BIODATABASE: Loading documentation

Manu Chassot

31 janvier 2020

Metadata

The spreadsheet ‘DDD’ of the XLSX file DDD__Database.xlsx is the Data Dictionary Design of the database. The ‘DDD’ 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 codes 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.

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

Talend Job (load_sample_bank):

- public.sample_bank
- public.sample_dryed_bank: will be loaded from the new file 'table Data_moisture.csv'
- public.sample_grinded_bank to remove from the database

Analysis

Talend Jobs (load_general_columns, load_analysis_infos, load_melting_data):

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

Spatial layers

- eez
- eez_boundaries
- eez_iho_union_v2
- eez_land_v2_201410
- fao_fishing_areas
- mahe_plateau
- rfmos
- world_borders