EMODnet Thematic Lot n°5 – Chemistry

EASME/EMFF/2020/1.3.1.11/Lot5/SI2.846161

Start date of the project: 03/10/2023 (24 months)

Harmonization and data validation on litter aggregated datasets
10.07.2025

Disclaimer

The European Marine Observation and Data Network (EMODnet) is financed by the European Union under Regulation (EU) 2021/1139 of the European Parliament and of the Council of 7 July 2021 establishing the European Maritime, Fisheries and Aquaculture Fund.

Document info

Title (and reference)	Harmonization and data validation on litter aggregated datasets (https://doi.org/10.13120/je91-sp43)
WP title (and reference number)	
Task (and reference number)	D. 3.3 Validated collections for beach and seafloor litter; D. 3.4 Validated collections for microlitter
Authors [affiliation]	M.E. Molina Jack, Matteo Vinci, Alessandro Altenburger, Nikola Holodkov, Elena Partescano, Alessandra Giorgetti (OGS)
Dissemination level	Public
Submission date	10.10.2025

Contents

Harmonization and format validation on litter aggregated datasets

1 History

Version	Authors	Date	Comments
1	M.E. Molina Jack,Matteo Vinci, Alessandro Altenburger, Nikola Holodkov, Elena Partescano, Alessandra Giorgetti (OGS)	10/10/2025	Creation

2 Introduction

Data related to marine litter available in EMODnet Chemistry is at present focused on beach macrolitter, seafloor litter from trawlings, microlitter in water bodies and microlitter in sediment. Each type of data is managed through specific format and standard vocabularies that have been developed according to the best practices in the field of marine litter.

Beach and seafloor litter data from trawlings are managed in two European central databases that are populated with data formatted according to the indications given in the EMODnet Chemistry Guidelines and forms for gathering marine litter data: beach and seafloor trawlings (https://doi.org/10.6092/15c0d34c-a01a-4091-91ac-7c4f561ab508). Instead, microlitter data is managed following the generic distributed data management approach established in EMODnet Chemistry and the format (ODV for microlitter) and vocabularies are described in EMODnet Chemistry Guidelines and formats for gathering and management of micro-litter data sets on a European scale (floating and sediment micro-litter) (https://doi.org/10.6092/d3e239ec-f790-4ee4-9bb4-c32ef39b426d).

Once a year, EMODnet Chemistry provides validated aggregated datasets concerning the above-mentioned data types. Data is centrally harvested at the CDI Data and Access Service to collect the most up to date version of the datasets available in the system. This data products generation process represents a further post-ingestion quality control step, relevant to improve the reliability of the managed data.

This document describes the steps of data aggregation and data and metadata validation needed to achieve the standardised, harmonised and validated datasets concerning **marine litter.**

3 Beach Litter data

Data is aggregated into regional datasets using an internal tool (Litter Merger) based on python (https://github.com/nodc-it/LitterMerger/), building on datasets that contain data for one sampling event (survey, CDI datasets granularity). The aggregated datasets keep the source data format (https://doi.org/10.6092/15c0d34c-a01a-4091-91ac-7c4f561ab508), i.e. EMODnet Beach litter data format for beach litter (https://vocab.nerc.ac.uk/collection/L24/current/EBL/) .

During the merge the software verifies the consistency of the date format in the file, the correctness of the field's labels and the coherence of the survey codes within each file. When an anomaly is found it is described in the log file.

Once all survey datasets are combined into a single regional dataset, a validation is performed using the existing validation tools based on a RESTful webservice (https://nodc.ogs.it/marinelitter/beachvalidator) as described in Partescano et al, 2021. Syntactic (format) and semantic (content) checks serve to ensure uniformity in the data format and to guarantee a minimum quality standard.

Syntactic controls include the verification of the number and position of columns and column names or reading the sheet lines and creation of Java objects for the subsequent validation phase.

Semantic tests verify the coherence between coordinates of surveys and survey lengths (highlighted with warnings), the presence of at least two coordinates or one coordinate and the length to be able to evaluate the sampling effort or the coherence of terms used to describe the beach survey or litter items according to the vocabularies (e.g. https://nodc.ogs.it/marinelitter/vocab).

The complete list of verifications is available in an additional report (Cociancich et al., 2021a).

This validation process is performed offline running the code on a local system. This is needed to face the long processing time required in the case of large aggregated datasets that may conflict with the time out of the online validation tool.

If the validation provides warnings or errors, data is further checked using additional tools such as QGIS, internal python scripts and further postgreSQL queries. Following this process the data originators are contacted to evaluate the anomalies and decide about possible corrections.

Additional checks are made based on the type of warning. If the warnings correspond to one of six specific categories (Beach far from coastline, Survey far from coastline, Start coordinate discrepancies, Survey length exceeding beach length, Survey length far from calculated distance, Survey too short/long), a Python script automatically generates interactive HTML maps using the Folium library for visual inspection.

The mapping process uses a European coastline shapefile to calculate minimum distances to the mainland, applying a >100m threshold for coastline distance issues. For each problematic record, the script displays beach/survey points, nearest coastline locations, and distance measurements with connecting lines. The tool also compares survey coordinates with beach locations, validates reported versus calculated survey lengths, and visualizes start/end points with comprehensive distance indicators.

This automated process creates a structured validation workflow with organized folders containing both problematic data files and their corresponding interactive HTML maps, enabling systematic visual inspection of each identified data quality issue and supporting evidence-based discussions with data originators about potential corrections.

4 Seafloor Litter data from trawlings

Data is aggregated into regional datasets using the above-mentioned LitterMerger tool, starting from files that contain data for one sampling event (haul, CDI datasets granularity). As for beach litter, the aggregated files maintain the original data format, that is EMODnet Seafloor litter data format for seafloor litter from trawlings (https://vocab.nerc.ac.uk/collection/L24/current/ESFL/).

Data aggregated in the regional datasets is validated using the RESTful webservice (https://nodc.ogs.it/marinelitter/bottomvalidator) described previously. The validation output highlights if some information is missing or incorrect. During the validation process the presence of duplicates is checked through the verification of the coordinates, date of the survey and types and quantities of items found. The coherence of the information related to the survey lengths and the presence of coordinates on land are also verified. More details are available in a separate report (Cociancich et al., 2021b).

As before, if the validation results in warnings or errors it is fundamental to further check the discrepancies by means of SQL queries directly in the database or through QGIS in the case of spatial inconsistencies. The collaboration of data providers is essential to update/correct the data, both in the validated aggregated dataset and in the database. Once corrections or updates are done, CDIs are resubmitted to maintain the alignment between the collections, the data in the PostgreSQL database and the Data Access System.

5 Microlitter data

Ocean Data View (ODV) software is used to combine the different datasets (data + metadata) into a single one in extended ODV format as follows:

1. Import ODVs and metadata files. Beforehand, check the range of duplicate detection: if it is too high, stations can be considered as duplicates by ODV (specially if metadata are the same: cruise, station, type, local cdi, edmo); 0.2 hour and 0.005 degrees might be reasonable.
2. Check if there are empty variables. If so, check the reason: is it any problem during the import? Or originators report empty variables?
3. Perform a harmonised export of the collection. The harmonisation process using ODV performs unit conversion into a common P06, merges the same P01 expressed by different local names allowing an easier retrieval of variable names from P01 standard codes. The common units are:
    - Depth variables expressed in metres
    - Length variables expressed in kilometres
    - Area of sample collector (aperture) {mouth area} in square centimetres
    - Width of sample collector (aperture) {mouth width} in centimetres
    - Mesh Size in micrometres
    - Sample volume in cubic metres
    - Wind direction in Degrees True
    - Wind speed in metres per second
4. Check which variables have not been harmonised:
    - Because units are different
    - Because the P01s or P06 conversions are not present in the “sdn_harmonization_settings.xml” file
5. If needed, create a derived variable to aggregate variables/units when feasible. Copy the P01, P06 to add it in the comment of the properties of the new created variable.

6. Check that the number of decimals is adequate for the float variables (e.g. weight and volume might need a larger amount of decimals, around 8 as the numbers might be really small).
7. Export the collection with the spreadsheet format non-compact, long between -180-180.

To compare the updated data between the previous collection and the present one:

1. Together with the harvest, a complementary file with the differences in the collections is provided and it highlights (old-updated-new) data; therefore it is possible to filter the cdi that have been updated. Afterwards, through the text join function in excel it is possible to create an array with the CDI numbers like:
   <CDI1> || <CDI2> || <CDI3> || … (don't forget the spaces)
2. With this array in ODV, it is possible to filter through the metadata field "CDI id" inserting the array with the list of CDI that have been updated. This must be done both in the previous collection and in the present one, specially if the differences seem consistent.
3. Once both collections are filtered they can be exported
4. ODV allows to compare the 2 exported collections through:

- Tools → Other collection → Compare data availability

This will give the differences on the cruise, station and sample level to check whether they are reasonable.

The spreadsheet export is then used to harmonise the H vocabulary terms for the reporting of the microlitter properties according to the identifier and preferred label as defined in the Nerc Vocabulary server (e.g. SDN:H01::H0100001 (microplastic items)). This is done by means of scripts in Python outside ODV.

Data validation

Data validation is performed with the spreadsheet export. The following controls are performed:

1. Check coordinates likelihood
2. Check that primary variable is always present
3. Check consistency between bottom_depth and water depth
4. Check units consistency and non-imported or non-harmonised variables
5. If there are empty variables inside the collection check the reason. Is it due to a problem in the import or is it because the originator included empty variables in the ODV files? Address the issues with the originator
6. Check that litter type is always present. Even if count is 0, a generic type must be present
7. Check that size range is always present. Even if count is 0, a generic size range must be present
8. Check that count is always present
9. Check that count and weight are consistent (e.g. weight 0 and count 0)
10. Check that H vocabularies are consistent with the reported property (e.g. litter type: H01; size range: H03…)
11. Check constant values in different columns
12. Check constant values in litter properties reporting (H values or count)
13. Check that sampling effort is always present.
14. Check sampling effort consistency: if count !=0, sampling effort cannot be 0

15. Check QF=0. Data providers must QC their data

If there are any issues during data validation, data providers are contacted and with their help corrections of anomalies are evaluated and performed; afterwards, the spreadsheet export with the correct H labels is re-imported into ODV.

Export

1. Open the collection. Order the variables and rename them as follows:

MinDepBelowSeaSurf [m]
MaxDepBelowSeaSurf [m]
End_Lat
End_Lon
EventEndDateTime* [years since 0000-01-01]
mouth_width [cm]
Mouth_area [cm2]
Mesh_size [micrometres]
Sampling_protocol
SampTrkLen [km]
VolWBodySamp [m3]
sampID
SubsampleID
Microlitter_count_water
Microlitter_count_water_VOR
Microlitter_weight_water [g]
Microlitter_type_water
Microlitter_size_class_water
Microlitter_polymer_type_water
Microlitter_colour_class_water
Microlitter_transparency_water
Microlitter_shape_class_water
WMOSeaState
WindSpd [m/s]
WindDirFrom [deg T]

***** To export EndDateTime as ISO string it needs to be named exactly as EventEndDateTime

2. Export as needed.

6 References

- A. Cociancich, A. Brosich, A. Altenburger, E. Geletti, M.D.M. Chaves Montero, M. Vinci, M.E. Molina Jack, E. Partescano. EMODnet Chemistry Beach Litter Data Validation Process (V.1). Zenodo (2021), 10.5281/zenodo.4557648

- A. Cociancich, A. Brosich, A. Altenburger, E. Geletti, M.D.M. Chaves Montero, M. Vinci, M.E. Molina Jack, E. Partescano. EMODnet Chemistry SeaFloor Litter Data Validation Process (V.1). Zenodo (2021), 10.5281/zenodo.4588994

- E. Partescano, M.E. Molina Jack, M. Vinci, A. Cociancich, A. Altenburger, A. Giorgetti, F. Galgani. Data quality and FAIR principles applied to marine litter data in Europe. Marine Pollution Bulletin (2021). https://doi.org/10.1016/j.marpolbul.2021.112965.