Stockholm Convention
Regional Centre for Capacity Building
and the Transfer of Technology

Stockholm Convention
on persistent organic
pollutants (POPs)

FACTSHEET 4 ON GLOBAL MONITORING PLAN DATA WAREHOUSE

DATA MANAGEMENT ISSUES FOR DATA PROVIDERS

GMP DWH DATA STRUCTURE

Global Monitoring Plan Data Warehouse (GMP DWH) has been designed for importing both **primary and aggregated** POPs data from four environmental matrices – **ambient air, human blood, human milk and water.**

Data structure is **fully standardized** into three key items: **site, sampling attributes and measurement** (see Figure 1). POPs data are imported to the GMP DWH Data Repository through individual import branches organized per environmental matrix (see Factsheet 3).

Imported data fields have predefined form and are either **text, number or item selected from a particular code list.** Code lists are available in **analytical data reporting spreadsheets** (MS Excel format) on the website http://www.pops-gmp.org/dwh or are embedded in the GMP DWH Data Repository.

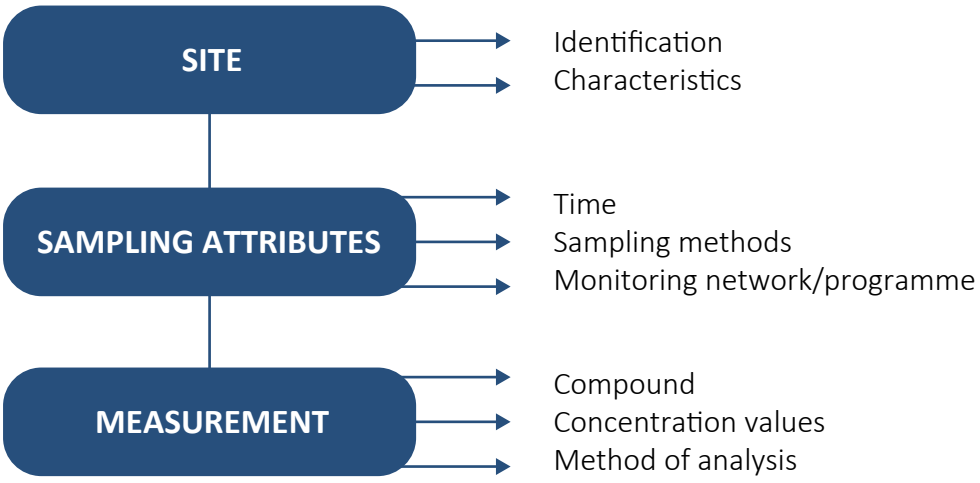

Figure 1- Key items of the data structure in GMP DWH

All data shown in GMP DWH Data Visualization are annually aggregated data. The data management strives for GMP DWH to process any primary data that are made available in a transparent and consistent manner according to agreed assessment methodologies.

Respecting the instructions summarized in the Guidance on the Global Monitoring Plan for Persistent Organic Pollutants and its Technical Note to Chapter 3 Statistical Considerations, the following three steps are performed to get aggregated values (see next page):

1. PARAMETERS OF AGGREGATION (GROUPING ELEMENTS)

The following set of parameters must be identical:

1. Matrix (air, water, human tissues);
2. Sampling site;
3. Year (for samples that were taken over the end/start of the calendar year, the day in the middle the sampling period was used for assigning the relevant year to the sample);
4. Type of sampling (in case of the ambient air matrix: passive, active);
5. Blood source and fraction (in case of blood);
6. Parameter (detailed specification of a substance).

2. TREATMENT OF VALUES BELOW LIMIT OF QUANTIFICATION

Values in a dataset that are below limit of Quantification (LOQ) are substituted by a constant. The GMP DWH uses 1/2 LOQ as the "left censoring limit", due to simplicity of this substitution method and its widespread use.

3. COMPUTATION OF AGGREGATED VALUES AND MEASURES OF VARIABILITY

The following aggregated values and measures of variability are then computed:

1. **Arithmetic mean** – mean of all concentration values. If the original value is lower than limit of quantification, a substitution value computed as ½ of the limit is used instead.
2. **Median** – non-parametric analogue of the mean computed in the same way as a 50th percentile.
3. **Geometric mean** – a parametric statistic used for estimation of a central tendency of log-normally distributed data, which is suitable especially for air pollution measurements.
4. **Standard deviation** – a parametric measure of variation. If only one record is used for computing the aggregation, standard deviation is not determined.
5. **5th and 95th percentiles** are computed as non-parametric measures of variation.
6. **Minimum and maximum** are computed as 0th and 100th percentile.
7. **Start/end of the sampling** in a particular year are determined as a start date of an initial sampling and an end date of a final sampling within the year. If the sampling period exceeded start/end of the year, the value of 1 January/31 December is used instead.
8. **Sampling frequency** is determined as a characteristic period between the two successive samplings. The term "characteristic" means that at least 50% of the time between two successive samplings was in this period. In case of months, some margin of tolerance is added due to uneven length of calendar months. For non-periodic sampling and sampling with only one sample in a year, the value of "12 months" is used as the characteristic period.
9. **Largest gap** is determined as a longest period between the subsequent samplings and start/end of the year in days. i.e. if a weekly sampling starts at the beginning of the year and ends in February, the value will be higher than 300.
10. **Minimum/maximum depth** are computed as a minimum/maximum value of sampling depth in case of surface water (otherwise it is not defined).
11. **Number of records** are determined as quantity of primary values used for computing the aggregation.
12. **Number of records below LoQ** are determined as the number of primary data records whose concentration value is under the censoring limit.

Please note that aggregation may also affect the three following items:

- **Type of passive sampling** is set to "not classified" if aggregated items differ in this option within one calendar year.
- **Recalculation** is set to "not classified" if aggregated items differ in this option within one calendar year.
- **Method** is set to "not classified" if aggregated items differ in this option within one calendar year.