Introduction to the IOTC tuna factory purchases data flow and database

Lucia Pierre[1], Nathalie Bodin[2], Fabio Fiorellato[1]

[1] IOTC Secretariat, Le Chantier Mall, PO Box 1011, Victoria, Mahé, Seychelles
[2] Data for Ecosystem and Ocean Health consulting (DEAL), 66 rue Anatole France, 37540 Saint Cyr sur Loire, France

Abstract

Tuna factory purchases data constitute a complementary source of independent information in support of IOTC tuna fisheries analyses. This novel data source is aimed to be used routinely for future assessment and to cross-verify and reduce uncertainties in the currently available statistical data (in particular for what concerns species composition). A total of 45 companies have been submitting tuna quarterly reports to IOTC secretariat since 2010. Here, we present the IOTC tuna factory purchases data flow and database, including the different steps of data harmonization, compilation and preliminary curation undertaken on the quarterly reports to improve the overall data quality and traceability to the original information source. Between 2010 and 2020, 72% of the total number of purchases records received at the Secretariat came from the Indian Ocean, and the rest from the Atlantic and Pacific Oceans. The Indian Ocean 2010-2020 reports revealed that purse-seine and pole-and-line represented the majority of the total number of records (63% and 21%, respectively), and were dominated by skipjack (37%), followed by yellowfin (31%), bigeye (15%), and albacore (14%) tunas. Moreover, around 75% of purchases data from the Indian Ocean reports were harmonized into four species-specific commercial weight categories for each of the four major tuna species. Next project steps aim to finalize data curation and conduct analyses including comparisons of the tuna purchases data against the IOTC fisheries statistics by species and gear.

1. Introduction

The estimation of the species composition of tuna catch is required to provide accurate species-specific data to the IOTC Secretariat for fishery data analysis. In the Indian Ocean, the catch composition for most tuna fisheries, and in particular for tropical tuna species, is derived from tuna size samples collected during unloading operations, and possibly during the fishing activity by onboard observers, and used within a processing procedure that has been recently challenged (Fonteneau et al., 2017; Herrera and Báez, 2018).
Data from tuna factory purchases are thought to provide a useful ancillary source of independent information to compare and validate the species and catch size category composition breakdowns by fleet, as these data are retrieved from vessels' logbooks and landings as well as from tuna size samples. Hence, the assessment of tuna factory purchases data as a potential validation tool for longline and purse-seine catch data has been initiated in the Western-Central Pacific Ocean (Lewis and Williams, 2016; Williams, 2020), Indian Ocean (IOTC Secretariat 2013, 2020), and Atlantic Ocean (Bodin et al., 2021). Recently, the use of tuna factory purchases data has been proposed to be included as an integral component of the purse seine data processing (Báez et al., 2020).

Since 2010, the IOTC Secretariat has been receiving tuna factory reports on a quarterly basis from ISSF-participating companies that source tuna caught in the Indian Ocean. For example, a quarterly report sent by the company during quarter 2 (between April and June) of a specific year is thought to gather all tuna purchases

that occurred during quarter 1 (January-March) of the same year. The tuna factory reports include species-specific data in total weight and, in most cases, by commercial categories (e.g., 1.8-3.4kg, >10kg, etc.) collected by the company, as well as information on its origin, i.e., fishing and unloading operations (dates, gear, ocean where the fish was caught, fishing and carrier vessel's name identification, unloading location, etc.). Overall, the data structure and contents of the quarterly reports vary greatly between submitting companies and years, with several data reports that appear to suffer from a generalized lack of standardization and quality control at the source, resulting in frequent inconsistencies or incomplete data (e.g., lack of mandatory information on fishing and/or unloading operations, different date formats which are used often incomplete, etc.).

To overcome these issues encountered by IOTC and other tuna Regional Fisheries Management Organizations (tRFMOs) receiving ISSF-participating companies' tuna factory reports, different data management tools were developed (IOTC Secretariat, 2020). First, a new uniform submission template required by ISSF since 2019, generally improved the harmonization of the structure and contents of the ISSF-participating companies' quarterly reports submitted between mid-2019 and mid-2021. Second, an interactive electronic tool was developed by the IOTC Secretariat to help harmonizing the tuna factory quarterly reports. Starting from a generalized data model of the expected information, the tool allows end-users to import original data submissions and format their content (one file at a time) to produce reports with standardized data fields and code lists. Third, a relational database was designed to improve data quality and reusability.

In such a context, an IOTC project was set up with the overall objective of assessing the use of tuna factory purchases as a complementary data source, both in terms of volume and the breakdown of catch by species, in support of IOTC tuna fisheries analyses. This novel data source is aimed to be used routinely for future assessment and for reducing uncertainties in the currently available statistical data. The project is composed of three main components: (i) component 1 aims to harmonize and compile all tuna factory quarterly reports submitted by the ISSF-participating companies during 2010-2021; (ii) component 2 aims to upload harmonized reports into a relational tuna factory purchases database and proceed to further data curation so as to improve the overall data quality and traceability to the original information source; and finally (iii) component 3 corresponds to an ongoing MSc research project which aims to compare the species and size/weight composition available from the tuna factory purchases with the annual catch data reported to the IOTC secretariat for each fishing gear (IOTC Secretariat, 2020)

Here, we present the preliminary results of the components 1 and 2 of the project. We first summarize the steps required for the harmonization and compilation of the ISSF-participating companies' tuna factory quarterly reports, and present the model of the relational database. We then provide a description of the tuna factory purchases data and related metadata available at the IOTC secretariat for the period 2010-2020. Finally, we present the next steps, in terms of data curation, that will be undertaken as part of the IOTC tuna factory data project, and we propose potential improvements in the data submission and processing with the aim to facilitate the work for all stakeholders and ensuring a better quality, traceability, and reusability of the tuna factory purchases data.

2. Harmonization and compilation of the quarterly reports into a relational database for tuna factory purchases data

A total of 1,479 quarterly reports from 45 ISSF-participating companies were submitted to IOTC during the study period, including 522 emails and 957 data files (using either .xls, .xlsx. or .xlm formats).
First, a review of the ISFF-participating companies' metadata collated from the title and/or content of the emails and data files was conducted, allowing the identification of report duplicates. In that case, the more recent report was selected for further processing while the other reports were archived (and subsequently, not processed).

The remaining (i.e., non-archived) reports were then harmonized through the IOTC electronic tool specifically designed for this task, so that the variables were organized into tables based on their functional dependencies to fit the relational database model and provide results that would be ready to be analyzed (**Annex 1**) (see IOTC Secretariat, 2020 for examples of the different formats of reports submitted at the IOTC). For that purpose, each quarterly report was uploaded into the software and each column (i.e., variable) of the original report was first re-defined according to standards (e.g., location, country, gear and species standardized with FAO codes and names), and then re-structured following the tidy framework (each observation a row, each column a variable, each table a type of observational unit) (Wickham, 2014). The IOTC electronic tool also allows to identify, correct and track obvious data errors and inconsistencies in the entries (see IOTC Secretariat, 2020 for a presentation of the electronic tool during processing of report). A configuration file is thus created for each quarterly report, and the results of processing the submitted quarterly report with the specified configuration are finally uploaded into a relational database composed of 13 tables (**Annex 2**).

As a result, a total of 622 reports (i.e., 65% of the total quarterly reports submitted for the study period) were successfully harmonized and compiled into the database; the rest of the quarterly reports (n = 335; 35%) were archived because purchases data were missing from the report or an entire report was duplicated. Among the successfully harmonized quarterly reports, 71% (n = 439) were processed directly and 29% (n = 183) required additional editing of the original data files before further processing. Three main types of corrections were made to fit the minimum data requirement and structure of the software: (i) correction of date format; (ii) modification of sheet format (unmerged cells); (iii) other (e.g., separation of purchases according to, for example, product type or type of unloading).

3. Description of the raw content of the tuna factory purchases database for the period 2010-2020

The number of ISSF-participating companies providing tuna factory quarterly reports to IOTC increased from 4 in 2010 to 24 in 2019 (**Figure 1A**). The number of successfully harmonized quarterly reports increased from 10 in 2015 to 103 in 2017-2018, and decreased to 75 in 2020 (**Figure 1B**). This resulted in a total of 114,637 raw tuna factory purchases records uploaded into the database. At this stage, the data include some duplicate records, i.e., records that have been repeated in several distinct quarterly reports together with non-duplicate records. The curation of the tuna factory purchases database will ensure the removal of these duplicates as well as other data quality improvements (see section 4).

Over the study period, 81% of the total number of quarterly reports to IOTC, i.e., 72% of the total number of raw tuna factory purchase records (n = 82,775), came from the Indian Ocean, and the rest concerned other tRFMOs (ICCAT, IATTC and WCPFC) (**Figure 1B**).

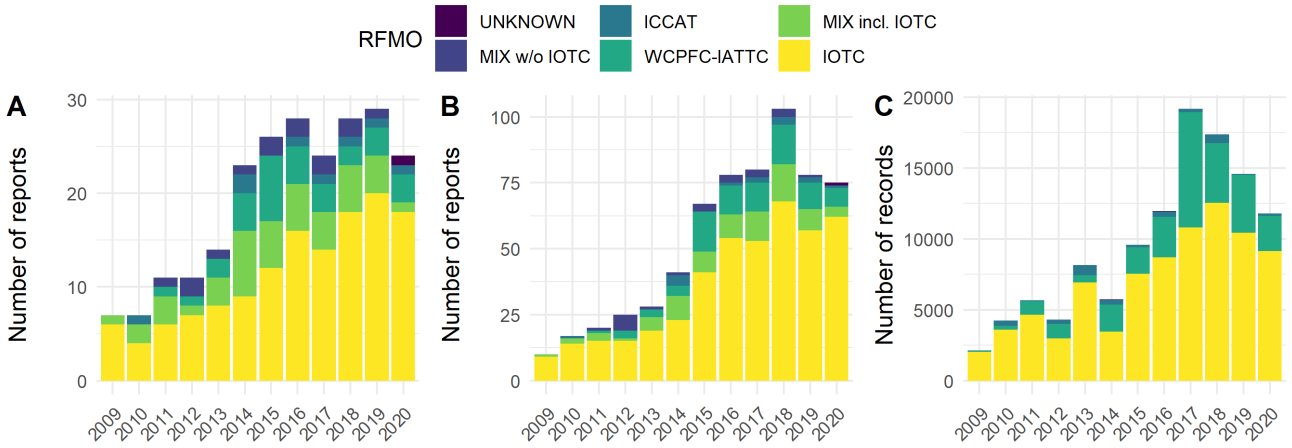


Figure 1. Number of ISSF-participating companies that provided data to the IOTC secretariat (A), number of reports received and successfully processed (B), and derived number of raw tuna factory records (C), per year and per tRFMO area during 2010-2020.

The analysis of the Indian Ocean quarterly reports revealed that 44%, 15% and 13% of the total number of reports for the 2010-2020 period came from purse-seine, longline and pole-and-line purchases, respectively, and 27% from a mix of gears (i.e., purse-seine, longline, pole-and-line, handline and troll line) (**Figure 2A**). With regards to the number of records, purse-seine represented the majority of the tuna factory purchases (63%), followed by pole-and-line (21%), longline (15%), and handline and troll line (< 1%) (**Figure 2B**).

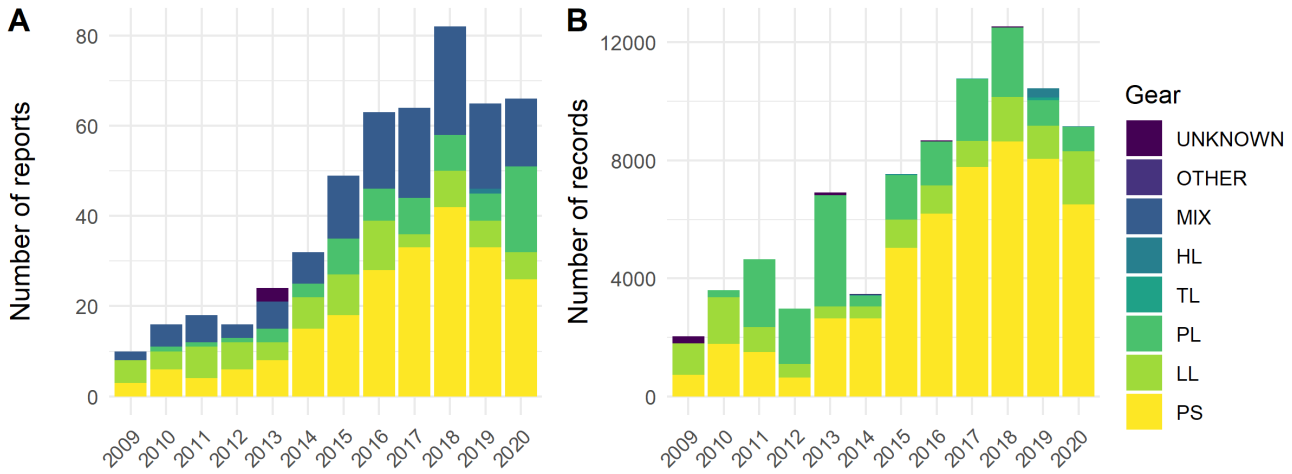


Figure 2. Number of processed reports (A) and derived number of raw tuna factory records (B) for the IOTC area per year and per type of fishing gear during 2010-2020. LL: Longline; PL: Pole-and-line; PS: purse-seine; HL: Handline; TL: Troll line.

Although the information on the type of product was missing from the reports and records (39% and 44%, respectively) for a majority of the tuna factory quarterly reports, the remaining ones included whole round tuna (48% and 52%) or tuna loins (2% and 4%) purchases (**Figure 3**).

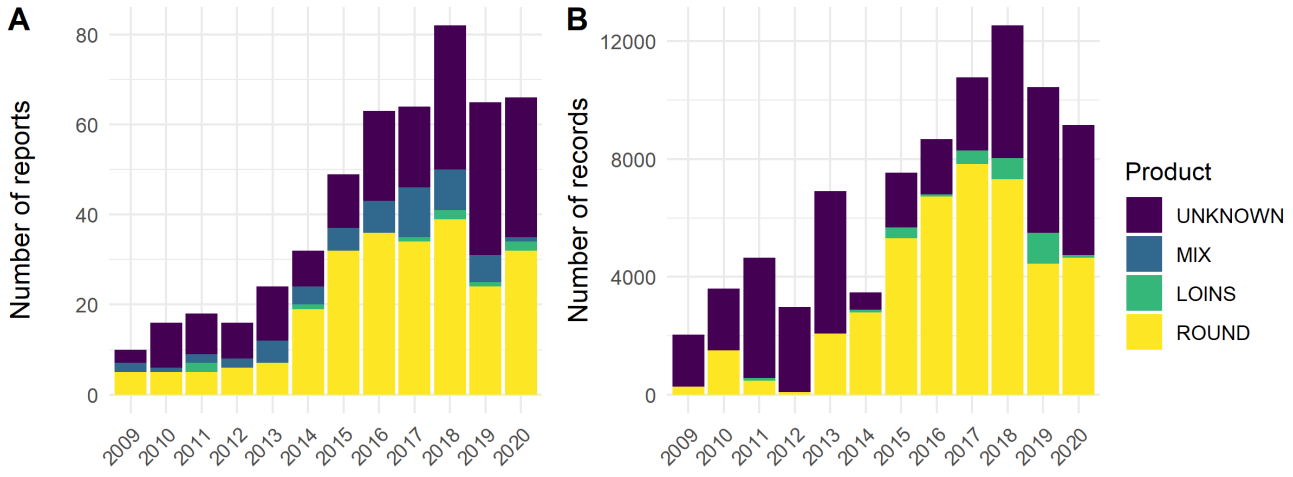


Figure 3. Number of processed reports (A) and derived number of raw tuna factory records (B) for the IOTC area per year and per type of product (loins, round, mix -loins and round-, and unknown) during 2010-2020.

Overall, the IOTC related tuna factory purchases were dominated by skipjack and yellowfin tuna (37% and 31% of the total number of tuna factory records, respectively), followed by bigeye tuna (15%) and albacore tuna (14%) (**Figure 4A**). The analysis per gear type revealed the dominance of albacore tuna purchases (97%) for longline, skipjack and yellowfin tuna purchases (59% and 40%, respectively) for pole-and-line, and a mix of skipjack (37%), yellowfin (35%) and bigeye (24%) tuna purchases for purse-seine vessels (**Figure 4B-D**).

Figure 4. Number of raw tuna factory records for IOTC area per year and per species during 2010-2020. SKJ: skipjack tuna; YFT: yellowfin tuna; BET: bigeye tuna; ALB: albacore tuna; Other sp: includes frigate tuna, longtail tuna, kawakawa and other species.

Most of the tuna purchases data included information on commercial weight categories, although a large number of different categories and units were used by the companies, as well as within the same company (i.e., for a given company, purchase and commercial weight category units may differ between the quarterly reports and within a unique quarterly report making even more challenging the harmonization and curation of the factory data). Most of the submitted factory reports were successfully harmonized into four species-specific commercial weight categories for the major tuna species (i.e., 79% of the total skipjack tuna records, 87% of the total yellowfin tuna records, 89% of the total bigeye tuna records and 85% of the total albacore tuna records); the rest was grouped as *other* when only total weight was available or when the weight category used by the company was not usable (e.g., the range of the weight category was too large such as <10kg for yellowfin and bigeye tunas) (**Table 1**). **Figure 5** shows for example the composition of the tuna factory purchases for the purse-seine vessels that caught the major tropical tuna species in the Indian Ocean during 2010-2020.

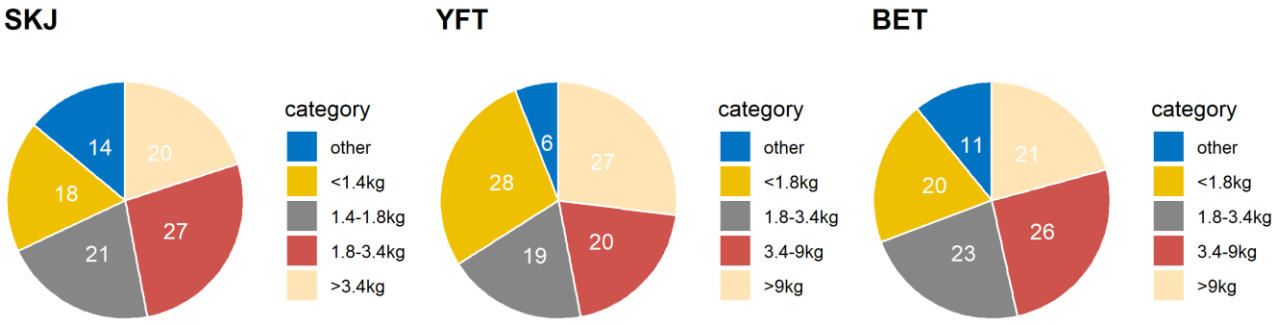


Figure 5. Share (%) of harmonized weight categories in the factory records of tropical tunas caught with purse-seine in the Indian Ocean during 2010-2020. Commercial weight categories were standardized into four categories (kg) for each species. "other" corresponds to records where only total weight was available or where the weight category used by the company was not exploitable. SKJ = skipjack tuna; YFT = yellowfin tuna; BET = bigeye tuna.

4. Next steps: curation of the tuna factory purchases database and connection with other tuna fisheries data sources

The final clean-up of the tuna factory purchases database is essential to ensure good data quality before proceeding with any further data analysis, and indeed constitutes the next step of component 2 of the project. Data curation will include automatic data standardization and compilation, and identification/suppression of the duplicate data that could not be identified and corrected with the IOTC electronic tool. Finally, data from the tuna factory database (properly anonymized) will be connected with other tuna fisheries data sources for scientists to undertake required comparative data analysis (component 3).

5. Recommendations for the improvement of the quality and management of the factory purchases data

Here we present potential recommendations at different steps of the data management process aimed at improving the quality, traceability and reusability of the tuna factory purchases data and facilitate future work for all involved stakeholders.

Submission of the quarterly reports

The 2019 submission template for tuna factory purchases data generally improved the harmonization of the structure and contents of the ISSF-participating companies' quarterly reports. However, issues remain with the quality of the data (e.g., date intervals are wrong or overlap with each other, fishing vessel and carrier information is wrong or incomplete), and additional information would be required to fit the management database requirements and maximize the quantity of information to be usable for data analysis. **Table 2** lists some recommendations proposed for the improvement of the submitted quarterly reports.

Table 2. List of recommendations proposed to improve ISSF-participating company quarterly submission reports.

Recommendation	Description of the recommendation
Design a submission template for metadata	A submission template for the metadata should be designed, and the company should submit both the metadata and data quarterly reports. Also, a unique ID should be assigned to each company and factory to facilitate reports traceability. Metadata should include information on the company name, location and unique ID, email contact, RFMO of concern, year-quarter concerned by the report, type of unloading (Direct unloading vs Transshpiment-unloading), product type as required by the management tools. The format of the metadata template must be designed to directly feed the data management tools
Standardize file names of the submitted data and metadata reports	The standardization of the name of the quarterly report submitted by the company would greatly facilitate the identification of the duplicate reports and the control of regular reporting. This can be achieved by defining for example an hierarchical file name nomenclature of type "[CompanyCode][Year][Quarter][{new, rev}][…]".
Modify the format of the data submission template for dates	The separation and formatting (yyyy-mm-dd) of the start and end dates in the 2019 data submission template greatly limited the number of date format issues encountered during the data processing, however issues still remain mainly due to Excel date formatting options. The use of a validation rule on the dates could reduce date errors and ensure better data quality and traceability. Also, the use of separate columns for Day, Month and Year for each reported date in addition to the international standard for dates (yyyy-mm-dd) could constitute an additional data verification.
Modify the format of the data submission template for vessel information numbers	The use of tidy data for the unique vessel identifier -UVI- (i.e., one column for the UVI type and one column for the UVI value) restraints the company to provide only one identifier while several may be available. The data submission report should include at least three separate columns for IMO, IRCS and other registration numbers as a maximum information is highly recommended to ensure cross-checking of the vessel information
Include essential additional information to be recorded into the data submission report	Add information on the factory that processed catch (name and location - city/country) as the quarterly report may contain catch/purchases data from more than one location (several factories for one company) Add information on the type of product as required for the normalization and comparison of purchases data with fishing data (round fish) Add information on purchases data unit to avoid errors, especially in the case of data quarterly reports that may contain purchases data with different units within a same report
Separate data submission templates for each type of unloading	Use two separate data submission templates for each type of unloading (Direct unloading vs Transshpiment-unloading) to avoid data entry errors and confusion for the data curator
Modify the format of the data submission template for purchases data	The 2019 data submission template imposes the type of species and commercial weight categories to be reported by the company. Purchases data for other species are thus lost while they were provided by several companies prior to 2019. The information may be of interest for future data analysis and could be added into the data submission template Information on fishing mode (FAD/FREE) and catch certification (MSC) was provided by some companies prior to the 2019 data submission template. This information may be of interest for future data analysis and could be added into the data submission template
Define a formal workflow for the submission of corrections and updates	Currently, companies might re-submit data for factories and / or periods of times that were already submitted in the past, to account for corrections or updates. As there's no clear indication of whether an update should be *partial* (i.e. include only records that have been updated) or *complete*, as well as how to deal with the request of *removing* records that were submitted by mistake in the past, a data submission workflow should be defined (by ISSF, with support from the t-RFMOs and / or the companies) in order to streamline the process of providing data updates in the future.

A database for the ISSF-participating companies

The increasing number of ISSF-participating companies submitting purchase information and contributing to the tuna factory database and the complexity and regular change of the companies' affiliation challenges the traceability of the submitted quarterly reports and data. A company database was created including information on the name, acronym and location of the factory and of the ownership company(ies). The company database will have to be regularly updated in collaboration with the companies and could also be completed with additional information (e.g., main processing activities in terms of product and species of interest). Moreover, an estimation of the coverage of the ISSF-participating companies compared to the total number of tuna processing companies operating in the Indian Ocean would inform on the limits of the tuna factory purchases database and allow to apply appropriate correction and analysis. In the Pacific Ocean, ISSF-participating companies' tuna factory data represented on average ~37% of the tropical WCPFC purse seine catch in recent years (Williams, 2020), and it was recommended to encourage the voluntary participation of non-ISSF-participating companies to the factory purchases data flow (Lewis, 2017; Williams, 2020). Finally, it is proposed to strengthen the collaboration and communication with participating tuna companies, for instance through annual summary reports providing an overview of the quantity and quality of the submitted quarterly data and metadata reports.

6. Conclusion

The provision of scientific management advice has been recently challenged with regards to the quality of the data currently used by tRFMOs for stock assessments. The use of complementary independent data sources, such as the tuna factory purchases already collected by the companies, would strengthen data confidence and improve tuna data analysis.

With the collaboration of the ISSF-participating companies, a large data set of more than 114,000 tuna purchases records for the 2010-2020 period, was reviewed, organized, harmonized, compiled and stored into a specifically-designed database that shall simplify future analysis. Being routinely collected by the companies as part of their trading operations, tuna purchases data constitute undoubtedly a unique and essential source of complementary information for tuna fisheries stock assessment and management. The present study however highlighted the crucial step of transforming the raw data into standardized curated tuna factory data before further analysis. During the next months, the project will finalize the curation of the tuna factory purchases database, and proceed with the exploratory analyses work which includes preliminary comparisons of the tuna purchases against the fisheries statistics available at the IOTC Secretariat by species and gear, both at a global level (total catches) and disaggregated level (e.g., weight commercial categories *vs* biological size sampling and/or catch-at-size).

Acknowledgments

This work was co-funded by the IOTC Secretariat and the International Seafood Sustainability Foundation (ISSF). The work was conducted independently by the authors, and its results, professional opinions and conclusions are solely the work of the authors. There are no contractual obligations between ISSF and the authors that might influence the results, professional opinions, and conclusions.

References

Báez, J.C., Ramos, M.L., Herrera, M., Murua, H., Cort, J.L., Déniz, S., Rojo, V., Ruiz, J., Pascual-Alayón, P.J., Muniategi, A., San Juan, A.P., Ariz, J., Fernández, F., Abascal, F., 2020. Monitoring of Spanish flagged purse seine fishery targeting tropical tuna in the Indian ocean: Timeline and history. Marine Policy 119, 104094.

Bodin, N., Fiorellato, F., Palma, C., Mayor, C., 2021. Introduction to the ICCAT tuna factory unloadings data flow and database. Collect. Vol. Sci. Pap. ICCAT SCRS/2021/064.

Fonteneau, A., Alayón, P.P., Chassot, E., 2017. From large fixed to small mobile spatio-temporal strata: improving estimates of species and size composition of the landings of the European purse seine fishery in the Atlantic Ocean. Collect. Vol. Sci. Pap. ICCAT 73, 829-849.

Herrera, M., Báez, J.-C., 2018. On the potential biases of scientific estimates of catches of tropical tunas of purse seiners of the EU and other countries report to the ICCAT and IOTC. IOTC Working Party on Tropical Tunas (WPTT20), Mahé, Seychelles, 29 October - 3 November 2018, p. 40.

IOTC Secretariat, 2013. Report and documentation of the Indian Ocean Tuna Fisheries of Indonesia Albacore Catch Estimation Workshop: Review of Issues and Considerations. Bogor-Jakarta, 21-25 June 2013, 40 pp.

IOTC Secretariat, 2020. Research proposal: an evaluation of data from ISSF-affiliated canneries for use in tuna fisheries management. Presented at the IOTC Working Party on Data Collection and Statistics (WPDCS16), Online, 30 November - 3 December 2020, p. 10.

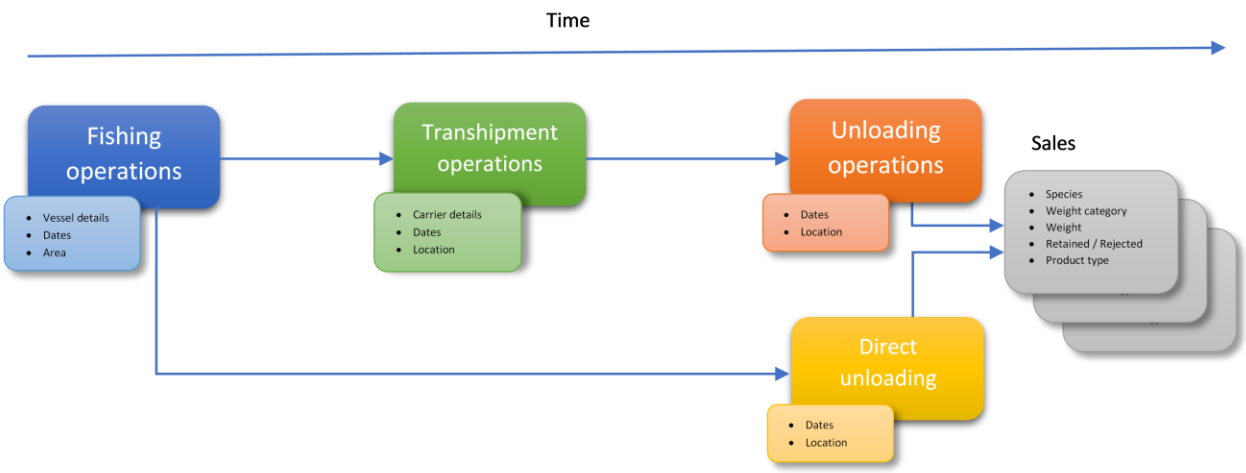
Lewis, A., 2017. Pilot Study of the Potential for using Non-ISSF Associated Cannery Receipt Data for the work of the WCPFC. WCPFC, Rarotonga, Cook Islands 9-17 August 2017, p. 23.

Lewis, A., Williams, P., 2016. Potential use of cannery receipt data for the scientific work of the WCPFC. WCPFC, Bali, Indonesia, 3-11 August 2016, p. 19.

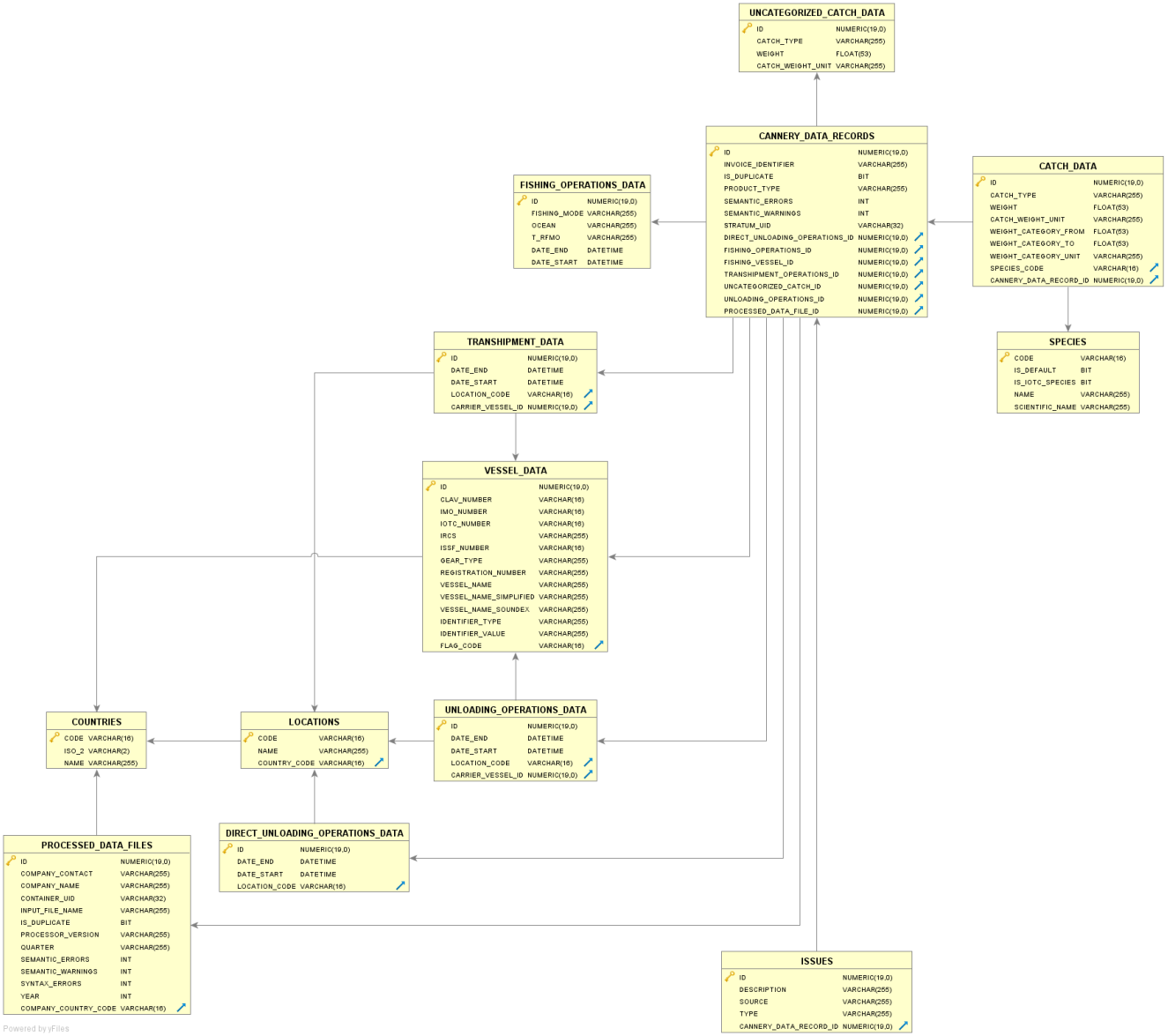
Wickham, H., 2014. Tidy Data. Journal of Statistical Software 59, 1–23.

Williams, P., 2020. An update on cannery data with potential use to the WCPFC. WCPFC, Online Meeting 11-20 August 2020, p. 11.

ANNEX 1: Overview of the tuna purchases data structure and timeline of its collection and reporting workflow



ANNEX 2: Relational model of the tuna factory purchases database



ANNEX 3: Number of records for the IOTC area per species and commercial weight categories

Species	Category	Number of records	% of the total number of records
BET	<1.8kg	2554	20
	>9kg	2637	21
	1.8-3.4kg	2869	23
	3.4-9kg	3199	25
	other	1392	11
YFT	<1.8kg	5923	23
	>9kg	6633	26
	1.8-3.4kg	5529	21
	3.4-9kg	4282	17
	other	3484	13
SKJ	<1.4kg	3842	13
	>3.4kg	4204	14
	1.4-1.8kg	4324	14
	1.8-3.4kg	11634	38
	other	6378	21
ALB	<4kg	62	1
	>10kg	8055	68
	4-7kg	425	4
	7-10kg	1377	12
	other	2001	17