETC-ICM. This database is included in the EEA data service as version 12, and is publicly available. The database and metadata are available at the following URL:
http://www.eea.europa.eu/data-and-maps/data/waterbase-groundwater-8

Waterbase – Groundwater dataset contains four data tables:
- BODIES
- STATIONS
- QUALITY
- SALTWATER_INTRUSION

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_Groundwater_v12_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:

- BODIES: Country Code, GWBcode_WFD or GWBcode_EIONET
- STATIONS: Country Code, StationID
- QUALITY: Country Code, GWBcode_WFD or GWBcode_EIONET, Year, Determinand Code
- SALTWATER_INTRUSION: Country Code, SALTcode

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

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:

- BODIES: Country Code, Waterbase_GWBcode
- STATIONS: Country Code, StationID
- QUALITY: Country Code, Waterbase_GWBcode, Year, Aggregation Period, Determinand Name
- SALTWATER_INTRUSION: Country Code, Waterbase_SALTcode

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

A special QA field (QA_LRviolations) has been added to the data tables. Information of the rules violated in the respective record is 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. Outlier detection tests

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

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)

531 station or saltwater intrusion area can’t be linked to any of the groundwater body provided in the respective table

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.

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.