Automatic and manual SoE data quality and representativity, country reviews (and consequencies for reporting and data publication)

Presenters: Lidija Globevnik and Vit Kodes

Event / date: Freshwater Eionet Workshop / 19.9.2013
Authors: Marko Kovačič, Lidija Globevnik, Miroslav Fanta, Vit Kodes

Part I

Automatic and manual SoE data quality and representativity – general overview

WISE SoE annual reporting workflow

Step 1: NRC

- data is collected from national databases and transformed to DD template
- data is uploaded on CDR

Step 2: CDR Automatic QA

- invoked when CDR envelope closed
- check the data with basic QA and display possible errors

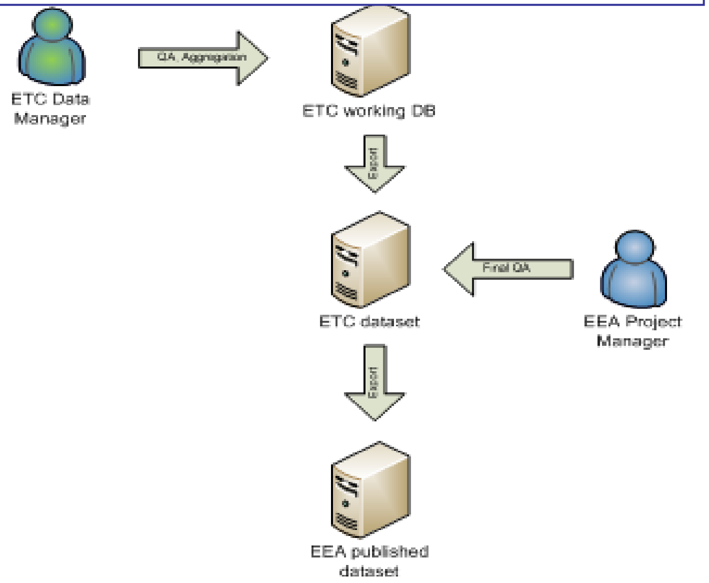
Step 3: ETC Data Manager

- collect data from CDR country envelopes and merge into ETC working database
- full QA is performed on the data
- quality assured data is exported to EEA

Step 4: EEA Project Manager

- does the final QA check
- content experts check for additional outliers
- final dataset is published on EEA website

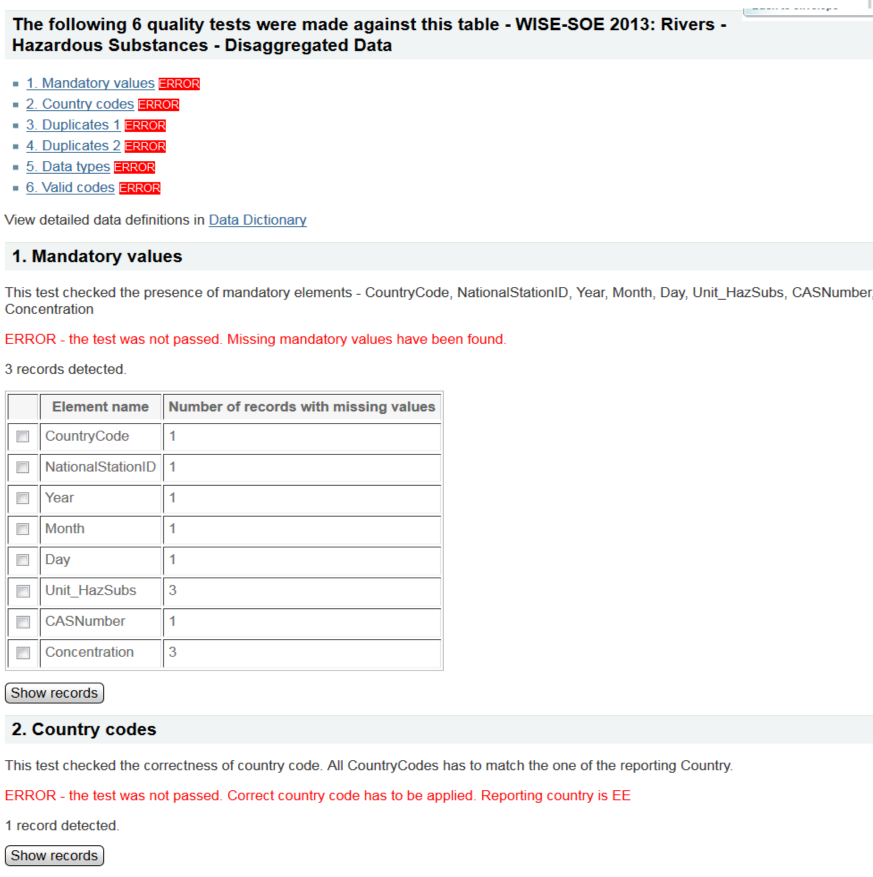
Reporter: uploads data into relevant country folder on **cdr.eionet.europa.eu;** invokes automatic QA tests; checks results, corrects data and re-uploads the corrected dataset; re-runs QA



Step 2: CDR Automatic QA

Data inserted in the latest template

- Basic QA checks
  - Mandatory values
  - Duplicates
  - Standard values according to *data dictionary*:
    - Determinand_HazSubs, CASNumber, CEN_ISO, Unit_HazSubs
  - etc.
- **New in 2013**: provides a summary of the QA result first, “Show records” shows the complete lists of detected records



The following 6 quality tests were made against this table - WISE-SOE 2013: Rivers - Hazardous Substances - Disaggregated Data

- 1. Mandatory values ERROR
- 2. Country codes ERROR
- 3. Duplicates 1 ERROR
- 4. Duplicates 2 ERROR
- 5. Data types ERROR
- 6. Valid codes ERROR

View detailed data definitions in Data Dictionary

1. Mandatory values

This test checked the presence of mandatory elements - CountryCode, NationalStationID, Year, Month, Day, Unit_HazSubs, CASNumber, Concentration

ERROR - the test was not passed. Missing mandatory values have been found.

3 records detected.

	Element name	Number of records with missing values
☐	CountryCode	1
☐	NationalStationID	1
☐	Year	1
☐	Month	1
☐	Day	1
☐	Unit_HazSubs	3
☐	CASNumber	1
☐	Concentration	3

2. Country codes

This test checked the correctness of country code. All CountryCodes has to match the one of the reporting Country.

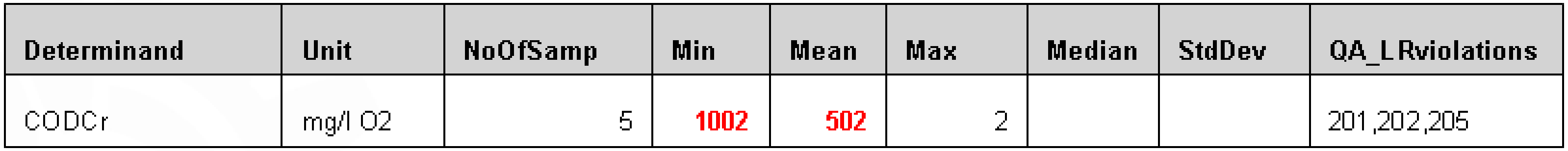
ERROR - the test was not passed. Correct country code has to be applied. Reporting country is EE

1 record detected.

Step 3: Quality assurance / Quality control (QA/QC by ETC Data managers)

List of QA rules is available in Validation rules, which are annually updated and available on NRC EIONET Freshwater interest group:

- **Logical rules** (applied on aggregated data):



Determinand	Unit	NoOfSamp	Min	Mean	Max	Median	StdDev	QA_LRviolations
CODCr	mg/l O2	5	1002	502	2			201,202,205

- **Data consistency rules**

  monitoring station ID used in the concentration table must be available in the stations table (or already stored in the working database)

  coordinates of stations must be located within a country

  consistency of reported data with the available codelists (pre-defined text values)

  ....

- **Outliers**

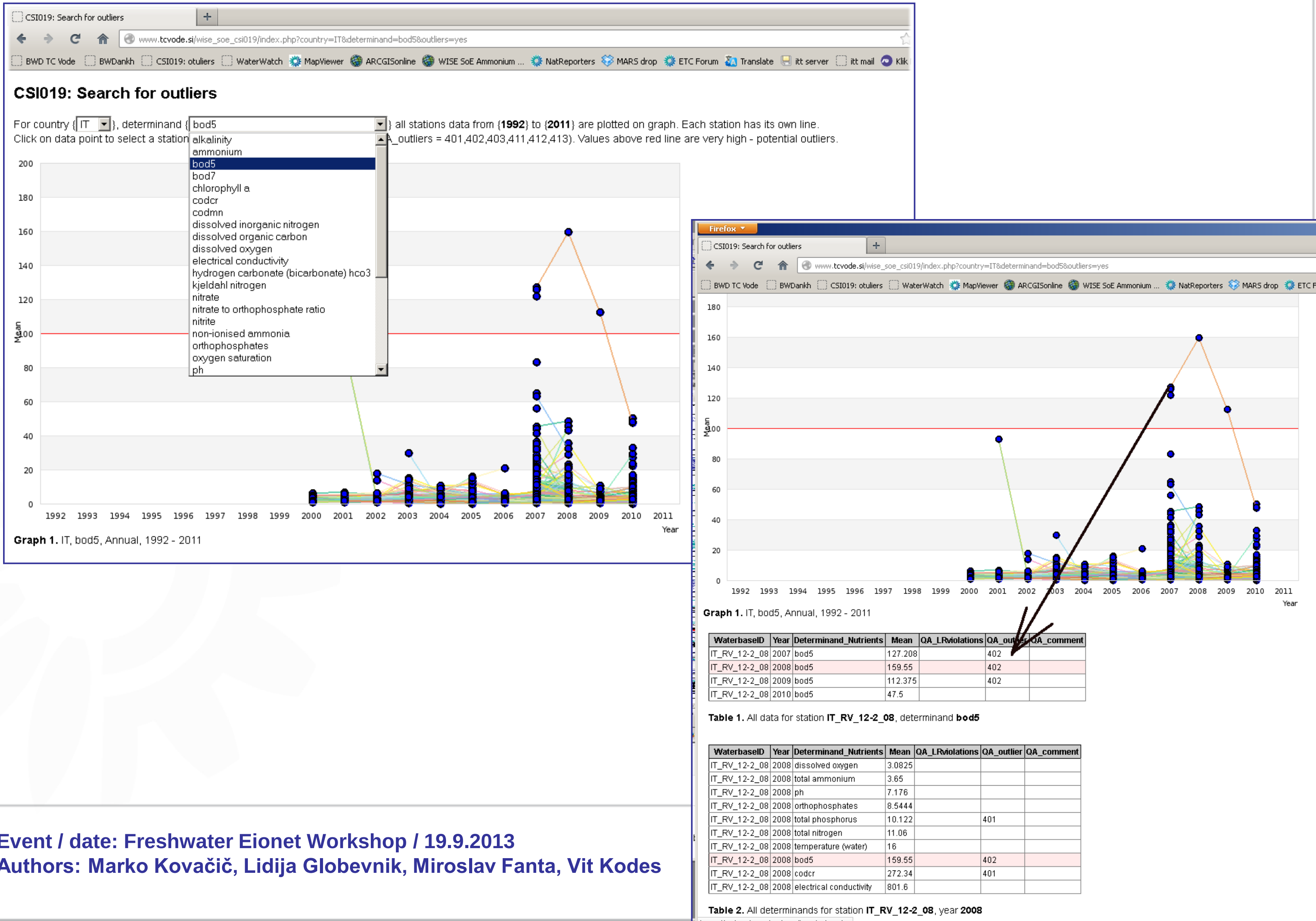
  Simple outliers: potentially extremely low / high values to detect unit errors, decimal order errors, typing errors, etc. (e.g. *pH > 14, BOD, CODcr >100 mg/l O_2 , …DO>20 mg/l O_2)*

  Complex outliers: low / high values suspicious in the context of other values provided for given substance in a given station (e.g. v*alue **± 3x standard deviation** from average in one year; **± 5.5 x standard deviation** from an average of a time series) …Q-test, Z-test, quartile test*

- **Checking of spatial data**

Checking of outliers:

Visualisation of data in common platform, viewing outliers, comparison with other determinands



CSI019: Search for outliers

For country {IT}, determinand {bod5} all stations data from {**1992**} to {**2011**} are plotted on graph. Each station has its own line.
Click on data point to select a station ... A_outliers = 401,402,403,411,412,413). Values above red line are very high - potential outliers.

Graph 1. IT, bod5, Annual, 1992 - 2011

Graph 1. IT, bod5, Annual, 1992 - 2011

WaterbaseID	Year	Determinand_Nutrients	Mean	QA_LRviolations	QA_outlier	QA_comment
IT_RV_12-2_08	2007	bod5	127.208		402	
IT_RV_12-2_08	2008	bod5	159.55		402	
IT_RV_12-2_08	2009	bod5	112.375		402	
IT_RV_12-2_08	2010	bod5	47.5			

Table 1. All data for station **IT_RV_12-2_08**, determinand **bod5**

WaterbaseID	Year	Determinand_Nutrients	Mean	QA_LRviolations	QA_outlier	QA_comment
IT_RV_12-2_08	2008	dissolved oxygen	3.0825			
IT_RV_12-2_08	2008	total ammonium	3.65			
IT_RV_12-2_08	2008	ph	7.176			
IT_RV_12-2_08	2008	orthophosphates	8.5444			
IT_RV_12-2_08	2008	total phosphorus	10.122		401	
IT_RV_12-2_08	2008	total nitrogen	11.06			
IT_RV_12-2_08	2008	temperature (water)	16			
IT_RV_12-2_08	2008	bod5	159.55		402	
IT_RV_12-2_08	2008	codcr	272.34		401	
IT_RV_12-2_08	2008	electrical conductivity	801.6			

Table 2. All determinands for station **IT_RV_12-2_08**, year **2008**

Main problems in QA/QC – Step 2 and 3

- CDR Automatic QA is not always invoked (P*repared for specific envelopes! And for the last templates*)

- Cross versions: differences in redelivered data

- Cross table checks (example: *if a station is defined in stations table, it should also be present in data tables and vice versa; if missing in station table, data cannot be used – no location data…)*

- Countries use different (national) names for attributes, especially for hazsubs names; ETCICM developed **aliases** (internal mapping tables)

  Example of different country names (aliases) for hazsubs determinand ***1,1,2,2-tetrachloroethene:***

country reported Determinand	CAS Number	correct Determinand
Perchloroethylene (tetrachloroethylene)	127-18-4	1,1,2,2-tetrachloroethene
TETRACHLOROETHENE (PER/TETRACHLOROETHYLENE)	127-18-4	1,1,2,2-tetrachloroethene
Tetrachlorethylen (Tetrachlorethen)	127-18-4	1,1,2,2-tetrachloroethene
Tetrachloroethene	127-18-4	1,1,2,2-tetrachloroethene
Tétrachloroéthylène-1,1,2,2	127-18-4	1,1,2,2-tetrachloroethene

Step 3: Aggregation

- Temporal aggregation (hazsubs) – measurements taken in the same location, same parameter at different sample dates:

Determinand_Hazsubs	Year	Month	Day	LOQ_Flag	Concentration
Copper dissolved	2011	1	5		1.1
Copper dissolved	2011	2	1		1.1
Copper dissolved	2011	3	2	<	1
Copper dissolved	2011	3	31	<	1
Copper dissolved	2011	5	9		1.8
Copper dissolved	2011	6	7		1.7
Copper dissolved	2011	7	5		1.4
Copper dissolved	2011	8	2		1.4
Copper dissolved	2011	8	29		2.9
Copper dissolved	2011	9	29		1.2
Copper dissolved	2011	11	3		1.3
Copper dissolved	2011	11	30		1.6

***Min**, **Mean** aggregation: for values < LOQ, LOQ/2 is taken*

Determinand_Hazsubs	Year	NoOfSamp	NoOfSampBelLOQ	LOQ	Min	Mean	Max	StdDev
Copper dissolved	2011	12	2	1	0.5	1.375	2.9	0.604

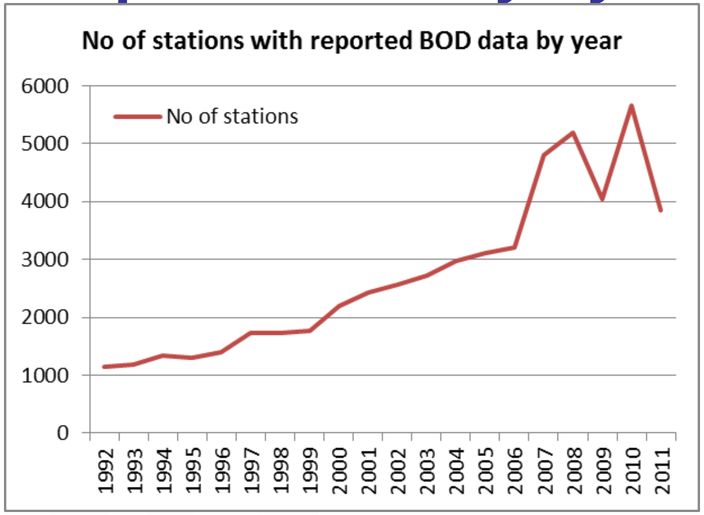
From step 3 to data use

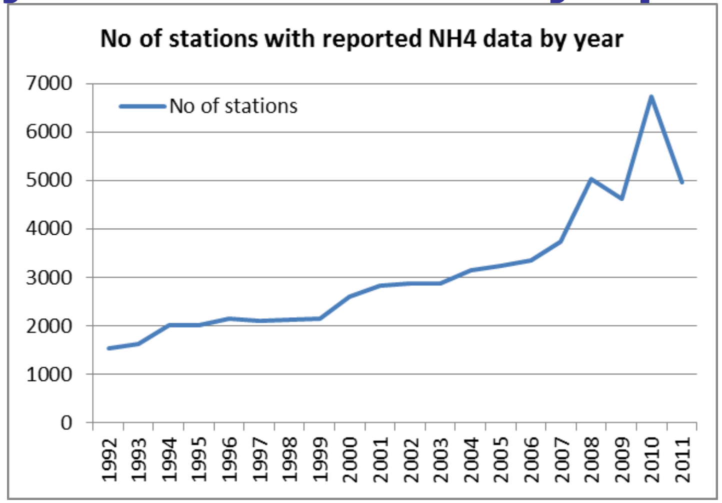
- ETC Data managers communicate QA/QC issues with countries and ask for confirmation or corrections of reported data
- Reported data are put into working database; records containing detected and uncomfirmed errors or other issues are tagged; ETC prepares aggregated data for EEA from working database (Disaggregated data are only stored in ETC working database);
- EEA controls and performs final QA/QC and publish data base (http://www.eea.europa.eu/data-and-maps/data/waterbase-rivers-9).
- QA checks are repeated through the process and by each data delivery
- Datasets are used for freshwater assessments:
  - experts decide which tagged data to include; may detect additional data quality issues/problems that are communicated further with countries
  - Representativity is important: by time, by space, by determinands

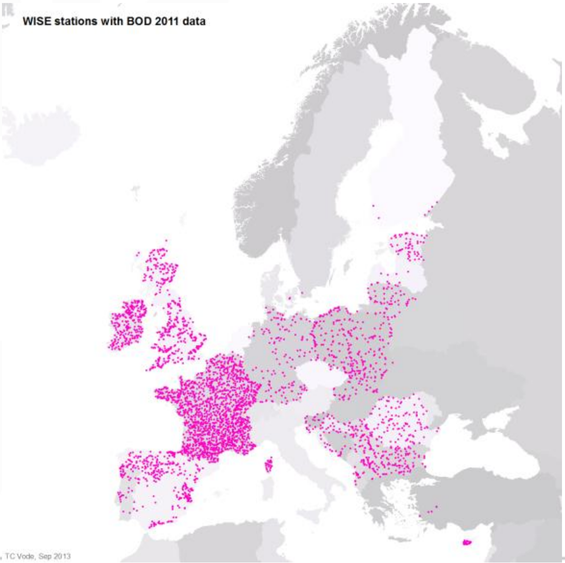
WISE SoE Rivers dataset statistics:

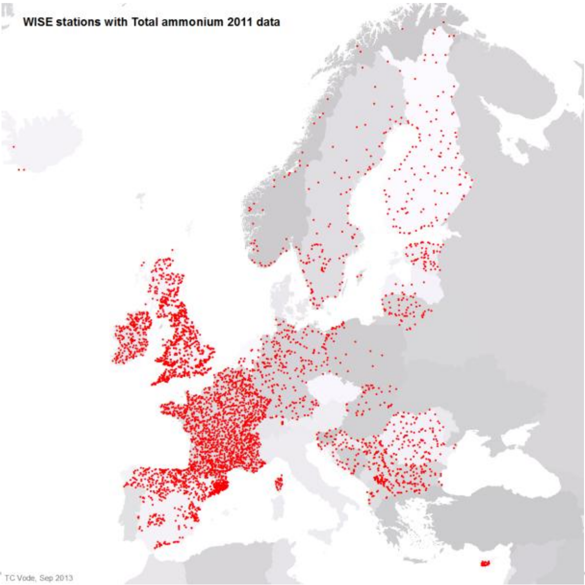
	Total records	QC/QA issues	QC/QA issues [%]
Stations	15.308	1.383	9.0%
Pressures	8.401	526	6.3%
Nutrients	1.082.336	116.194	10.7%
Hazsubs	856.144	54.787	6.4%
Hazsubs disaggregated	7.022.987	520.098	7.4%
Supportive Determinands	98.769	4.140	4.2%

SoE data on BOD and Total Ammonium - representativity by time/by determinands/ by space

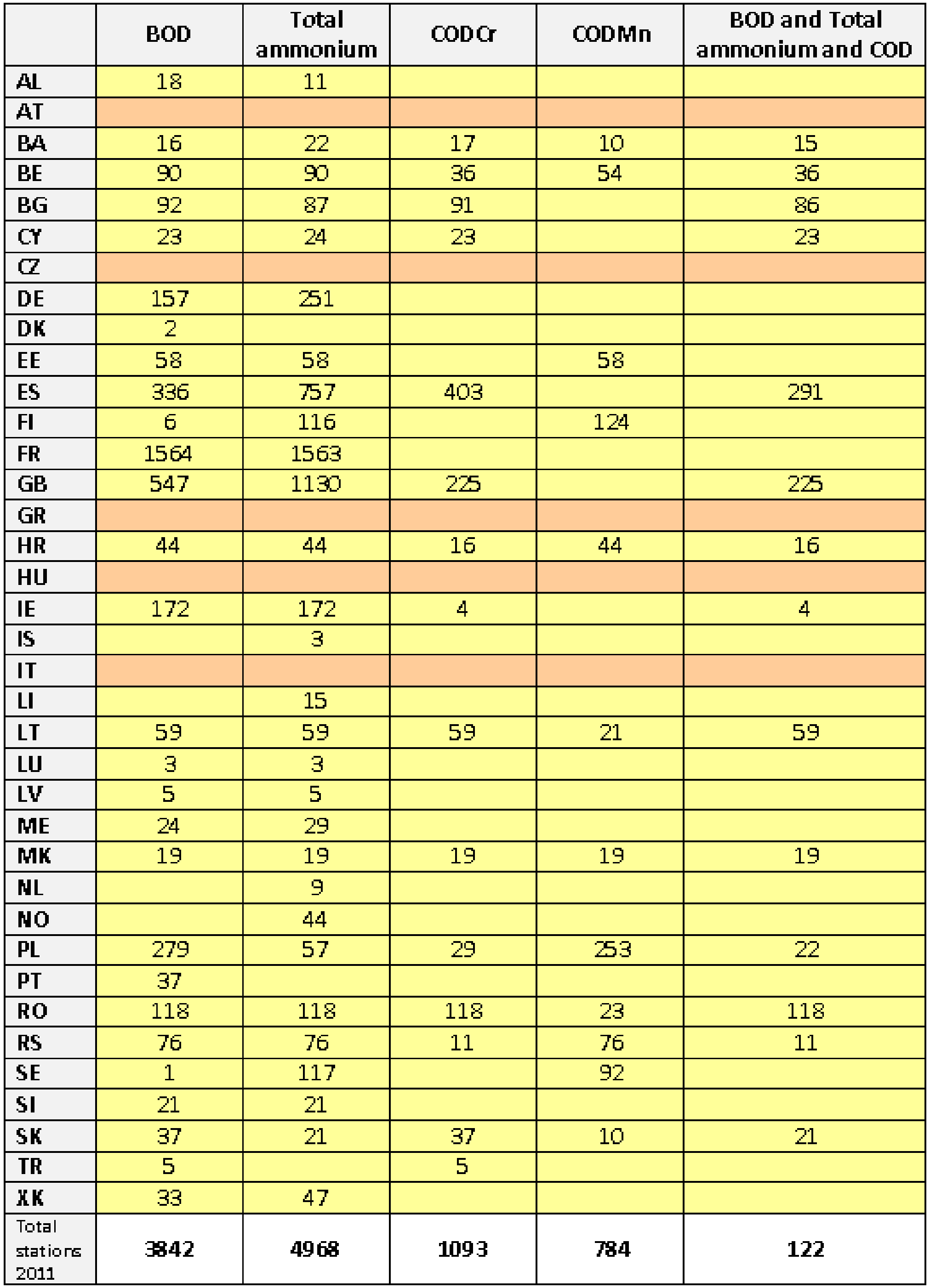








Number of stations with BOD, Total Ammonium and COD data for 2011



	BOD	Total ammonium	CODCr	CODMn	BOD and Total ammonium and COD
AL	18	11			
AT					
BA	16	22	17	10	15
BE	90	90	36	54	36
BG	92	87	91		86
CY	23	24	23		23
CZ					
DE	157	251			
DK	2				
EE	58	58		58	
ES	336	757	403		291
FI	6	116		124	
FR	1564	1563			
GB	547	1130	225		225
GR					
HR	44	44	16	44	16
HU					
IE	172	172	4		4
IS		3			
IT					
LI		15			
LT	59	59	59	21	59
LU	3	3			
LV	5	5			
ME	24	29			
MK	19	19	19	19	19
NL		9			
NO		44			
PL	279	57	29	253	22
PT	37				
RO	118	118	118	23	118
RS	76	76	11	76	11
SE	1	117		92	
SI	21	21			
SK	37	21	37	10	21
TR	5		5		
XK	33	47			
Total stations 2011	**3842**	**4968**	**1093**	**784**	**122**

Length of data series

- We are loosing timeseries. In the last CSI019 (BOD and Total Ammonium) assessment 702 (18% from 3899) stations for BOD and 921 (18% from 5049) stations for Total ammonium are included (stations with time series 1992 – 2011).

- Some countries have not reported lately, e.g.
  - Hungary has not reported since 2007
  - Czech Republic has not reported since 2008
  - Austria has not reported since 2011

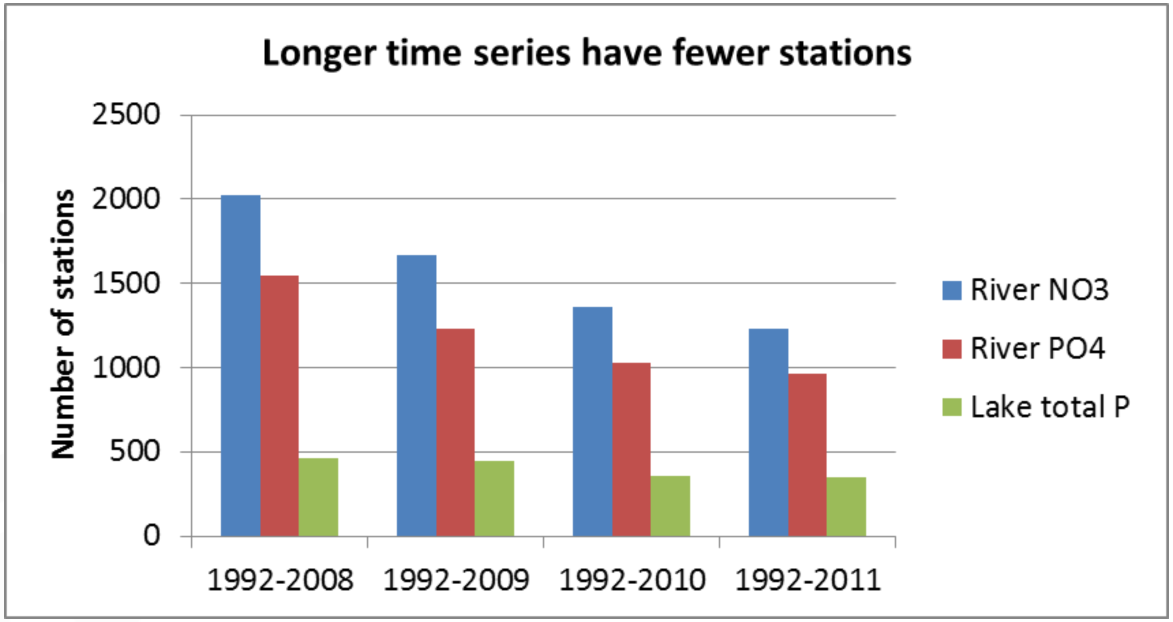
- Some countries have stoped reporting some determinands or under different changed name (Ammonium -Total ammonium).

- **Loss of stations as time series get longer reduces representativity!**

SoE nutrients:
Fewer stations as time series get longer

By Anne Lyche Solheim (ETCICM - NIVA):

- Time series analysis requires consistency
  - Only stations with complete series after inter/extrapolation can be used
- Loss of stations as time series get longer reduces representativity
  - Monitoring stopped (or changed?) or Reporting stopped?
  - Reporting errors (changes in station coding = new station)?



Part II

Part II

Hazardous substances:

Hazardous Substances data report and country reviews – consequences for reporting and data publication

Hazardous Substances data report and country reviews – consequences for reporting and data publication

- The ETC/ICM Technical Report 1/2013 is a **complementary** report to the European Environment Agency (EEA) Report No. 8/2011

- The **first attempt** to compile the SoE data on selected hazardous substances

- Provides information on the **status of the ETC/ICM hazardous substances database**, **SoE data availability and the occurrence** of hazardous substances throughout Europe including spatial and temporal changes

Background

- a systematic **summary** presentation of the data giving a quick overview of the **state and availability** of hazardous substances SoE data, **occurrence**, **concentrations levels** and **trends** over time a **compact display** of the thousands of data records for each substance
- **not an assessment** of the situation between the reporting countries
- a **periodical** Technical Report updated every second year
- next issue of this report will cover the period **2002–2011**, including **lake** data
- on-going thorough **QA/QC** procedures

Hazardous data quality issues and QA/QC

Issues:

- Outliers
- Units
- LOQ, LOD
- Identification of substances (Names, CAS) – rivers,lakes
- Supportive determinands (hardness) – rivers,lakes
- Disaggregated x aggregated data
- SUMs (DDTs, HCHs, PAHs) – rivers,lakes
- Too much data excluded from an assessment due to QA issues

QA/QC:

- Databases clean up and unification
- QA/QC procedures enhancement
- Common QA/QC procedures across water categories

Country comments

13 countries participated in commenting in 2012:

- AT, CH, CY, DE, FR, GR, LV, NL, PL, SE - rivers
- AT, CH, CY, SE, - groundwater
- HR, IS, PL, UK – TCM

12 countries participated in commenting in 2013:

- DE, FI, FR, RO, SE, SI - rivers
- AT, CH, CY, DE, DK, FR, GR, IE, SI - groundwater
- DE, SE - TCM

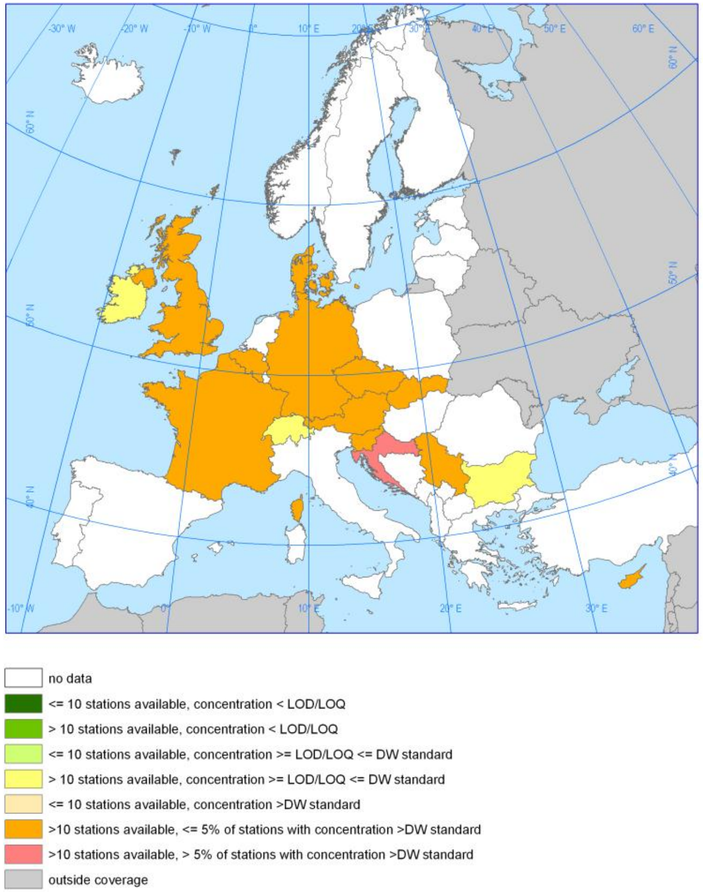
Consequences for reporting

- Preference of **disaggregated** data - rivers, lakes
- **Total x dissolved** concentrations identification (metals) - rivers, lakes, groundwater
- Provision of supportive determinand for cadmium (**hardness**) - rivers, lakes
- **LOQ** specification in aggregated data - rivers, lakes
- Provision of **threshold values** - groundwater

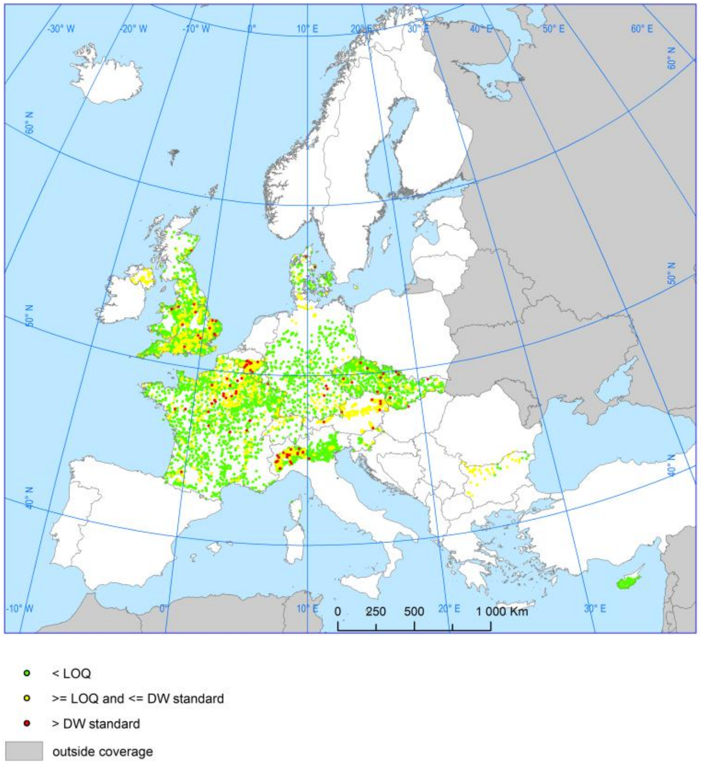
Consequences for data publication

- Assessment based on **disaggregated data** (preferably) and aggregated data where disaggregated data not available instead of aggregated data publication - rivers, lakes
- Separate assessment for **total** and **dissolved** concentrations (metals) instead of a combined assessment for both types of concentrations - rivers, lakes, groundwater
- Handling of missing supportive determinand for cadmium (**hardness**) - rivers, lakes?
- Handling of **missing LOQ** in aggregated data - rivers, lakes?
- Assessment based on either **threshold values** or **drinking water standards** - groundwater
- **Station data** presentation in the maps instead of country aggregated maps - groundwater

An example: atrazine in groundwater



🟦 no data
🟩 <= 10 stations available, concentration < LOD/LOQ
🟩 > 10 stations available, concentration < LOD/LOQ
🟩 <= 10 stations available, concentration >= LOD/LOQ <= DW standard
🟨 > 10 stations available, concentration >= LOD/LOQ <= DW standard
🟨 <= 10 stations available, concentration >DW standard
🟧 >10 stations available, <= 5% of stations with concentration >DW standard
🟥 >10 stations available, > 5% of stations with concentration >DW standard
⬜ outside coverage



Country aggregation

Station data

Part III

Part III

Questions to NRCs and discussion

1) How to efficiently follow changes in station codes?

Since guidance is not always followed (codes in all tables), due mistakes and slight relocations of stations, codes change (*example: GB_RV_GBF10028 and GB_RV_F10028*): Suggestion: Countries report stations in all tables and provide stations mapping tables

2) How to increase accurancy of hazardous substances reporting?

Sometimes countries use national names (example: *Aldriini, Aldrine, etc. for determinand Aldrin) and put wrong units (*mg/l instead of µg/l or vice versa*;* Suggestions: Always use correct CAS number, follow DD templates and internally check units.

3) Are countries willing to:

- report disaggregated hazardous substance rivers and lakes;
- report national threshold values for groundwater per station;
- update HS dataset (total and dissolved concentrations of metals;
- update/report nutrient data (2011, 2012…)?

4) How to motivate countries to answer critical validation questions and how to motivate countries to communicate with data managers? We know that

- CDR feedback features are hard to work with
- E-mailing is hard for archiving and control
- NRCs are preoccupied to promptly answer and regularly communicate with ETCICM

Possible way forward:

- **“Communication tracking” system could ease communication between countries and ETC ICM**
- Use **common platform for visualisations** of tabular and spatial datasets to check data (online maps, tables)

example: ETCICM SoE nutrient data platform

Thank you!