DRAFT- The SOFIA-TSAF database: a source for global fisheries stocks and area health status records data

1 Introduction

SOFIA-TSAF Database has been designed to document the input, output and compilation of global fishery status health records that will provide reproducibility, visibility, and reliability of this data.   The database is not the database of record or source of record but a tool to locate the data source of record and reproduce similar results. The database is described below.

2 Methods

2.1 Data Collection

The database is a compilation of three tiers of fisheries data and aggregated by Fishery Areas and do not include exact locations.
The tiers include the following:
Tier 1 – Data products from research completed on Fisheries data with the goal of determining the health of a fishery management area or stock
Tier 2 – Data products are comprised of FishStatJ Capture data
Tier 3 – Professional opinion and other sources on the health of a specific fishery
We are currently using Fishery Area 37 as a pilot project to show the process of data import to data reporting. Fishery Area 37 includes the Mediterranean and Black Sea.

2.1 Database Metadata

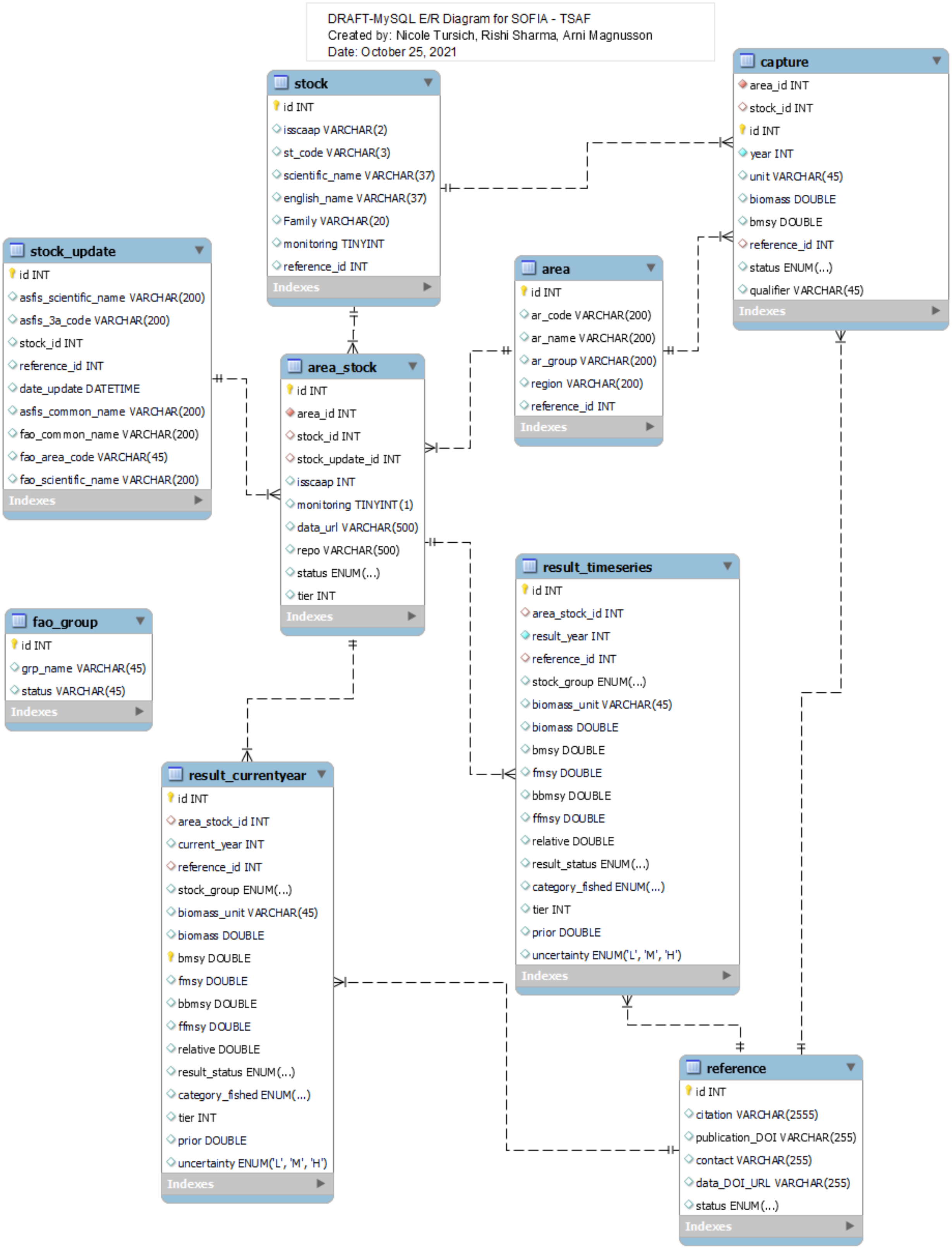
The database, although still in pilot phase, is stored on Oracle MySQL version 8.0.26 Service (MySQL). MySQL is operating system independent and can be installed via command line to both systems with minimal impact. Currently, MySQL sofia-tsaf database includes a relational database structure with 9 linked tables. The tables include: area, stock, stock_update, area_stock, capture, fao_group, reference, result_currentyear and result_timeseries. The database will be stored in a git repository for others to download and update.

The relational database has been created using MySQL recommendations. MySQL recommends using auto-incrementing integers as primary keys thus every table contains a column called ID. In addition to MySQL recommendations, the database has requirements that data records are uploaded with a citation. The tools in this database can be used to look up the data records themselves and then if willing use the MySQL database attachment to compile the data records on personal systems, thereby reproducing the data. The metadata structure is shown in Figure 1.
The tables are described in detail below and can be recreated by following the commands in the order that they appear. NOTE: the order of operation is important due to the relational data structure:

2.2.1 Area

An area is defined in the Global Capture Production dataset downloaded as major fishing area. A zip file containing Capture production data on area as csv can be obtained from: https://www.fao.org/fishery/statistics/global-capture-production/en (add citation).
The process for importing the Area data is to unzip the capture data and run the import shell script for Area pointing to the file downloaded and specifically the file called: CL_FL_WATERAREA_GROUPS.csv. The import script runs the following MySQL commands:

```
LOAD DATA
  LOCAL INFILE 'C:\\tmp\\fish\\Capture_2021.1.2\\CL_FI_WATERAREA_GROUPS.csv'
  INTO TABLE tafp.area
      FIELDS TERMINATED BY ',' ENCLOSED BY '"'
      LINES TERMINATED BY '\r\n'
      IGNORE 1 LINES
      (@ar_Code, @Name_En, @Name_Fr, @Name_Es, @Name_Ar, @Name_Cn, @Name_Ru, @Ocean_Group_En,
      @Ocean_Group_Fr, @Ocean_Group_Es, @Ocean_Group_Ar, @Ocean_Group_Cn, @Ocean_Group_Ru,
      @InlandMarine_Group_En, @InlandMarine_Group_Fr, @InlandMarine_Group_Es, @InlandMarine_Group_Ar,
      @InlandMarine_Group_Cn, @InlandMarine_Group_Ru, @FARegion_Group_En, @FARegion_Group_Fr,
@FARegion_Group_Es,
      @FARegion_Group_Ar, @FARegion_Group_Cn, @FARegion_Group_Ru)
      SET ar_code = @ar_Code, ar_name = @Name_En, ar_group = @InlandMarine_Group_En, region = @FARegion_Group_En,
reference_id = 445;
```



DRAFT-MySQL E/R Diagram for SOFIA - TSAF
Created by: Nicole Tursich, Rishi Sharma, Arni Magnusson
Date: October 25, 2021

Figure 1. Entity/Relationship Diagram for SOFIA-TSAF Database

2.2.2 Stock

Stock data records are found in the AFSIS database and are imported by downloading the 2021 version of the ASFIS_sp.zip file.csv files, unzipping the file and running an import script. The download is found in the following location:
https://www.fao.org/fishery/collection/asfis/en (add citation)
After unzipping the file called ASFIS_sp_year.txt the import routine is started similar to above using a shell script for importing stocks. The import script runs the following MySQL Commands:

```
LOAD DATA
 LOCAL INFILE 'C:\\tmp\\ASFIS_sp\\ASFIS_sp_2021.txt'
 INTO TABLE tafp.stock
     FIELDS TERMINATED BY ',' ENCLOSED BY '"'
     LINES TERMINATED BY '\n'
     IGNORE 1 LINES
     (ISSCAAP, @txocode, 3A_CODE, Scientific_name, English_name,
     @french_name, @spanish_name, @arabic_name,
     @chinese_name, @russian_name, @author, Family, @order, @stat)
     SET isscaap = ISSCAAP, 3a_code = 3A_CODE, scientific_name = Scientific_name,
     english_name = English_name, family = Family, reference_id = 446;
```

2.2.3 Stock_Update

Variations on scientific name were found between what FAO reports using FishStatJ and what ASFIS reports in stocks tables. To resolve this variation, a manual process of importing the species that didn't have a corresponding record in ASFIS to the stock_update table. This way the FAO can continue to report the same species yet we have the most recent scientific name for the species. Comparison of the following files will produce the mismatch on the scientific name for 28 records:
ASFIS_sp_2021.txt' and FAOCatches37 (add citation) and ASFIS_sp_2021.txt and C.tab.all.csv (add citation)
Only 2 of the FAO records for area 37 could not be resolved to ASFIS species and those include:

1. *Psetta maxima*
2. *Venerupis pullastra*

Only *Venerupis Pullastra* was listed in CL_FI_SPECIES_GROUPS.csv (found from the zip file downloaded during the area table creation above)

2.2.4 Area_Stock

TODO

2.2.5 Capture

Capture data records are found in the Global Capture Production dataset downloaded as major fishing area. A zip file containing Capture production data on area as csv can be obtained from: https://www.fao.org/fishery/statistics/global-capture-production/en (add citation).
The process for importing the Area data is to unzip the capture data and run the import shell script for capturedata pointing to the file downloaded and specifically the file called: CAPTURE_QUANTITY.csv. The import script runs the following MySQL commands:

```
DROP TABLE IF EXISTS tafp.tmp_capture;
DROP TABLE IF EXISTS tafp.capture;

CREATE TABLE `capture` (
  `area_id` int unsigned NOT NULL,
  `stock_id` int unsigned DEFAULT NULL,
  `id` int unsigned NOT NULL AUTO_INCREMENT,
  `year` int unsigned NOT NULL,
  `unit` varchar(45) DEFAULT NULL,
  `biomass` double unsigned DEFAULT NULL,
  `bmsy` double unsigned DEFAULT NULL,
  `reference_id` int unsigned DEFAULT NULL,
  `status` enum('current','superseded','modified') DEFAULT NULL,
  `qualifier` varchar(45) DEFAULT NULL,
  PRIMARY KEY (`id`),
  KEY `fk_capture_area_idx1` (`area_id`),
  KEY `fk_capture_stock_idx2` (`stock_id`),
  KEY `fk_capture_ref_idx2` (`reference_id`),
  CONSTRAINT `fk_capture_area_id` FOREIGN KEY (`area_id`) REFERENCES `area` (`id`) ON DELETE CASCADE ON UPDATE
CASCADE,
  CONSTRAINT `fk_capture_ref_id` FOREIGN KEY (`reference_id`) REFERENCES `reference` (`id`) ON DELETE CASCADE ON
UPDATE CASCADE,
  CONSTRAINT `fk_capture_stock_id` FOREIGN KEY (`stock_id`) REFERENCES `stock` (`id`) ON DELETE CASCADE ON UPDATE
CASCADE
) ENGINE=InnoDB AUTO_INCREMENT=6396639 DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_0900_ai_ci;

CREATE TEMPORARY TABLE tafp.tmp_capture(
  id int(11) unsigned NOT NULL AUTO_INCREMENT,
  cap_stock VARCHAR(200) DEFAULT NULL,
  cap_area VARCHAR(200) DEFAULT NULL,
  cap_year INT DEFAULT NULL,
  cap_value DOUBLE DEFAULT NULL,
  cap_measure VARCHAR(45) DEFAULT NULL,
  cap_qualifier VARCHAR(45) DEFAULT NULL,
  PRIMARY KEY (`id`)
) ENGINE=InnoDB AUTO_INCREMENT=295 DEFAULT CHARSET=utf8mb4;
LOAD DATA
  LOCAL INFILE 'C:\\tmp\\Capture_2021.1.2\\CAPTURE_QUANTITY.csv'
  INTO TABLE tafp.tmp_capture
      FIELDS TERMINATED BY ',' ENCLOSED BY '"'
      LINES TERMINATED BY '\r\n'
      IGNORE 1 LINES
      (@UN_CODE,@ALPHA_3_CODE,@AREACODE,@MEASURE,@PERIOD,@VALUE,@STATUS)
      SET cap_stock = @ALPHA_3_CODE, cap_area = @AREACODE, cap_year = @PERIOD,
      cap_value = @VALUE, cap_measure = @MEASURE, cap_qualifier = @STATUS;

INSERT INTO tafp.capture(area_id, stock_id, `year`, biomass, unit, qualifier, reference_id)
```

```
SELECT ar.id, st.id, tc.cap_year, tc.cap_value, tc.cap_measure, tc.cap_qualifier, 445
FROM tafp.tmp_capture tc
LEFT JOIN tafp.area ar on tc.cap_area = ar.ar_code
LEFT JOIN tafp.stock st on tc.cap_stock = st.st_code;
```

2.2.6 FAO-Group

Group data records are found in the AFSIS database and subsequently in column isscaap column in the Stock Table. This is related back to the group table via the id. The names of the group data was found in the CL_FI_SPECIES_GROUPS.csv (found from the zip file downloaded during the area table creation above). An MySQL script was created to import the data records and includes the following:

```
insert into tafp.fao_group(id, grp_name, status) values (11,'Carps, barbels and other cyprinids','current');
insert into tafp.fao_group(id, grp_name, status) values (12,'Tilapias and other cichlids','current');
insert into tafp.fao_group(id, grp_name, status) values (13,'Miscellaneous freshwater fishes','current');
insert into tafp.fao_group(id, grp_name, status) values (21,'Sturgeons, paddlefishes','current');
insert into tafp.fao_group(id, grp_name, status) values (22,'River eels','current');
insert into tafp.fao_group(id, grp_name, status) values (23,'Salmons, trouts, smelts','current');
insert into tafp.fao_group(id, grp_name, status) values (24,'Shads','current');
insert into tafp.fao_group(id, grp_name, status) values (25,'Miscellaneous diadromous fishes','current');
insert into tafp.fao_group(id, grp_name, status) values (31,'Flounders, halibuts, soles','current');
insert into tafp.fao_group(id, grp_name, status) values (32,'Cods, hakes, haddocks','current');
insert into tafp.fao_group(id, grp_name, status) values (33,'Miscellaneous coastal fishes','current');
insert into tafp.fao_group(id, grp_name, status) values (34,'Miscellaneous demersal fishes','current');
insert into tafp.fao_group(id, grp_name, status) values (35,'Herrings, sardines, anchovies','current');
insert into tafp.fao_group(id, grp_name, status) values (36,'Tunas, bonitos, billfishes','current');
insert into tafp.fao_group(id, grp_name, status) values (37,'Miscellaneous pelagic fishes','current');
insert into tafp.fao_group(id, grp_name, status) values (38,'Sharks, rays, chimaeras','current');
insert into tafp.fao_group(id, grp_name, status) values (39,'Marine fishes not identified','current');
insert into tafp.fao_group(id, grp_name, status) values (41,'Freshwater crustaceans','current');
insert into tafp.fao_group(id, grp_name, status) values (42,'Crabs, sea-spiders','current');
insert into tafp.fao_group(id, grp_name, status) values (43,'Lobsters, spiny-rock lobsters','current');
insert into tafp.fao_group(id, grp_name, status) values (44,'King crabs, squat-lobsters','current');
insert into tafp.fao_group(id, grp_name, status) values (45,'Shrimps, prawns','current');
insert into tafp.fao_group(id, grp_name, status) values (46,'Krill, planktonic crustaceans','current');
insert into tafp.fao_group(id, grp_name, status) values (47,'Miscellaneous marine crustaceans','current');
insert into tafp.fao_group(id, grp_name, status) values (51,'Freshwater molluscs','current');
insert into tafp.fao_group(id, grp_name, status) values (52,'Abalones, winkles, conchs','current');
insert into tafp.fao_group(id, grp_name, status) values (53,'Oysters','current');
insert into tafp.fao_group(id, grp_name, status) values (54,'Mussels','current');
insert into tafp.fao_group(id, grp_name, status) values (55,'Scallops, pectens','current');
insert into tafp.fao_group(id, grp_name, status) values (56,'Clams, cockles, arkshells','current');
insert into tafp.fao_group(id, grp_name, status) values (57,'Squids, cuttlefishes, octopuses','current');
insert into tafp.fao_group(id, grp_name, status) values (58,'Miscellaneous marine molluscs','current');
insert into tafp.fao_group(id, grp_name, status) values (61,'Blue-whales, fin-whales','current');
insert into tafp.fao_group(id, grp_name, status) values (62,'Sperm-whales, pilot-whales','current');
insert into tafp.fao_group(id, grp_name, status) values (63,'Eared seals, hair seals, walruses','current');
insert into tafp.fao_group(id, grp_name, status) values (64,'Miscellaneous aquatic mammals','current');
insert into tafp.fao_group(id, grp_name, status) values (71,'Frogs and other amphibians','current');
insert into tafp.fao_group(id, grp_name, status) values (72,'Turtles','current');
insert into tafp.fao_group(id, grp_name, status) values (73,'Crocodiles and alligators','current');
insert into tafp.fao_group(id, grp_name, status) values (74,'Sea-squirts and other tunicates','current');
insert into tafp.fao_group(id, grp_name, status) values (75,'Horseshoe crabs and other arachnoids','current');
insert into tafp.fao_group(id, grp_name, status) values (76,'Sea-urchins and other echinoderms','current');
insert into tafp.fao_group(id, grp_name, status) values (77,'Miscellaneous aquatic invertebrates','current');
insert into tafp.fao_group(id, grp_name, status) values (81,'Pearls, mother-of-pearl, shells','current');
insert into tafp.fao_group(id, grp_name, status) values (82,'Corals','current');
insert into tafp.fao_group(id, grp_name, status) values (83,'Sponges','current');
insert into tafp.fao_group(id, grp_name, status) values (91,'Brown seaweeds','current');
insert into tafp.fao_group(id, grp_name, status) values (92,'Red seaweeds','current');
insert into tafp.fao_group(id, grp_name, status) values (93,'Green seaweeds','current');
insert into tafp.fao_group(id, grp_name, status) values (94,'Miscellaneous aquatic plants','current');
```

2.2.7 Reference

This section is the most important part of the database and it is imperative to stick to the data import process. The columns in this table include the citation, publication _DOI, contact, data url and status on if the reference is still used. A shell script is provided to import a file containing the following columns: citation, publication _DOI, contact, data url and status. The script requires a file location and directory as input parameters.

2.2.8 Result_CurrentYear

The data comes from a couple sources. For Tier 1 data, the records for this will have to be manually created and gathered from previous published reports. This involved creating a reference similar to above in the reference section for the data gathered and ensuring the stocks and areas are included in the area_stock table. The import process is described below in the import section.

For Tier 2 data, the data is stored in a github repository called sofia-tsaf. Each area has its own repository and output folder. In the output folder there is a file called: current_status.csv. use the import shell script to import these records.   This process is to be automated as the pilot progresses.

2.2.9 Result_TimeSeries

Similar to above, for Tier 1 data, the records for this will have to be manually created and gathered from previous published reports. This involved creating a reference similar to above in the reference section for the data gathered and ensuring the stocks and areas are included in the area_stock table. The import process is described below in the import section.

For Tier 2 data, there is a timeseries report that is housed the same location in the output folder as described above except the output is called stock_timeseries.csv.

2.3 Quality Control

The data included in the database has mostly been gathered from reputable sources including published sources from professionals in the fishery field. Each data record is assigned to a citation via the references table. Additionally, on creation of the database the data in this report was imported using shell scripts that checked the relational data. The data were also exported and checked row by row with each original source. One minor difference did appear and that was deemed non-issue and that was the rounding of data we received as a source to check data imported. The database does not round and although there can be truncation errors due to the nature of the double column data type we deem the data checked. There is only one significant difference found in the tier 1 data and that is being looked at currently and is shown below:

Area		
47	Southern African pilchard	Sardinops ocellatus
61	Japanese pilchard	Sardinops melanostictus
77	California pilchard	Sardinops caeruleus

2.4 Import process

There is a somewhat automated process for importing fisheries data as shown above in the table descriptions. The following steps should be followed in order to add more data to the data structure and allow for more tier data comparisons and stock health reporting.

2.4.1 Tier 1 Importing

Tier 1 data mostly will include data records that are reported in the Result_CurrentYear and Result_TimeSeries. As this data has already been compiled and summarized elsewhere it was not deemed necessary to report raw data.

1. Ensure that the stock and area of interest has records in the area_stock table for tier 1 using this command: area, scientific_name, category_fished, ffmsy, bbmsy, uncertainty, referenceid
2. Import reference to the reference table for each source of data to be used using reference import shell script as described above in the reference table metadata.
3. Obtain the reference ids for the data to be added to the Result_CurrentYear and Result_TimeSeries. Use this referenceid in step 1.

4. Construct a csv table with the following columns to be imported to Result_CurrentYear or Result_TimeSeries:
5. Run import shell script for Tier 1 data using the parameters for filename, file location.

2.4.2 Tier 2 Importing

The import process for tier 2 data comes directly from the FAO FishStatJ database and results from the sofia-tsaf analysis.

2.4.3 Tier 3 Importing

To be determined.

3 Conclusions

To be determined.

4 References

To be determined.