Waterbase – Rivers
Version 8

Quality control documentation

18 June 2008

Waterbase – Rivers

In the context of the implementation of the Water Framework Directive (WFD), the European Environment Agency (EEA) EIONET-Water annual data flow for waters is in the process of being transferred into the WISE 'State of the Environment' (SoE) voluntary data flow. With this it remains one of the EIONET Priority Data Flows, but gains full integration into the reporting under WISE as the single entry point of water information in Europe and complementarily with data collected under the WFD. Most information that is used for European level 'state of environment' assessments needs to be provided by member countries and there it usually comes from monitoring networks that are to meet several assessment purposes, SOE, as well as different legal requirements..

Data on rivers 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.

Rivers dataset include physical characteristics of the river monitoring stations, proxy pressures on the upstream catchments areas, as well as chemical quality data on nutrients and organic matter in rivers.

QA/QC activities

This document briefly presents the ETC/Water and the EEA activities focused on quality of the Waterbase - Rivers dataset and the results of these activities. In addition a warning is given on the use of certain records for analytical purposes (see section 2 and 3).
The Quality control tests have been performed on the Waterbase - Rivers database provided in April 2008 by ETC/WTR. This database is included in the EEA data service as version 8, and is publicly available. The database and metadata are available at the following URL:
http://dataservice.eea.europa.eu/dataservice/metadetails.asp?id=1038

A subset of the dataset is also used in the WISE (http://water.europa.eu/).

Waterbase – Rivers dataset contains three data tables:

- PRESSURES
- QUALITY
- STATIONS

Four types of the tests have been performed on the data tables. Basic tests, Logical rules violation test, Outlier detection and Station coordinates tests.

1. Basic tests

1.1 Summary

The summary presents number of records subdivided by country for each table of the dataset:

a) which were delivered by the country in the last delivery (very late deliveries are not included)
b) which was possible to process (reasons why some of the records was not possible to process are very various and comprehensive summarization is difficult)
c) total in the working database
d) which are present in the Waterbase - Rivers v8

Numbers of records excluded from the Waterbase subdivided by the reasons of the exclusion are also present.

1.1.1 Waterbase - Rivers: Quality

Country Code	Numbers of records							
	in the latest delivery		total in the working database	in the Waterbase	excluded from the Waterbase – reasons*			
	total	processed			1	2	3	4
AL			1395	1395				
AT	3114	2554	21656	21656				
BA	518	474	2208	2208				
BE	873	716	6023	6023				
BG	1057	627	9887	9881			6	
CH	64	64	766	766				
CY			97	97				
CZ	939	939	11832	11832				
DE	9042	1015	23442	23442				
DK	209	209	12417	12417				
EE	587	587	10083	10083				
ES	4021	4021	63396	63394		2		
FI	4940	4940	135347	135347				
FR	10973	10973	95846	95846				
GB			26045	26045				
GR			1833	1833				
HR	2012	1877	4900	4900				
HU	5949	4685	79573	79522			51	
IE	598	438	5894	4578		1176		140
IS	11	7	46	46				
IT			8762	8762				
LI	2		8	8				
LT	976	748	19491	19491				
LU	28	28	303	303				
LV	550	494	9610	9610				
ME			0	0				
MK	171	171	1690	1690				

MT			0	0				
NL			4353	4353				
NO	322	230	1657	1607				50
PL	1931	1931	23679	23679				
PT	767	767	767	754		13		
RO	762	761	5162	5068	2	92		
RS	987	987	3137	3137				
SE	1391	1276	36455	36455				
SI	134	134	4201	4201				
SK	762	535	6922	6915				7
TR			0	0				
Total	53690	42188	638883	637344	2	1283	57	197

*
1 – determinand is missing
2 – all of the determinand concentration fields (mean, minimum, maximum, median, standard deviation) and supportive determinands fields (Alkalinity, Conductivity, pH) are empty
3 – stations are not present in the stations table
4 – non-standard units are used, conversion to standard units is not possible

1.1.2 Waterbase - Rivers: Stations

Country Code	Numbers of records			
	in the latest delivery		total in the working database	in the Waterbase
	total	processed		
AL			30	30
AT	287	287	287	287
BA	44	44	53	53
BE	31	31	60	60
BG	110	110	110	110
CH	8	8	8	8
CY	9	9	9	9
CZ			72	72
DE			151	151
DK	42	42	42	42
EE	53	53	53	53
ES	1222	1222	1514	1514
FI	138	138	231	231
FR	1629	1629	1939	1939
GB			204	204
GR			94	94
HR	45	45	45	45
HU	99	99	101	101
IE	180	180	209	209
IS	1	1	1	1
IT			237	237
LI	1	1	1	1
LT	53	53	99	99
LU	4	4	4	4

LV	38	38	120	120
ME				
MK	20	20	20	20
MT				
NL			23	23
NO	46	46	48	48
PL	136	136	136	136
PT	59	59	59	59
RO	118	118	126	126
RS	76	76	77	77
SE	116	116	116	116
SI	21	21	30	30
SK			60	60
TR				
Total	4586	4586	6369	6369

1.1.3 Waterbase - Rivers: Pressures

Country Code	Numbers of records				
	in the latest delivery		total in the working database	in the Waterbase	excluded from the Waterbase – reasons*
	total	processed			1
AL			30	29	1
AT	289	289	289	289	
BA					
BE	31	31	59	59	
BG					
CH					
CY	9	9	9	9	
CZ			72	72	
DE			147	147	
DK			42	42	
EE	53	53	53	53	
ES	1222	1222	427	427	
FI			5	5	
FR	1569	1569	561	560	1
GB			190	190	
GR					
HR					
HU	98	98	101	101	
IE			74	74	
IS	1	1	1	1	
IT					
LI	1	1	1	1	
LT	53	53	99	99	
LU	4	4	4	4	
LV	38	38	120	120	
ME					
MK			20	20	

MT					
NL			12	12	
NO	46	46	48	48	
PL	136	136	136	136	
PT	59	59	59	59	
RO			124	124	
RS					
SE	116	116	116	116	
SI			24	24	
SK			55	55	
TR					
Total	3725	3725	2878	2876	2

*
1 – all of the pressure fields are empty (record is useless)

1.2 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.

List of data tables primary keys:
STATIONS: CountryCode, WaterbaseID
PRESSURES: CountryCode, WaterbaseID
QUALITY: CountryCode, WaterbaseID, Determinand, Year, AggregationPeriod

Result:
No primary key error has been detected.

1.3 Table relations tests

The unique Waterbase identifier (WaterbaseID) is contained in each of the data tables. It can be used to link data from one table to another. The able relations tests detect identifiers which are not present in some of the tables.

1.3.1 Number of stations without any data in the "QUALITY" table by country*

Country code	No. of stations	Percentage of total no. of stations
BA	6	11.32
BE	1	1.67
CY	1	11.11
ES	772	50.99
FI	1	0.43
FR	780	40.23
GR	9	9.57
IE	104	49.76
PT	1	1.69
SI	2	6.67
SK	3	5

Country code	No. of stations	Percentage of total no. of stations
Total	1680	26.38

*Some of the detected stations might be used for collecting information about hazardous substances in the water only. These data are not included in the Waterbase yet.

1.3.2 Number of stations without any data in the "PRESSURES" table by country

Country code	No. of stations	Percentage of total no. of stations
AL	1	3.33
BA	53	100
BE	1	1.67
BG	110	100
CH	8	100
DE	4	2.65
ES	1087	71.8
FI	226	97.84
FR	1379	71.12
GB	14	6.86
GR	94	100
HR	45	100
IE	135	64.59
IT	237	100
NL	11	47.83
RO	2	1.59
RS	77	100
SI	6	20
SK	5	8.33
Total	3495	54.88

1.3.3 “QUALITY” and “PRESSURES” table records where “WaterbaseID” is not present in the “STATIONS” table

Quality and Pressures records missing connection in the Stations table were been removed from the Waterbase. They will be included after the country will provide such information.

2. Logical rule violation tests

Logical rules were tested in the "QUALITY" data table. This table contains several measurement value fields, calculated in the aggregation process. Logical relations can be detected between them and mathematically transformed in a set of rules. Following rules have been detected and tested:

Rule	Basic validation rules
1	Mean >= Minimum
2	Mean <= Maximum
3	Median >= Minimum
4	Median <= Maximum
5	Minimum <= Maximum
6	StandardDeviation < Maximum
Rule	**Combined validation rules**
13	IF Minimum < Maximum THEN (StandardDeviation > 0)
14	IF NumberOfSamples = 1 THEN (Mean = Minimum = Maximum = Median)
15	IF NumberOfSamples = 1 THEN (StandardDeviation = 0)
16	IF NumberOfSamples = 0 THEN (AllValueType Is Null)
Rule	**Negative value validation rule**
17	All Values >= 0

The following exceptions and modifications were been applied:

IF Maximum = 0 AND StandardDeviation = 0 THEN rule 6 is not violated
IF Determinand = Temperature the values can be negative (exception of the rule 17)
IF Rule 13 is violated THEN change StandardDeviation to Null

A special QA field (QA_LRviolations) has been added to the data table. Information of the rules violated in the respective record are kept there as a coma separated list of those rules numbers (the numbers are the same as in the table above). It is recommended that the records where QA_LRviolation field is not empty (1634 records), should not be used in a further analysis. The detected data quality inconsistencies will be tried to be solved in the near future.
The records where the rules 1 and 2 are violated have not been used for the WISE. The records where the Median value was intended to be used instead of the missing Mean but where the rule 4 is violated have been also excluded.

3. Outlier detection

Detection of outliers was performed on the "QUALITY" data table. Following values were analyzed:
Measurement values: mean
Determinands: all
Aggregation periods: all
Years: all

Measurement values were compared with other values from the same time series. If the value was detected as an outlier it was analyzed whether it can be a possible error or whether it was caused by natural conditions.
Some of previously detected errors have been already corrected by countries or were approved as natural high/low values (noted in the Remarks field).

Some 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. These time series have not been included in the subset used for the WISE update.

Last part of the outlier detection process was detection of records where Mean value is not provided.

All types of the information mentioned above have been stored in a special QA field (QA_outlier) added to data table. Following QA flags have been used:

1 – record is a potential outlier (7139 records). It is recommended not to use these records in the further analysis until the issue is solved by the data suppliers.

2 – measurements are probably taken from a highly polluted locations (155 records). It is recommended not to use them for calculation of average concentrations of nutrients for broader areas like RBD or whole Country. The representativeness of the result can be negatively affected. Records have not been used for the WISE.

3 – mean value is missing (1452 records)

4. Station coordinates tests

Positions of the all reported monitoring stations have been tested using the coordinates provided. If the coordinates locate the station outside the respective country borders or if coordinates are missing this information is stored in a special QA field (QA_coordinates_err for the “Stations” table, QA_station_err for the “Quality” an “Pressures” table). Following QA flags have been used:

1 – monitoring station is located outside the respective country borders – either on the sea or in another country (17 stations, 3083 quality records, 17 pressures records)

2 – coordinates are missing (16 stations, 380 quality records, 0 pressures records)

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