BIODATABASE: Loading documentation

Manu Chassot

01 February 2020

Metadata

The CSV file *ddd.csv* exported from the spreadsheet 'DDD' of the XLSX file DDD_Database.xlsx is the Data Dictionary Design of the database. The data set includes the following fields: entity, variable, data_type, unit, basic_checks, comment, tracer, views_level. It contains the description ('comment') of 517 variables linked to 21 distinct entities. The column 'basic_checks' refers to the reference tables of the database (section Reference tables).

The column 'tracer' takes the value 1 for the variables corresponding to 351 quantitative ecological tracers (e.g. isotopic ratio δ^{15}N). It takes the value 2 for the 30 morphomometric measurements of the fish (e.g. fork length) and fish organs (e.g. liver weight). The column 'tracer' is used during the data loading process to generate the tables metadata.analysis_tracers_details and metadata.fish_measures_details, respectively. These two tables are exported as .CSV files and called by the R scripts melt_analysis_tracers.R and melt_fish_measurements.R which melt the pivot data prior to insertion in the database. The melted data sets melting_tracers.txt and melting_fish_measurements.txt are then inserted into the tables analysis_tables.analysis_measures and public.fish_measures, respectively (section Analysis).

The column views_level can be used to filter the 333 main ecological tracers of interest (value = 1) to extract from the database vs. the 18 that are not of direct interest to the user but were required for some computation of tracers, e.g. extraction_vial_empty_mass.

The insertion of the DDD contents into the table metadata.ddd and creation of the tables metadata.analysis_tracers_details and metadata.fish_measures_details are described in **load_metadata.Rmd**.

Reference tables

The XLSX file DDD_Database.xlsx includes 39 tables that provide information on the different metadata and data sets: AFAD, AMINO_ACIDS, ANALYSIS, ANALYSIS_LAB, ANALYSIS_MODE, ANALYSIS_REPLICATE, ANALYSIS_SAMP_DESCRIPTION, ATRESIA, CRM, DERIVATIZATION_MODE, DRYING_MODE, EXTRACTION_MODE, FATTY_ACIDS, FISHING_MODE, GEAR, GRINDING_MODE, LANDING, MACRO_MATURITY, MICRO_MATURITY, MINERALS, OCEAN, OPERATOR, ORGANIC_CONTAMINANTS, OTOLITHS, PACKAGING, POF, PREY_GROUPS, PROCESSING_REPLICATE, PROJECT, SAMPLE_POSITION, SAMPLING_PLATFORM, SEX, SPECIES, STOMACH_FULLNESS, STORAGE_MODE, TISSUE, VESSEL, VESSEL_STORAGE. The insertion of the reference tables into the schema references_tables of the database, including the creation of an additional table ANALYSIS_GROUPS, is described in **load_reference_tables.Rmd**.

Fish

The 'fish' data set includes information on the fish collected (e.g. species, sex) and is composed of two distinct data sets: (1) *csv_fish_iot_nb.csv* and (2) *Data_Sampling_fish.csv*. The first data set includes 37,812 fish collected at the cannery IOT Ltd. and the SFA lab between 1987-08-26 and 2015-11-12 through historical IRD-SFA projects. The second data set is a CSV export from the working XLSX file Data_Sampling.xlsx that includes 23,939 fish collected through different projects conducted in the Seychelles and other oceans in collaboration with IRD and SFA partners between 1974-01-01 and 2019-10-04.

The loading of the data in the database is performed with Talend and consists of three steps: (i) the merging of the two data sets into a CSV file *fish_emotion3.csv*, (ii) the insertion of some fields into the table public.fish (Job *Load_fish*), and (iii) the insertion of the variables described in the table metadata.fish_measures_details into the table public.fish_measures after melting the pivot data with the R code melt_fish_measurements.R (Job *Load_fish_measures*).

Fishing environment

The 'fishing_environment' data set includes the information about the origin and conditions of fish collection (e.g. fishing gear, location, date) and is composed of two distinct data sets: (1) *csv_fishing_env_iot.csv* and (2) *Data_Sampling_environment.csv*.

The first data set includes the information retrieved from purse seiners' logbooks and well plans for the fish historically collected at the Seychelles cannery while the second data set includes different kinds of information for all the fish available in the CSV file *Data_Sampling_fish.csv* (section Fish).

For the first data set, a WKT field 'geom_text' was used to aggregate and store the spatial information on the origin of the fish in a textual format. Information on fish origin comes from the location of the fishing operations conducted throughout a purse seiner trip or reported for a brine well where the fish was stored. Different geographic objects were used according to the resolution of information available:

- 1 fishing operation: POINT;
- 2 fishing operations: LINESTRING;
- $>$ 2 fishing operations: MULTIPOINT.
- No information: code WKT_IO used to represent the whole Indian Ocean.

The second data set includes the raw information on fish origin when available, i.e. the geographic position of the fish (e.g. sampling onboard the vessel) or of the fishing operation reported in the purse seiner's well plan. For most of the fish caught with longline, no accurate location of the fishing operations was made available and the extreme positions of the fishing trip were used to define a rectangle polygon of fish origin. For the fish without fishing positions, some qualitative information gathered at the time of sampling was included in the field 'remarks_fishing' and the Exclusive Economic Zone (EEZ) was indicated when possible.

The loading of the data in the database is performed with Talend and consists of three steps: (i) the merging of the two data sets into a CSV file *fishing_env_emotion3.csv*, (ii) the removal of duplicates from the combined fishing environment data set and insertion into the table public.fishing_environment, (iii) the loading of the table mapping the unique fish identifier with the unique identifier of fishing environment into the table public.fish_caught (Job *Load_fishing_env*). The trigger update_geom_fishing_env updates the geometry field 'geom' from 'geom_text' and updates 'geom_uncertainty' which aims to reflect the uncertainty associated with the geographic information available on fish origin.

Sample bank

The sample bank data set contains the information on the preparation and pre-processing of the samples (e.g. drying, storage) and is a CSV export (*Data_Prep.csv*) from the XLSX file Data_Prep.xlsx. The data set is currently loaded into four tables in the database with Talend: (1) public.samples_origin, (2) public.sample_bank, (3) public.sample_grinded_bank and (4) public.sample_dryed_bank (Job *load_sample_bank*).

The part of the data set that establishes the link between the fish and the samples is loaded into the table public.samples_origin[1]. The table public.sample_bank contains the information common to all samples. Information on grinding has been shown to greatly vary between analyses and the general table public.sample_grinded_bank should be removed from the database. The information on grinding will be included in the tables describing some of the analyses when available. Similarly, the table public.sample_dryed_bank should be removed from the analysis. Information on water contents derived from the sample drying will be collated and included in a new CSV file (*Data_moisture.csv*), that will be loaded as a new table analysis_tables.data_moisture.

Analysis

The 18 CSV files exported from the XLSX working files to load into different tables of the database are the following: Data_AminoAcids.csv, Data_Contaminants_Dioxin.csv, Data_Contaminants_HG.csv, Data_Contaminants_Musk.csv, Data_Contaminants_PCBDEOC.csv, Data_Contaminants_PFC.csv, Data_Contaminants_TM.csv, Data_Fatmeter.csv, Data_FattyAcids.csv, Data_LipidClasses.csv, Data_Moisture.csv, Data_Otoliths_counts.csv, Data_Otoliths_morpho.csv, Data_Proteins.csv, Data_Reproduction_fecundity.csv, Data_Reproduction_repro.csv, Data_StableIsotopes.csv, Data_StomachContents.csv. The general columns describing the process of the analyses (e.g. type, lab, operator) are currently loaded into the table analysis_tables.analysis with Talend (Job *load_general_columns*).

Technical details about the analyses (e.g. concentration unit, materials) are loaded into several tracer-specific tables of the schema analysis_tables with Talend (Job *load_analysis_infos*).

All the quantitative results of the analyses identified by the list of 351 tracers available in the table metadata.analysis_tracers_details (section Metadata) are loaded into the table analysis_tables.analysis_measures with Talend after having melted the pivot data with the R script melting_tracers.txt (Job *load_melting_data*).

Spatial layers

The 14 following spatial layers are currently included in the schema geo_data of the database: (1) countries, (2) cwp_grid, (3) eez, (4) eez_boundaries, (5) eez_iho_union_v2, (6) eez_land_v2_201410, (7) fao_fishing_areas, (8) mahe_plateau, (9) rfmos,

[1]{The field table_source is not used and should be removed from the table}

(10) seamounts, (11) seamounts_wessel, (12) world_borders, (13) zet, (14) zet_hors_continents. The spatial layers are used for three reasons:

1. To update the geometry field public.fishing_environment.geom from the WKT field public.fishing_environment.geom_text with the trigger 'update_geom_fishing_env' of the table public.fishing_environment. This concerns the historical data collected in the Seychelles available in the CSV file *csv_fishing_env_iot.csv* (section Fishing environment). The trigger calls the layer geo_data.rfmos but should be reviewed as it generates a convexhull from the MULTIPOINTS which can result in wrong geometries;

2. To update the geometry field public.fishing_environment.geom from the field public.fishing_environment.r_fishing (remarks) with the SQL script update_fishing_environment_when_missing_based_on_remarks_fishing.sql. This occurs when no accurate geographic information was available on the fish origin and keywords (i.e. Ivory Coast EEZ, Mahe Plateau, SYC EEZ, WKT_IO, WKT_AO) were included in the field remarks_fishing of the CSV file *Data_Sampling_environment.csv*. The SQL script calls the layers geo_data.eez and geo_data.rfmos;

3. To extract the data from the database, e.g. within a given exclusive economic zone (EEZ) or within the vicinity of a seamount. It is noteorthy that the layer geo_data.eez_land_v2_201410 is mostly used for the extractions within the Seychelles EEZ as it does not include any part of land (i.e. the islands) and enables to include all the samples collected within the Seychelles waters, i.e. when samples are very coastal and could be apparently come from the land due to the low resolution of the coasts.