# 2013 Freshwater Eionet Workshop - 19/20 September 2013, Copenhagen

## Session 2: Data quality and aggregation

## Document 2a: ETC SoE data automatic and manual quality assurance, Marko Kovacic, ETC/ICM

### Background

WISE SoE annual data request run usually from 1.8 to 31.10 each year. In case of rivers, water quality and biological quality elements are requested. The main reported tables are:

* stations
* pressures
* nutrients
* hazardous substances
* supportive determinands
* biology
* classification system for ecological status

Countries **report data on CDR** where automatic quality checking is invoked if the data is supplied in latest format according to Data Dictionary. The task for the rivers data manager is to collect this data from the CDR, put it together into the working database and make a thorough quality checking on this data. The datasets from the current SoE data request are published on the EEA website and interactive maps from the list by category are also available.

**Nutrients and Hazardous substances** data are the far largest part of SoE Rivers database. There are around 1M records in the Nutrients table and 800K records in the aggregated HazSubs table. In the working database (not published by EEA) there are more than 6M disaggregated HazSubs records.

The main **quality checking procedures** for nutrients and hazardous substances data and some examples of **common mistakes from countries are as follows**:

* Mandatory fields contain data. All field that are marked as mandatory (\* sign) in the Data Dictionary should contain data. If not, the record is not useful for further analysis. Such records are excluded from the final dataset and countries are asked for clarification.
* Characters in integer or float fields. Countries sometimes put sign “<” and units into float fields (i.e. < 10 mg/l) which is wrong. Sign < should be put into LOD\_LOQ\_flag field and units into Unit\_Hazsubs or Unit\_Nutrients field.
* Standardize values according to data dictionary. Fields like Determinand\_HazSubs, CASNumber, CEN\_ISO, Unit\_HazSubs should contain standardized values which are defined in Data Dictionary for these fields. Any other values are removed from this field or record is marked as useless.
* Logical tests (for aggregated data). Aggregated values should all conform to logical tests like Mean > Minimum, Mean < Maximum, etc.
* Orphaned stations: stations not defined in stations table. All station codes (NationalStationID) used in data tables (Nutrients, Hazsubs, …) have to be defined the in Stations table otherwise the data are useless.
* Alias table for Determinand\_HazSubs / CASNumber: automatically correct national determinand names. Many countries (cca. 40% of all reported Hazsubs data in SoE 2012) use their national names for Hazsubs determinands. These values have to be standardized. One approach used to standardize these values is an “alias table”, see the example below:

|  |  |  |
| --- | --- | --- |
| **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 |

**Document URL**: <http://taskman.eionet.europa.eu/ETCW/wiki/Data_Handling>

Note: There is no pdf or word document, SoE QA documentation is published on this URL and maintained by Marek Staron, EEA. This documentation is the basis for all SoE data managers and their QA/QC work.

### Issues to be discussed and questions to NRCs

1. **Station recodifications**. Many countries do not follow the guidelines or make mistakes when it comes to changing the station codes (for example: GB\_RV\_GBF10028 and GB\_RV\_F10028). Countries should have some internal checking for station codes – if a station is defined in nutrients, hazsubs tables it must also be defined in station table and vice versa. If the station codes have been changed, countries have to provide a mapping table, so that all old data can be recodified and time series remain.
2. **Aliases in hazardous substances**. Unlike in nutrients reporting, countries usually do not follow the DD templates in reporting the hazsubs data. They use national names instead of official DD names: e.g. Aldriini, Aldrine for determinant Aldrin. The ETC/ICM has created a mapping table with all these combinations for internal QA work, but it would be more proper if countries followed the DD templates.
3. **Units in hazardous substances**. Unlike in the nutrients reporting, countries do not always use units specified in DD, furthermore for some data one unit is used (i.e. mg/l) and for other part of data different unit is used (i.e. µg/l). It is no problem to convert units by data managers, but in many cases than these values become outliers - this usually indicates an error in selected unit (i.e. µg/l instead of mg/l).
4. **Hazardous substances, filtered and unfiltered samples**.According to Directive 2008/105/EC EQS for metals refers to the dissolved fraction, i.e. the dissolved phase of water sample obtained by filtration through a 0.45 µm filter. In the SoE database it is not clear if the countries have submitted data on metals on filtered or unfiltered samples.
5. **Improve communication with countries**. Ticketing system (Trac, Redmine) to be used also for country communication – CDR feedback feature is hard to work with, no automatic email. Each country has its own ticket, simple to track progress; the EEA project manager can always track communication. The other option is to improve CDR feedback feature and that countries start to use it.
6. **Validation questions**. Countries are not very eager to answer validation questions (only 5 answered this year, 4 last year).

### N.B. see also document 2c dealing with hazardous substances more specificly.