Improving the core IOTC data management processes

PREPARED BY: IOTC SECRETARIAT[1], 18TH NOVEMBER 2016

Summary

The current state of the art related to the internal IOTC core data management processes is described, depicting benefits and shortcomings as they emerged after more than one decade of adoption. Reasons for a radical change in the process implementation are listed, together with the improvements that the envisaged changes will bring to the internal data flow – as part of the Secretariat's daily operations – and outside its boundaries (targeting mostly scientists, data analysts, policy makers, country-level focal points as well as national and regional management bodies).

The proposed changes aim at rationalizing the entire data management chain, all the way up from the data ingestion to the data dissemination steps, at the same time enabling data consumers to have a simpler and more effective way to get access to the data while still enforcing the confidentiality policies currently adopted by the Commission.

The most ambitious goal of this exercise is to increase the overall value of the data, transforming raw information into a valuable asset from the very first stages of the process, at the same time reducing the time-to-market prior to the final dissemination of regular information updates.

A description of the core and ancillary tools that the new data management processes will make available is given, detailing the impact that these tools will have on the Secretariat's staff daily operations as well as on the broader community that relies on the disseminated information. The strong interactions between the new data management processes and the revised data collection forms are also highlighted, stressing out the need for the revised forms to be adopted to the largest extent possible.

An example of the extended functionalities that the new integrated data management system will provide is also shown, demonstrating the added value that these functionalities could bring if made publicly available through the IOTC website.

Also, preliminary results indicate that the time required to produce the expected data sets for assessment of data-poor species is greatly reduced with respect to the current processes. For more complex species (Tropical or Temperate tunas), we expect the same gain in terms of efficiency once all the involved processes are successfully implemented and tested.

As one of the key tasks required by the successful implementation of the new data management processes is the harmonization of the current reference data sets (including fleet codes, gear codes and species codes) we will also propose *ad-interim* procedures that we expect will be adopted by the end-users to ensure a smooth transition between the two systems.

[1] Fabio Fiorellato (fabio.fiorellato@iotc.org) , James Geehan (james.geehan@iotc.org) and Lucia Pierre (lucia.pierre@iotc.org)

Contents

Introduction

The current data management processes in place at the Secretariat are the results of more than one decade of evolution and refinements, following frequent changes in recommendations, data management procedures as well as data submission and requirement policies.

The principle driving the implementation of the original processes was to enable users (the Data and Science section of the Secretariat) into timely and effectively respond to data updates, validate and finalize the received information and produce the expected outcomes, for dissemination and scientific purposes, in a simple and repeatable way.

Although great efforts were put into ensuring that the overall processes should behave and be managed as an *integrated system*, this approach was not fully pursued since the beginning (mostly for contingent reasons) thus resulting in a collation of tools and data storage mechanisms that – though fit for the purpose – turned out to be less resilient to changes and updates than expected.

As a result, the processes grew in complexity over time, requiring a well-established experience in order to be successfully mastered. Furthermore, due to the necessity of favouring quick response times over safe and controlled access to the data, some of the process components were designed as *stateful* systems, preventing – *de facto* – concurrent operations on different, non-overlapping subsets of the data, and therefore implicitly increasing response times to any change or update.

The need of disseminating the curated information through the IOTC website (e.g. the *Online Query Panel* and the Working Parties-specific datasets) added further complexity to a system that was beginning to strive to reach its goals, at the same time introducing unneeded complexity that impacted on its maintenance and extension.

Acknowledging the limit (and benefits) of the current systems, and considering the needs for a faster and more streamlined production of the datasets required for stock assessment purposes, the Secretariat took the chance to revise and start redesigning the entire data management chain.

This redesign process – very demanding in terms of efforts – is still ongoing, although it has recently reached a very advanced stage. The results achieved so far are encouraging, and the new workflow has proven to be a concrete step forward towards the achievement of the goals and objectives that the system was originally designed for.

The changes introduced with this process do also have very relevant (and beneficial) side effects: the new data management systems have indeed been designed - since their inception - in order to be remotely accessible: this means that all entitled users could log-in to the data management console from any location, as long as they have network connectivity. In principle, although with some limitations mostly due to screen sizes, the data management processes could also be accessed through a *smartphone* or a *tablet*, increasing the flexibility of a data management system that is crucial for the Secretariat. At the same time, the inherent *remote* nature of the new system might enable data consumers to link their statistical and analytical processes directly to the new data dissemination services, thus gaining direct access (with the due limitations) to the data currently available at the Secretariat.

Current state of the art

Data are reported to the Secretariat (from each CPCs) at different times of the year, usually at the end of June and at the end of December, or are collected from other sources (published data, national institutions etc.) depending on their availability.

Data is usually provided in multiple formats, most of the times as custom Excel files that need to be transformed into the target format (still an Excel file, as per the expected data submission formats) prior to be validated and ingested within the system.

The IOTC Data Section receives the updates, performs the due conversion and validates the contents, following up with the data submitter if any request for clarification is needed or if alternative sources provide contrasting information.

Once data are ready for ingestion, they are incorporated within the system by means of *ad-hoc* forms and procedures that do also perform sanity checks and multiple validation tasks. Only when all checks are successfully completed, the reported data are stored within the IOTC database.

Besides enriching the content of the IOTC database, finalized data can serve multiple additional purposes:

- It can be made available through the Online Query Panel (confidential records are not published if they don't come in aggregated form);
- It can be used to produce the different datasets required by stock assessment scientists prior to each Working Party (here, as well, confidential records are published only if they come in aggregated form);
- It can be used to respond to specific requests coming from scientists, CPCs or other national / regional institutions.

The production of the stock assessment datasets is the most complex of the tasks, as it requires multiple processing steps, including:

- disaggregation of nominal catch records that come aggregated by species or gear;
- reallocation of reported catches and efforts on standard areas and timeframes;
- cleanup and conversion of size-frequency records to the default length or weight units (by species);
- raising of reported catches to nominal catches (including size-frequency samples to scale down the results);
- production of catch at size / catch at age datasets.

From a technical point of view, all the above processes are implemented as a collection of different *ad-hoc* solutions including multiple Relational Databases (Microsoft SQL Server and MySQL Server), Access databases with embedded forms, macros and VBA[2] scripts, Excel templates for the production of summaries and the dissemination of dataset-specific records, R scripts for the production of charts, geo-spatial plots and reports.

All the components of this technology stack are linked together by means of shared data sources and files, that due to their nature require that users are connected to the IOTC Local Area Network, as the involved databases and files cannot be accessed from the outside.

In terms of the number of single components involved, we can currently identify:

- 3 different SQL Server databases on the same server instance (one containing the raw data plus the reference codelists and configuration tables, two containing the outcomes of different processes – mostly related to the generation of raised catches and size-frequencies)
- 1 MySQL Server database for the storage of data to be disseminated through the Online Query Panel;
- 3 Access databases for the ingestion / management / disaggregation / dissemination of Nominal Catch data;
- 2 Access databases for the ingestion / management of Catch-and-Effort data and the dissemination of reallocated data;
- 2 Access databases for the ingestion / management of Size-Frequency data and the production and dissemination of converted length and weight data;
- 4 Access databases for the production of raised catches;
- 1 Access database for the production of Catch-at-Age data;
- 1 Access database for the production of each of the stock assessment files (excluding stocks being assessed with data-poor methodologies);
- 9 different Excel templates whose content is feeding from a subset of the above databases, used to produce the final Excel files disseminated for each working party.

The IOTC database is responsible for the storage and management of all codelists (including higher level aggregations of species, fleets and gears) as well as for the storage of the raw data.

Submitted data (mostly coming from CPCs) are managed through *ad-hoc* Access databases that serve the main purpose of providing a graphical interface for the ingestion and validation of the provided data. Once these are finalized, they become part of the datasets stored by the main IOTC database: all information related to the data submitter (including contact details for the CPC focal point) is lost, and at this stage it's not possible to get back to the original data if not by storing them aside (as archived files) for later reference.

Keeping track of historical records for all of the main datasets is only possible by either archiving the status of the different databases involved in the process (mostly the Access database files above) or by backing up the content of the main relational databases.

[2] Visual Basic for Application, Microsoft

Given the complexity of the tasks to be accomplished and the long lifespan of the current processes (since their initial inception) the proliferation of components and sub-systems depicted above should not come as a surprise and is perfectly legitimate (yet open to improvements).

The current efforts into redesigning the data management systems while still maintaining the same level of functionality of the legacy one, have currently materialized in the following components:

- 3 different SQL Server databases on the same server instance (one containing the raw data plus the user credentials and roles, the reference codelists, the configuration tables and the original reported data files, one containing the outcome of the different processes including nominal catch disaggregation, catch-and-effort / size-frequencies distribution and catch-at-size and catch-at-age datasets, and one containing the history log of all the operation performed on the reported data);
- One web application exposing the REST ([Thomas]) remote services that serve as backbone for any operation available through the system;
- One web application exposing the User Interface to interact with the system.

Furthermore, beside the need of having a SQL Server database license, there is no other vendor lock-in and all the components are implemented with fully open source technologies (namely Java 8, Spring, Jersey, Angular JS, Bootstrap) all available as production-level components.

As anticipated, the major benefits implied by this approach are:

- Limited number of components (reduces maintenance and evolution costs);
- Data processes are stateless (multiple users can perform the same process at the same time without interfering with each other and producing repeatable, deterministic results);
- The system is inherently accessible from anywhere (does not require being physically within the IOTC Local Area Network);
- Fully-fledged access control at record-level (users do generally have limited capabilities: they can perform only the operations they're entitled to);
- Inclusion of geospatial features as natively provided by the RDBMS[3];
- Data extraction could be made available to external users as well (depending on user roles and capabilities);
- Data consumers could use the REST services to get access to the data they need (see above). This means, in principle, that consumers should incorporate REST calls e.g. in their R scripts to get live data from the system instead of having to download the datasets once they're made publicly available;
- Extended features could be plugged-in within the IOTC website to complement what is currently available through the Online Query Panel;
- Dissemination of curated data sets through the IOTC website is greatly simplified.

Major changes

A preliminary step for the implementation of the new data management processes consisted in a complete redesign of the existing main database.

The data structure was almost completely *normalized* in order to reduce redundancy to a minimum. The definition and content of many core codelists was revised to be more consistent with the actual data: in particular, some gear and species aggregations have changed not only in terms of adopted codes (aggregates do now have a symbolic code starting with '`AGxx`') but also in terms of the entries that are part of such aggregations.

Furthermore, the gear definition does now incorporate the concept of 'school type' within the gear itself (*Purse Seines* are now split between *Purse Seines – Free School* and *Purse Seines – Log School*).

Area codes (fishing grounds) have been hugely revised as well: in addition, thanks to the native geospatial extension provided by the database engine, it is now possible to incorporate the proper geographical boundaries for each area, including most of the irregular ones. As a consequence, all operations on the data that require the availability of proper fishing ground geometries to be performed, can now take advantage of the geospatial features of the new database.

[3] Relational DataBase Management System

The notion of '*alternative effort*' has also been introduced: in the legacy database, multiple effort records with different effort units could refer to the same strata, and a prioritization of effort codes was necessary in order to tell which of the different effort units for the same strata should be considered as *primary*.

This was particularly evident when producing the reallocated catch-and-effort dataset, as efforts not referring to the main effort unit (by gear) were simply discarded. Now, with the addition of a complementary effort unit, we increased the number of effort records that could be managed by the database (as long as they refer either to the primary or to the alternative effort for a given gear) and this introduced changes in effort values stored within the disseminated datasets, enhancing their accuracy.

Data aliasing

To enable a proper transition from the *old* (legacy) to the *new* coding systems, IOTC will update and disseminate the transposition mappings between current and new codelists. These transpositions are not always 1:1 and under specific circumstances they might introduce *aliasing* that will prevent exact backward conversions between datasets.

As an example, the old '*Species*' codelist was listing these three distinct codes:

- **BAR** - *Barracudas*
- **BVV** - *Yellowmouth barracuda*
- **YRS** - *European Baracuda*

Now, under the new '*Species*' codelist, these three distinct types of barracudas are all equivalent to the aggregate:

- **AG32** – *Barracudas*

This means that it will not always possible to match records when comparing a dataset produced by the current process with the same dataset produced by the new system.

The same also happens with a few gear codes, namely:

- **HOOK** - *Hook and line*
- **HATR** - *Hand line and Troll line*

These are now, under the new '*Gear*' codelist, all equivalent to the aggregate:

- **AG08** – *Handline and Trolling*

Other relevant changes related to the '*Gears*' codelist include renaming the *Baitboat*-related gears to their *Pole-and-line* equivalent, namely:

Legacy database	New database
BB – ***Baitboat***	**PL** – ***Pole and line***
BBM - ***Baitboat mechanized***	**PLME** - ***Pole and line (mechanized boats)***
BBN - ***Baitboat non-mechanized***	**PLNM** - ***Pole and line (non-mechanized boats)***
BBPS - ***Baitboat and purse seine***	**AG01** - ***Baitboat and purse seine***

Table 1. Updates to gear codes and categorizations

Features overview

Here we will present a brief list of the features as currently available within the new integrated IOTC statistical data management system.

The data management console is accessible as a remote web application and as such it just requires a web browser (we recommend using recent, W3C[4] compliant browsers such as Google Chrome and Mozilla Firefox) and a working, DSL-grade Internet connection. A screen resolution of at least 1600x900 pixels is also recommended for a better browsing experience.

Dashboard

The entry point of the system is the *Tasks dashboard*, listing the available tasks grouped by category.

It is complemented by a chart showing the size over time (in terms of number of records available per year) for each dataset, and can display the data as different types of charts filtered by dataset and time interval.

The bottom-left panel of the Tasks dashboard displays the overall details for each dataset, reporting – for each of them – the total number of records currently stored in the database, the number of records marked as *final* or *confidential* and the number of records with remarks set, plus the date of last update of each dataset.

Users can have a quick glance at the current status of each dataset and directly access its content by clicking on the dataset label within the 'Current dataset' section.

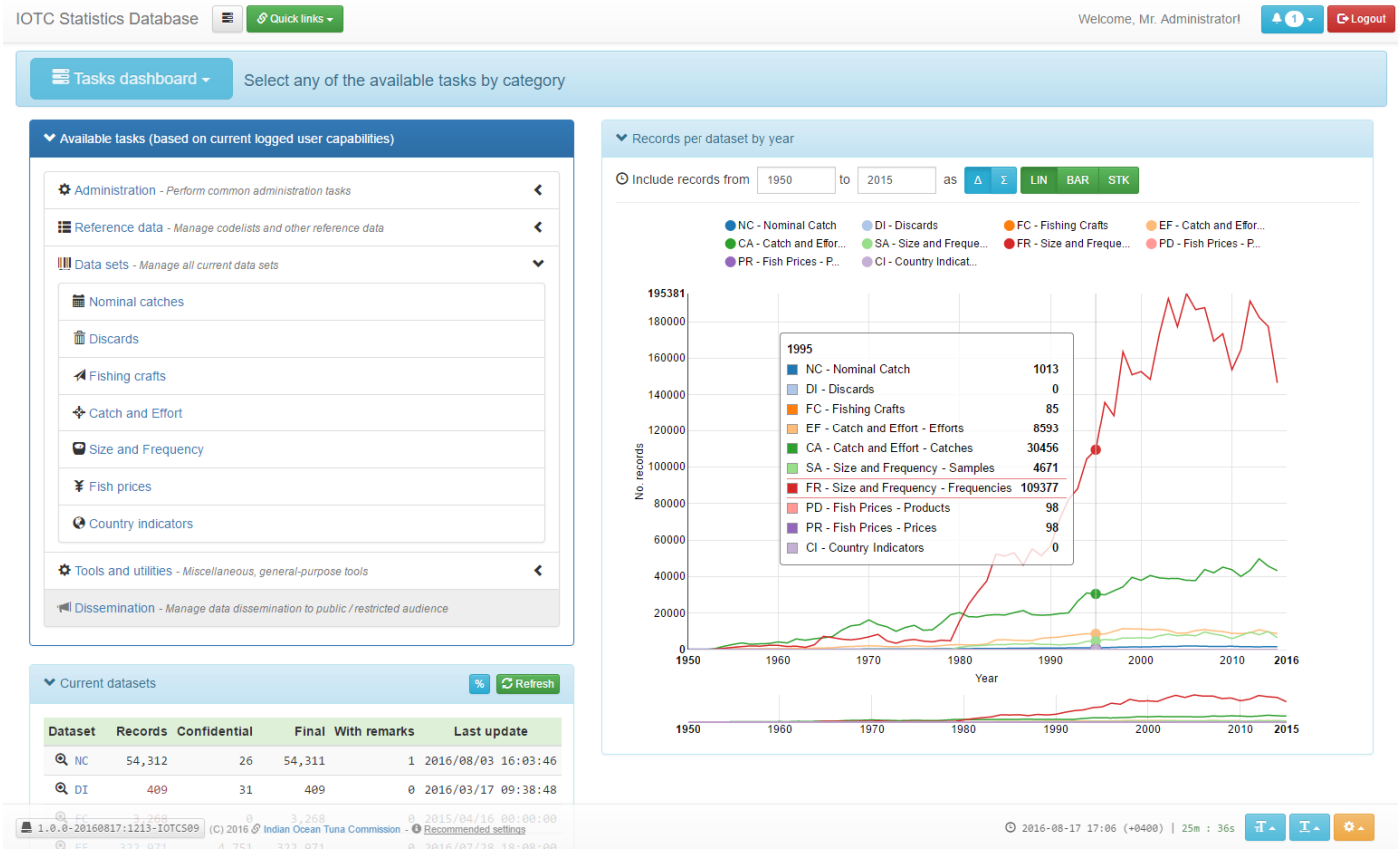

Figure 1. The *Tasks dashboard*

The overall list of currently available tasks is the following:

- **Administration**
  - **Users** – list / filter / create / update / enable / disable users, assign API keys
  - **User logins** – list / filter / delete recorded user logins
  - **User actions** – list / filter / delete recorded user actions
  - **User roles** – list / create / update / delete user roles
  - **User capabilities** – list / create / update / delete user capabilities
  - **User sessions** – list / terminate user sessions

[4] World-Wide-Web Consortium

- **Reference data**
  - **Codelists** – list available codelists, create / update / delete codelists entries
  - **Other reference data** – list other available reference data (aggregations, mappings etc.), create / update / delete other reference data entries
- **Data sets**
  - **Nominal catch** – manage nominal catch records / produce nominal catch summaries and reports / upload nominal catch data / export nominal catch records / disaggregate nominal catches
  - **Discards –** manage discard records / upload discards data
  - **Fishing crafts –** manage fishing crafts records / upload fishing crafts data
  - **Catch and Effort –** manage catch and effort records / upload catch and effort data / export catch and effort records summary / reallocate catch and effort records / display reported catch – effort – CPUE data on an interactive map / produce catch – effort and CPUE data reports
  - **Size / Frequency –** manage size and frequency records / upload size and frequency data / convert and distribute size and frequency records / display reported sample distributions on an interactive map / produce size / frequency summary plots
  - **Fish prices –** manage fish prices records / upload fish prices data
  - **Country Indicators –** manage country indicator records / upload fish prices data
- **Tools and utilities**
  - **Geospatial tools** – browse fishing ground geometries and filter fishing grounds by their current Indian Ocean area size

The list of available tasks will differ, based on currently logged user grants (see Figures 2a and 2b).

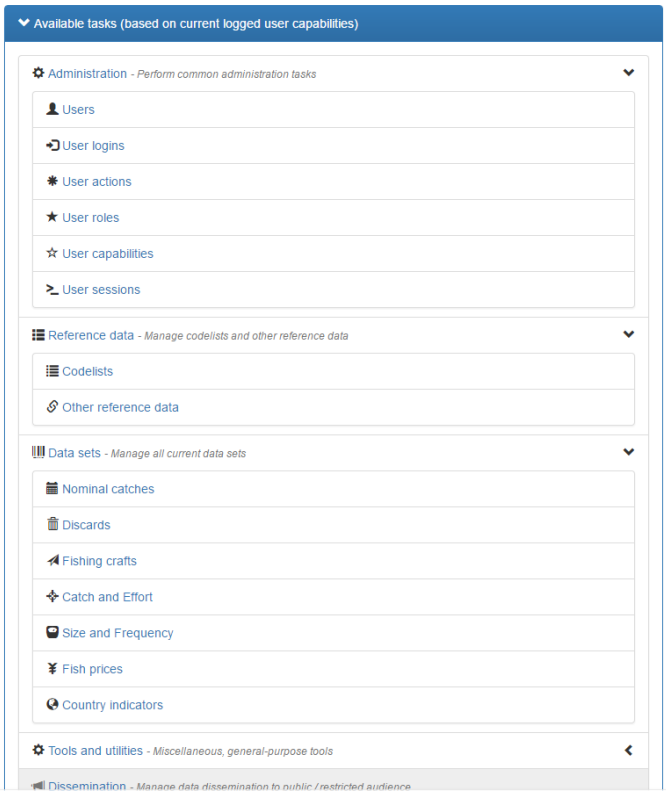

Figure 2a. Available tasks for an Administrative user

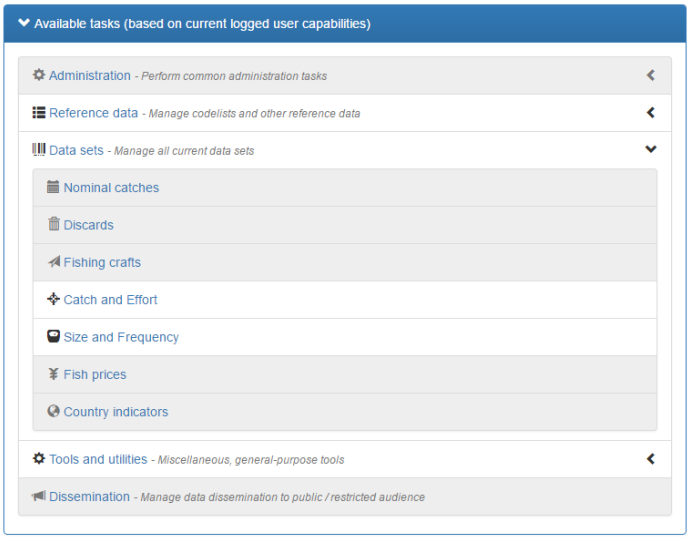

Figure 2b. Available tasks for another, limited user

User management

This section is accessible only to users with proper administrative grants. It provides all the required features to create new users, roles and capabilities as well as update user details and their grants / capabilities. Additionally, it gives administrative users the option to disable / enable an existing user as well as assigning / revoke API keys for non-interactive access to a subset of the remote data services provided by the application.

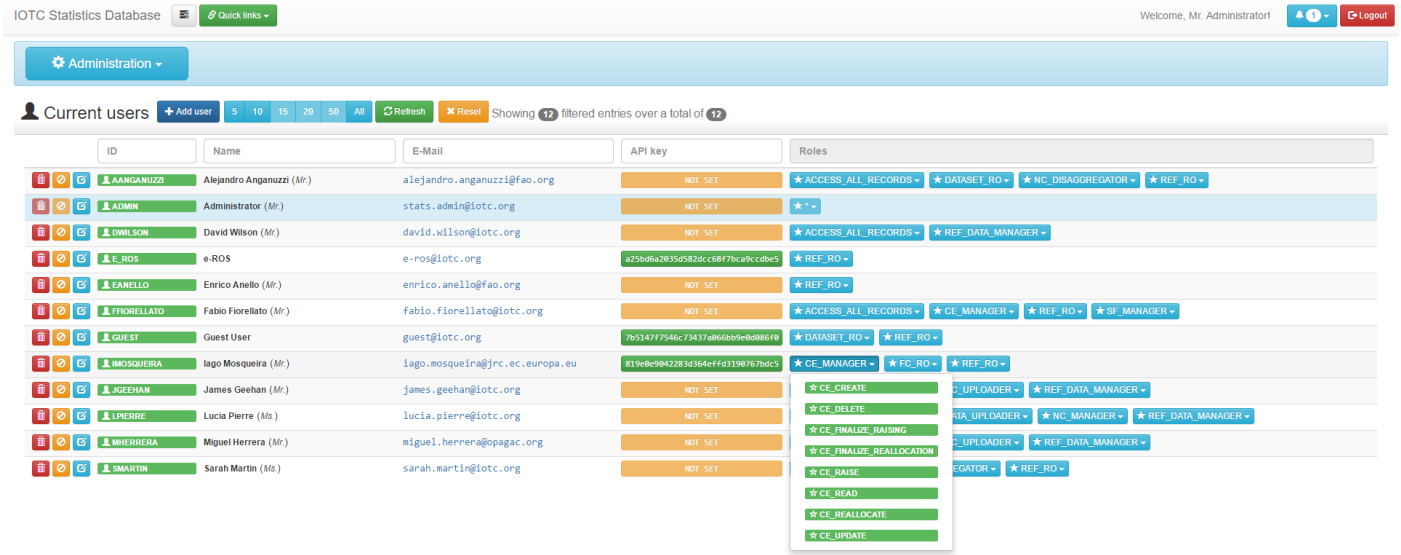

Figure 3. The current users' list

The tracing of user logins (*who* accessed the system, at *what time* and from *which location*) and user actions (*who* performed an action on the system, at *what time* and from *which location* – see Figure 4) does further increase the level of security and ensures that any non-legitimate access to the application could be easily identified, analyzed and accounted.

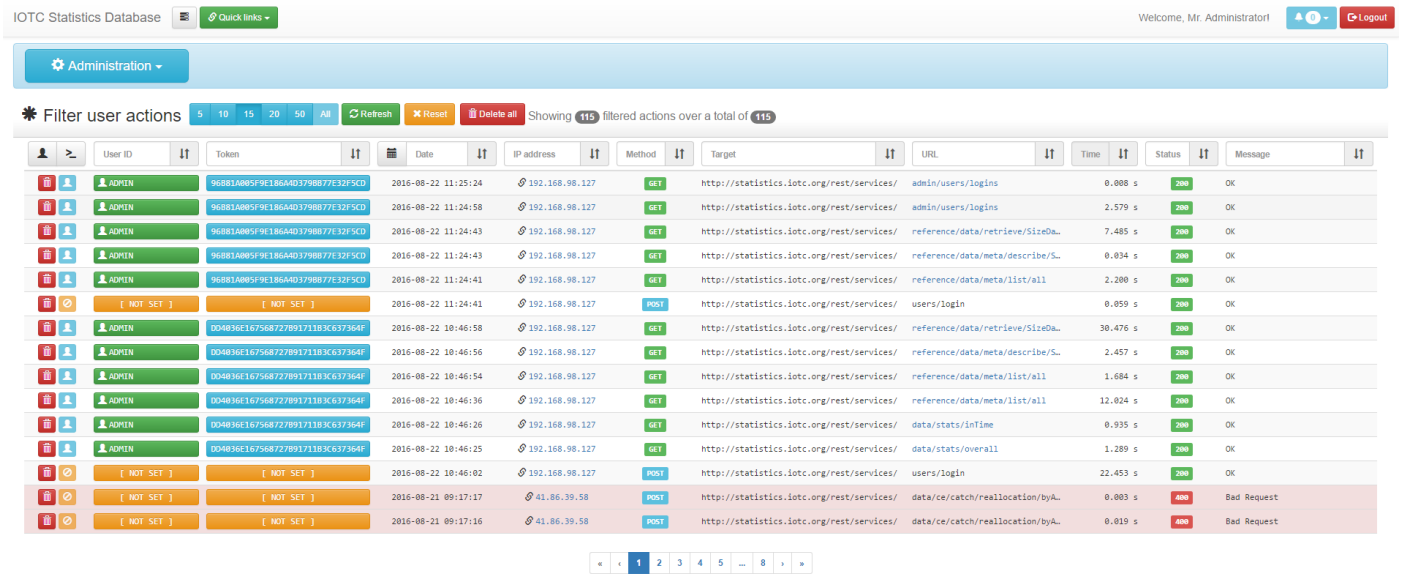

Figure 4. Listing recorded user actions

Codelists and other reference data management

Users with access to this section can perform changes to existing codelists or other types of reference data (including mappings, aggregations and configuration tables). These operations have an impact across the entire system, as changes to either codelists or reference data will indirectly introduce changes to the data dissemination and management processes (e.g. changing a higher level species classification will include / exclude the species from appearing within the disseminated data).

Codelists and *reference data* are different types of entities: the first are usually referenced by all datasets, whereas the second are mostly used to relate codelists among themselves and provide mappings (for backward compatibilities with the legacy codes) and configuration parameters to the data management processes.

Examples of codelists are:

- The SPECIES codelist (lists species by codes, scientific name, English name, French name etc.);
- The FLEETS codelist (lists fleets by codes, country code, reporting country code etc.);
- The FISHERIES codelist (lists fisheries by codes, fishery type etc.)

Examples of other reference data are instead:

- The SPECIES_AGGREGATIONS reference data (listing species – by code – that are part of a given aggregation);
- The FISHERIES_AGGREGATIONS reference data (listing fisheries – by code – that are part of a given aggregation);
- The SPECIES_MAPPING reference data (listing the mapping between legacy species codes and new species codes);
- The SIZE_DATA_LIMITS reference data (listing the minimum / maximum valid measurements – by type – for any relevant species)

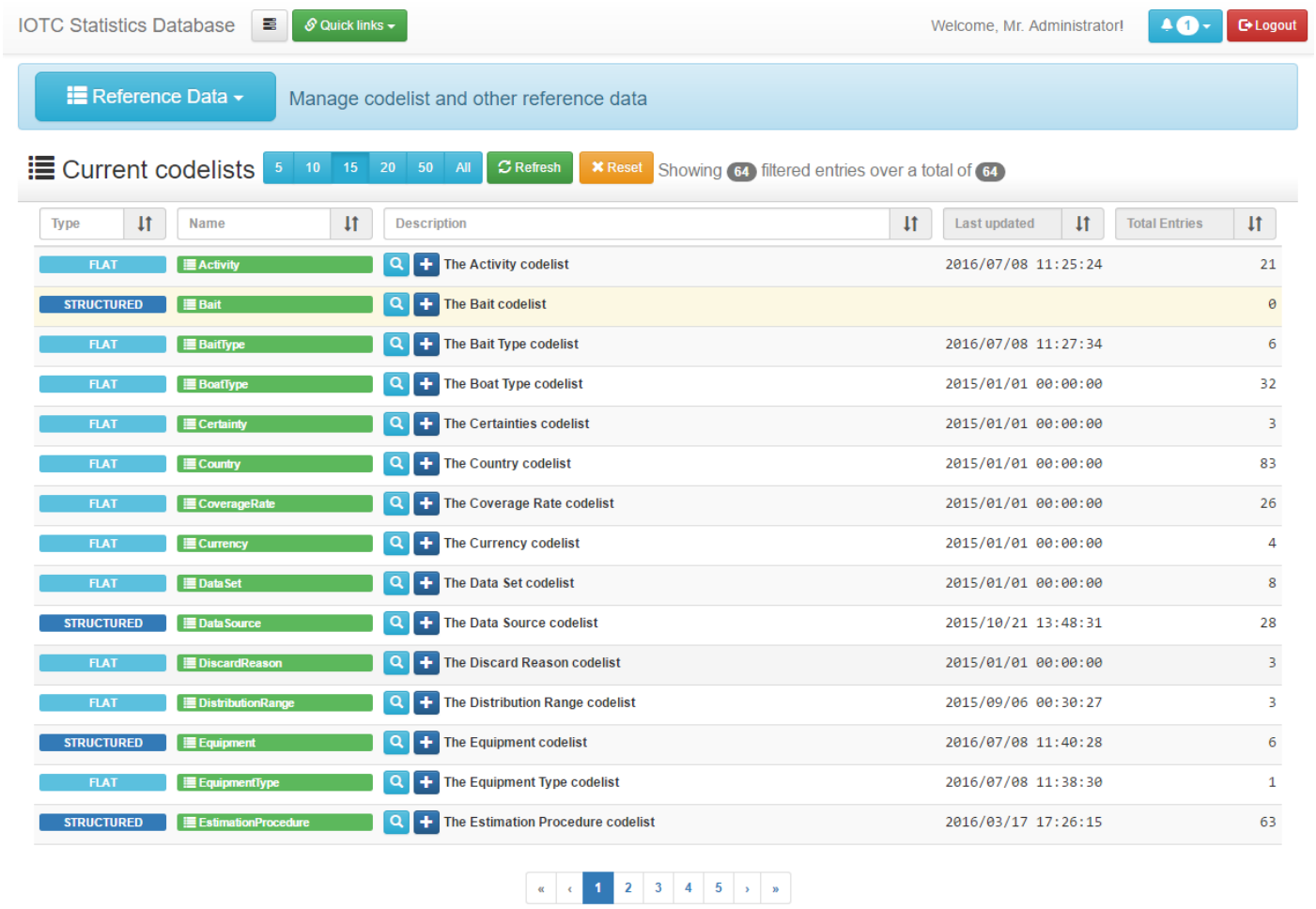

Figure 5. Listing the currently available codelists

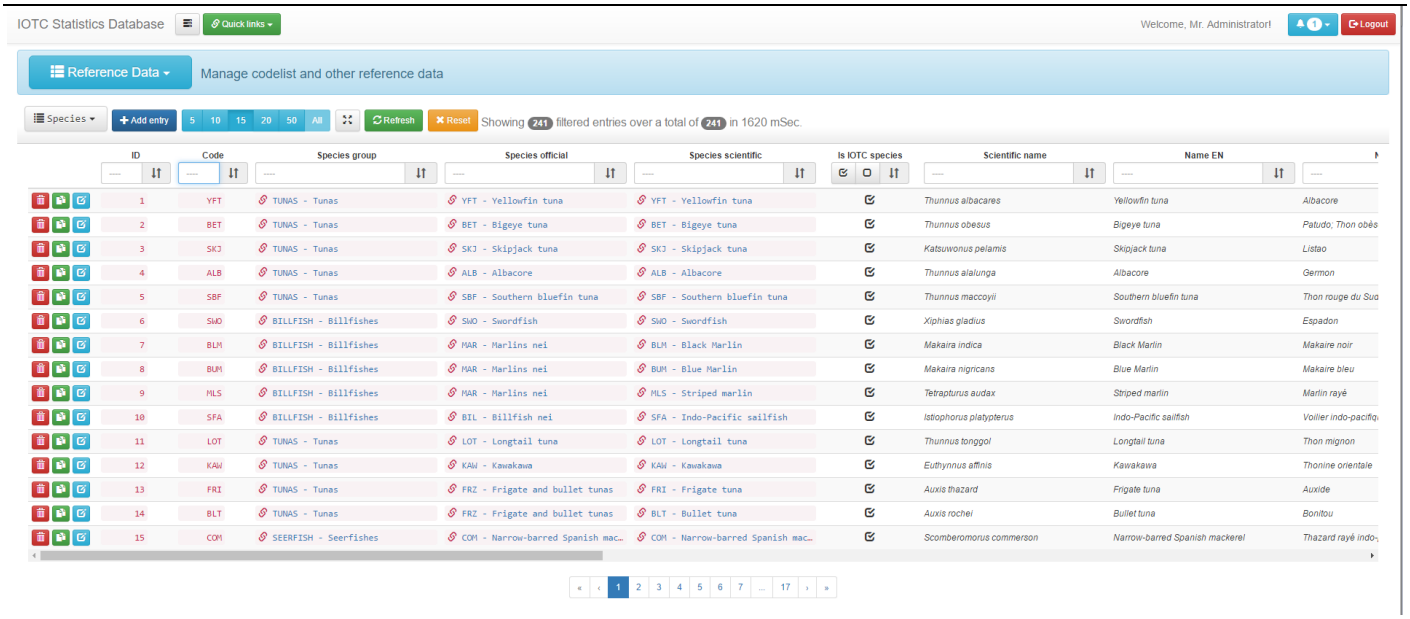

Figure 6. Displaying the content of a specific codelist (SPECIES)

The available tasks for codelist and other reference data management include changing (updating / deleting) existing entries as well as creating new ones.

During the creation or editing of an entry, the system will support users into ensuring that all mandatory fields are properly set, and will provide '*search as you type*' facilities to identify related reference data or codelists (see the 'Species official' field in Figure 7) that the entry being edited might depend from.

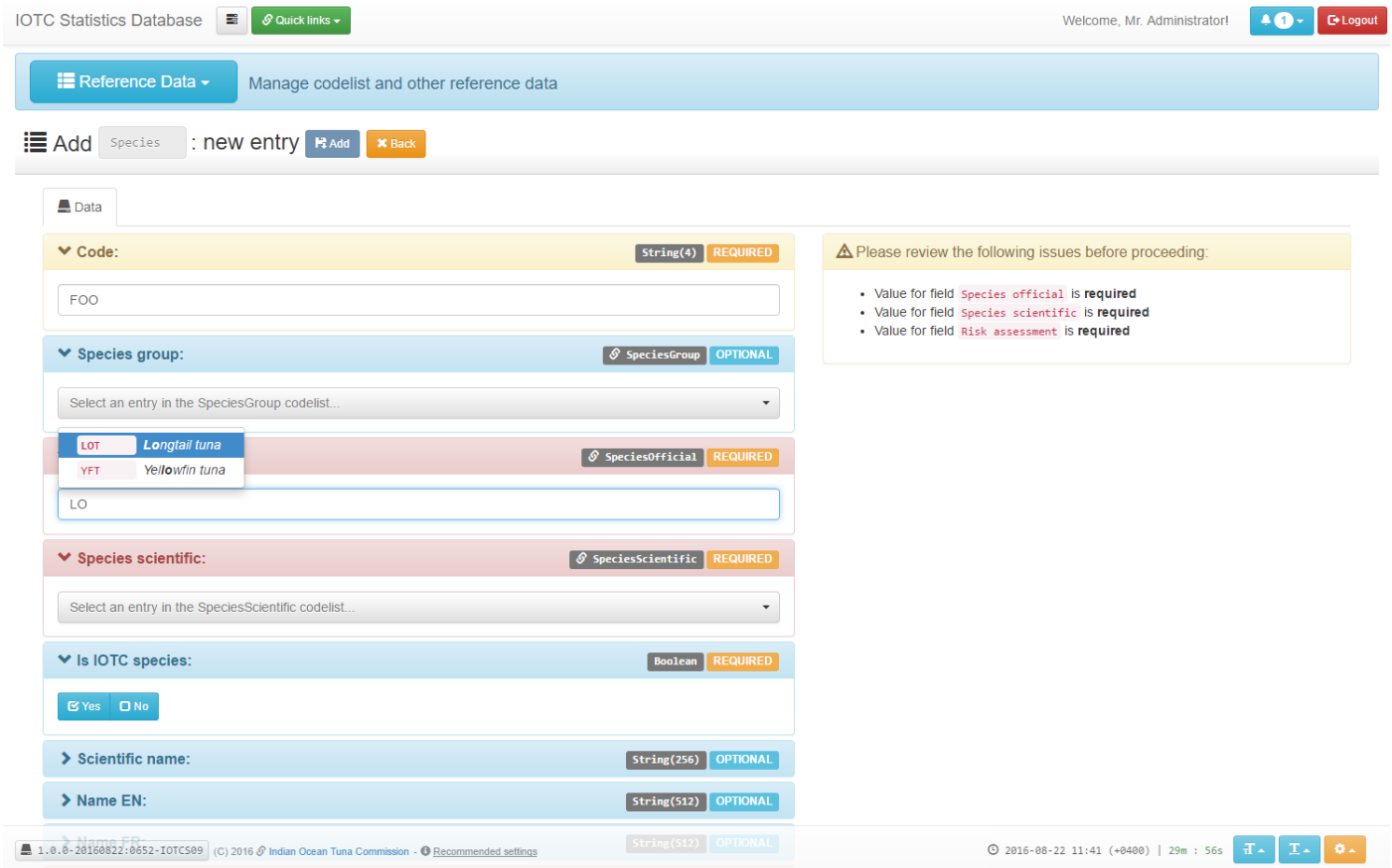

Figure 7. Editing a codelist entry (SPECIES)

Datasets management

The *dataset management tasks* are common to all datasets currently managed by the system and include operations and functionalities as:

- *Filtering* a dataset content
- *Creating* / *Editing* / *Deleting* a dataset entry
- *Uploading* new data (by using one of the forms currently available for download through the IOTC website)

Additional operations could also be available to entitled users, depending on the dataset type and on current user grants.

The new interfaces used to filter the content of a given dataset (see Figure 8) and edit any of its record (see Figure 9) are consistent across all datasets: in particular, when editing a dataset entry, the system provides support to ensure that all mandatory fields are properly set, and will display searchable *dropdown lists* to simplify the selection or update of any referenced codelist (as required by the specific dataset being edited).

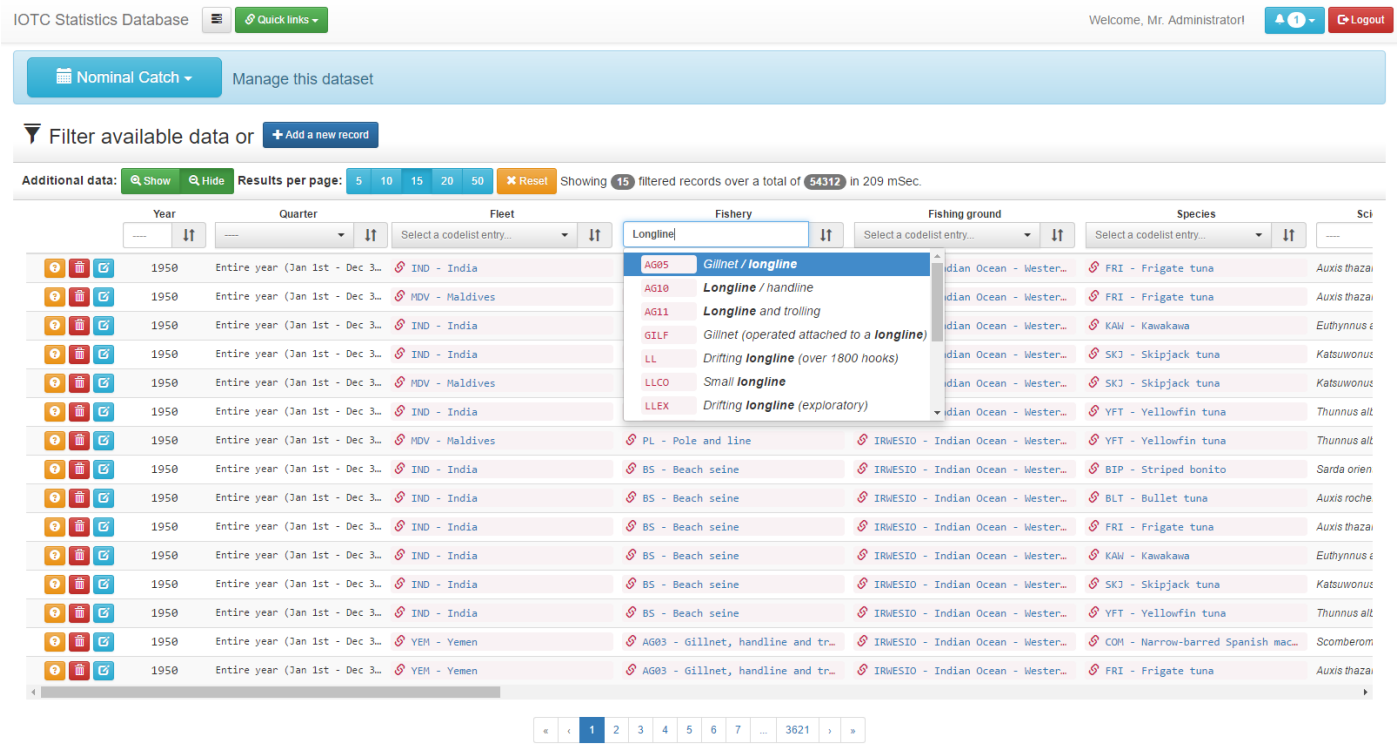

Figure 8. Filtering the content of a dataset (Nominal Catch)

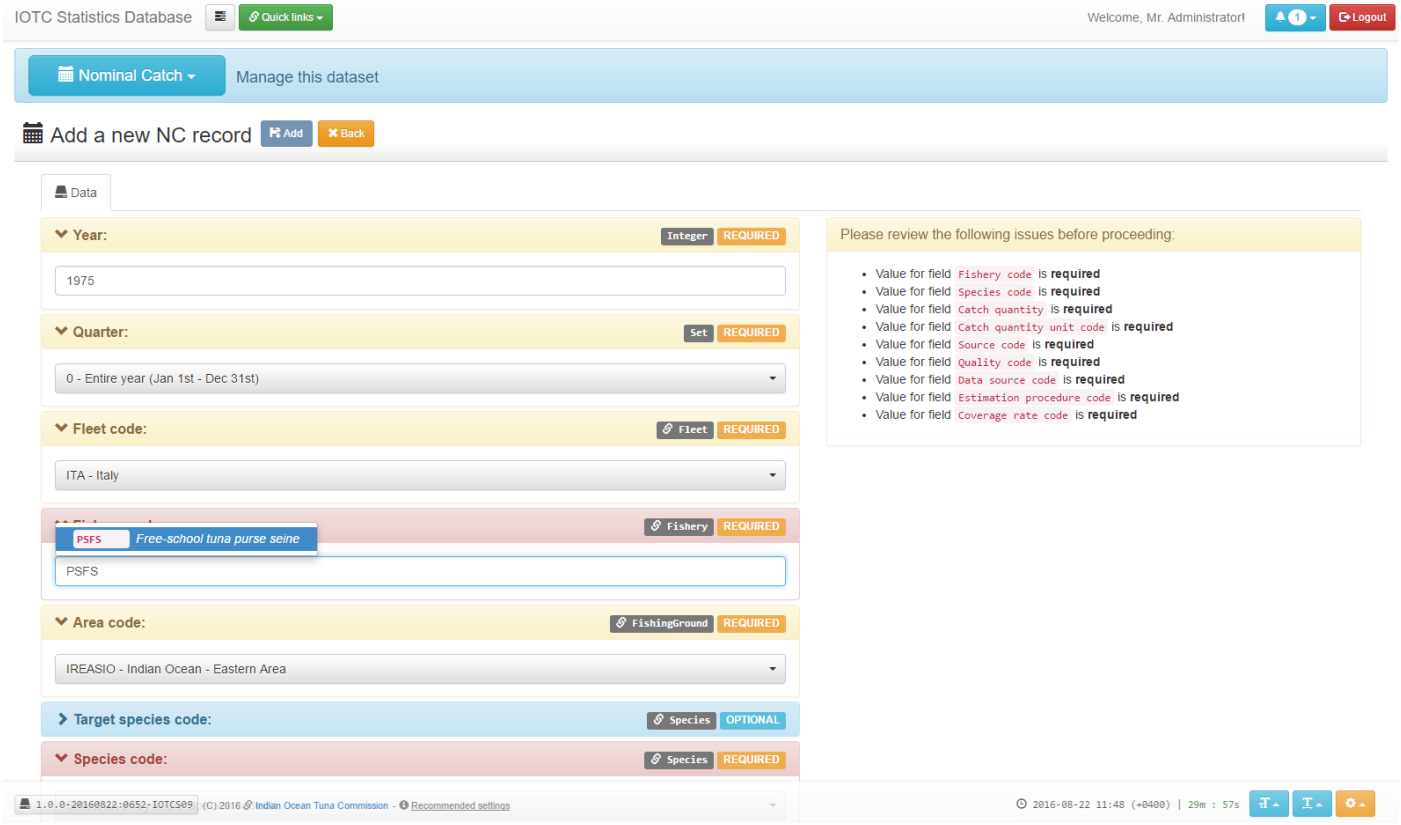

Figure 9. Manually editing a dataset entry (Nominal Catch)

The bulk upload of new dataset records to the system is a process that is basically dataset independent. The only changes required, by dataset, are related to the adoption of a different data upload template among those currently available for download under the IOTC website[5] and by the type of checks applied to verify the integrity and completeness of the data.

As of today, it is not mandatory for CPCs to provide data to the Secretariat using these dataset-dependent, standard forms. Nevertheless, we strongly *encourage* the adoption of this standard for all the required mandatory data reporting tasks: adopting these forms will have the beneficial effect of ensuring that the sanity checks and error fixing procedures available within the system could be effectively used to perform a first quality assessment of the provided information as soon as data is received by the Secretariat and reduce – consequently – the time needed to successfully import new data in the system.

The preliminary step of the upload process consists in displaying the metadata available within the uploaded file (submitter name / organization / contact details etc.) and present a preview of the uploaded data, including the first ten rows of the form content (see Figure 10).

[5] http://www.iotc.org/data/requested-statistics-and-submission-forms

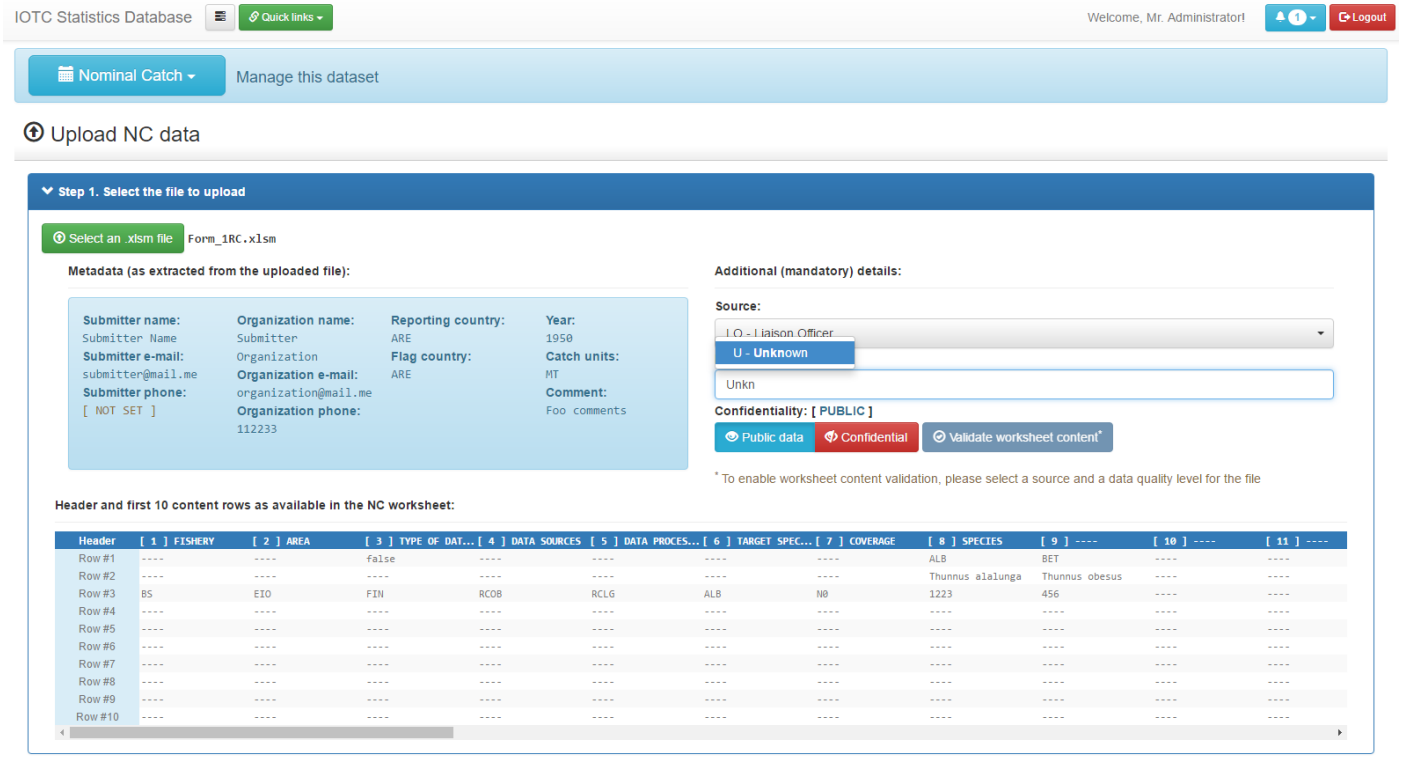

Figure 10. Displaying / editing metadata for an uploaded dataset (Nominal Catches)

Provided metadata are stored and linked to each corresponding record within the dataset (once finalized): this ensures that it will always be possible to trace back the origin of each and every record stored in the database and possibly contact the data provider in order to get further clarifications in case discrepancies are noted at a later date.

Once the required information is provided by the data-entry user, data are ready to be formally validated by applying preliminary checks verifying that:

- all mandatory fields are provided;
- there is no gap in the records;
- all codelist references are correctly set.

If any of the uploaded records is not triggering one of these rules, the upload process halts and presents the user with the list of identified issues, some of which might require updating the content of the involved codelists (e.g. to add a new species or a new fishing ground that is not yet part of the reference data) whereas other types of issues might require an update to the data stored in the file sent for upload.

As an example, the system may identify that within an uploaded form there are five fields that were supposed to be mandatory and that have not been provided: it's up to the data-entry user to decide whether these can be properly filled (e.g. by looking at past data sent by the same CPC) or they require getting back to the original data submitter and issue a request for clarification.

Once everything has been fixed and all validation rules are successfully triggered, the system will check whether the uploaded data are conflicting with existing records of the same dataset. This in order to prevent overwriting data that is already consolidated and furthermore to assess the trends of the data being updated: in some circumstances it is perfectly legitimate for a data submitter to provide updates to past records. For this reason, the system allows users to decide – on a colliding record by record basis – whether updates should be committed or not.

Nominal catches

Users entitled to perform additional tasks on the Nominal Catch dataset can access a growing set of features mostly meant to support the identification of possible inconsistencies within the data (prior or following the bulk upload of new nominal catch records) and produce ***reports showing trends and composition of current records*** within the dataset.

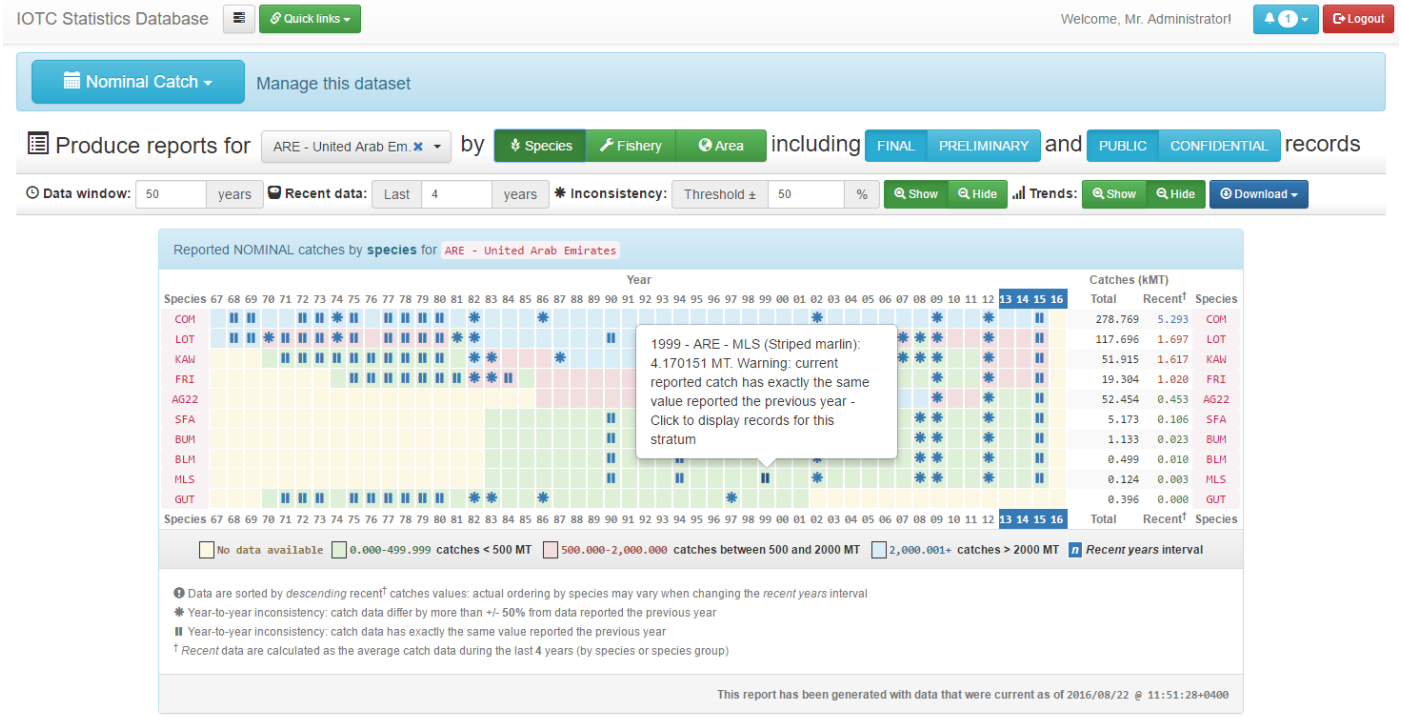

Figure 11. Nominal Catch reports by fleet and species

These type of reports can display catch amounts (for a given fleet) during a specific user-selectable timeframe (the last 50 years, by default). Cell colors vary with the range of catches recorded for each species / year, and the superimposed symbols show any relevant discrepancy (in terms of catch quantities) in data reported between two consecutive years as well as repeated patterns and trends.

Figure 11, as an example, shows that catches of Striped marlin for United Arab Emirates reported for 1999 have exactly the same value as for 1998 (~ 4.17 MT).

Similar reports by fleet can also be limited to specific gears or fishing zones (eastern or western Indian Ocean) and in all cases they can be exported as PNG[6] images (for inclusion in other reports) or as raw CSV[7] data for later analysis.

Users with the proper grants can also display ***summary reports for currently available Nominal Catch records***: this type of summary reports provide the same type of information as currently available in the Excel files disseminated – as part of the Nominal Catch dataset – prior to each working party.

With this specific feature, users can select a time range, a subset of the available fleets, gears, gear types (industrial / semi industrial / artisanal), catch areas and species, and have the corresponding reported Nominal Catches displayed as a line chart (see Figure 12a), a cumulative bar chart (see Figure 12b), a stream chart (see Figure 12c) or a table (see Figure 12d).

Data is grouped by species by default: nevertheless, users can choose to display nominal catches for a given set of filtering criteria by different grouping modes including areas (Eastern / Western Indian Ocean – see Figure 12e), fishery type (see Figure 12f), fishery group (see Figure 12g), fishery (see Figure 12h), fleet (see Figure 12i), species group (see Figure 12j) and record quality (see Figure 12k).

Grouping information by different criteria other than Species is available for all types of charts (not just bar charts as in the figures above).

[6] Portable Network Graphic, an open-source format for lossless image compression
[7] Comma-Separated Values, a textual format for the storage of tabular data

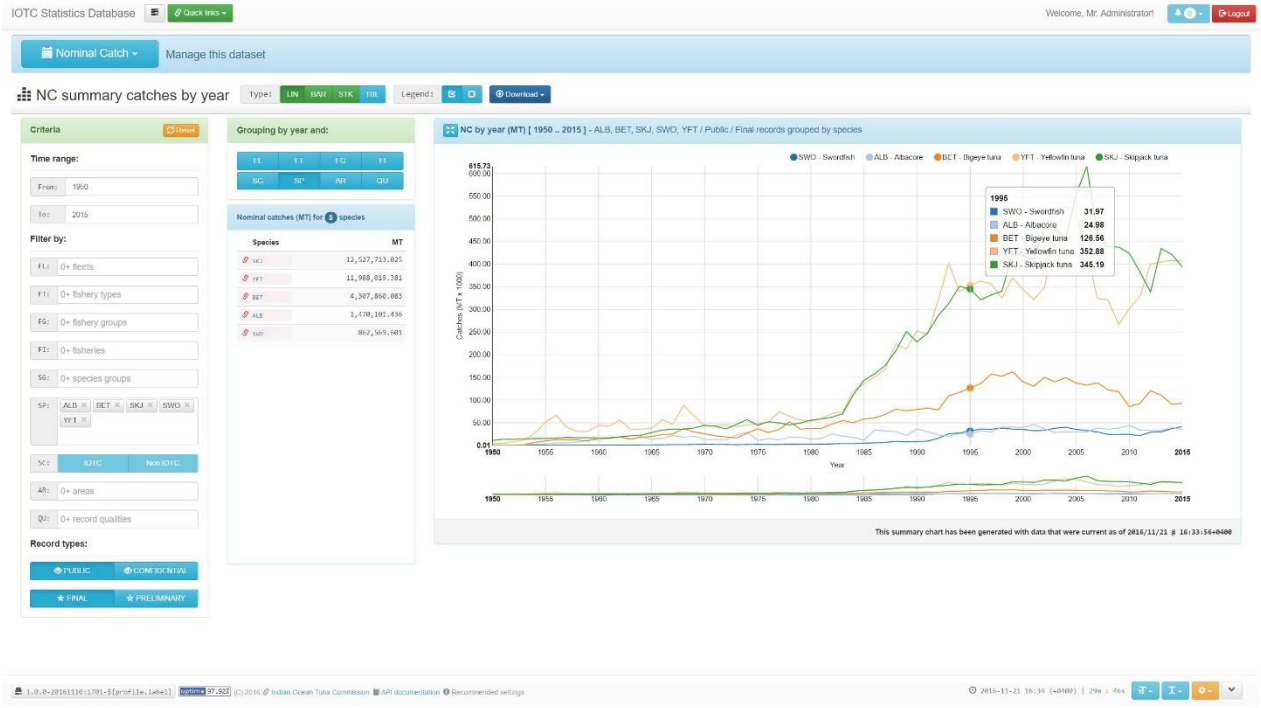

Figure 12a. Nominal Catch summary report as a line-chart

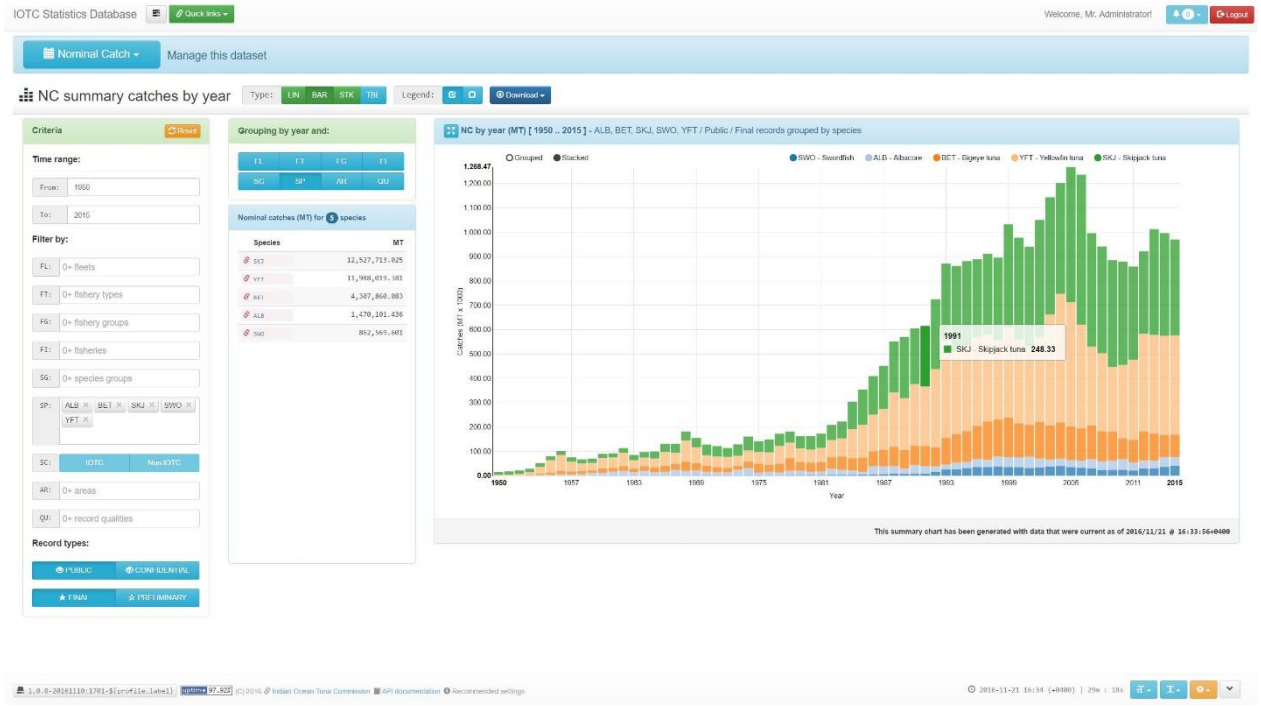

Figure 12b. Nominal Catch summary report as a stacked bar-chart

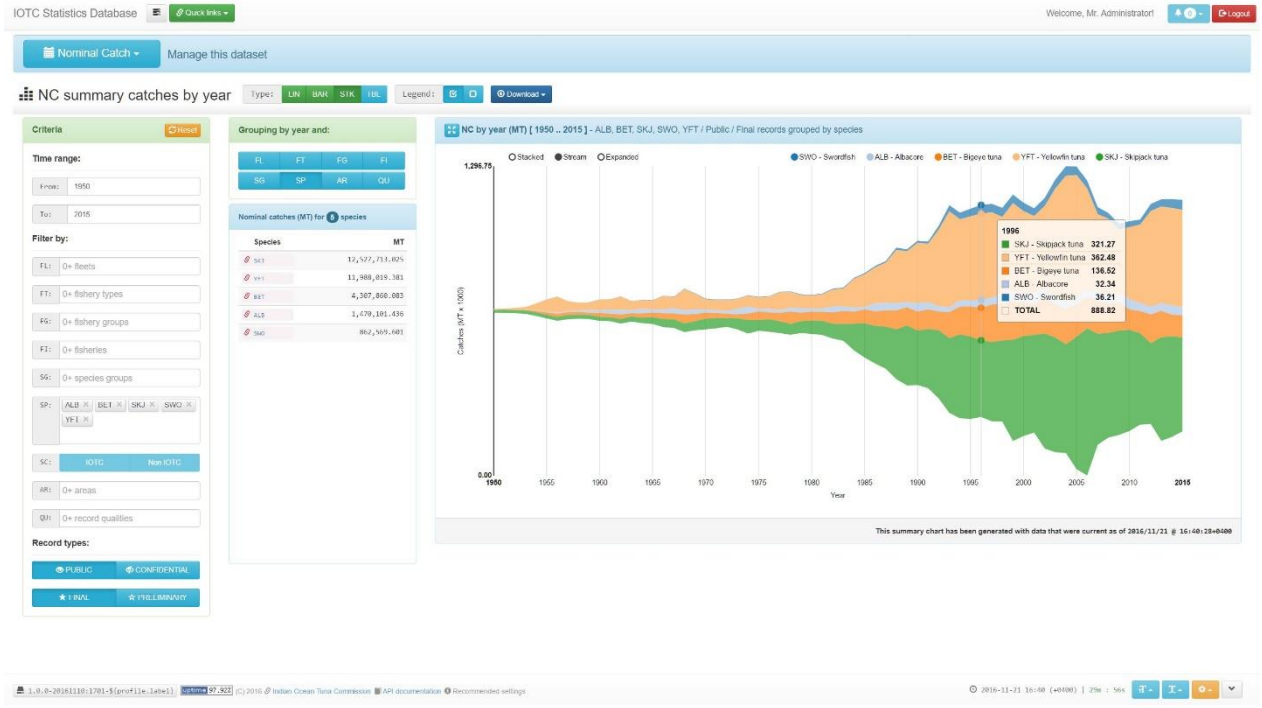

Figure 12c. Nominal Catch summary report as a stream-chart

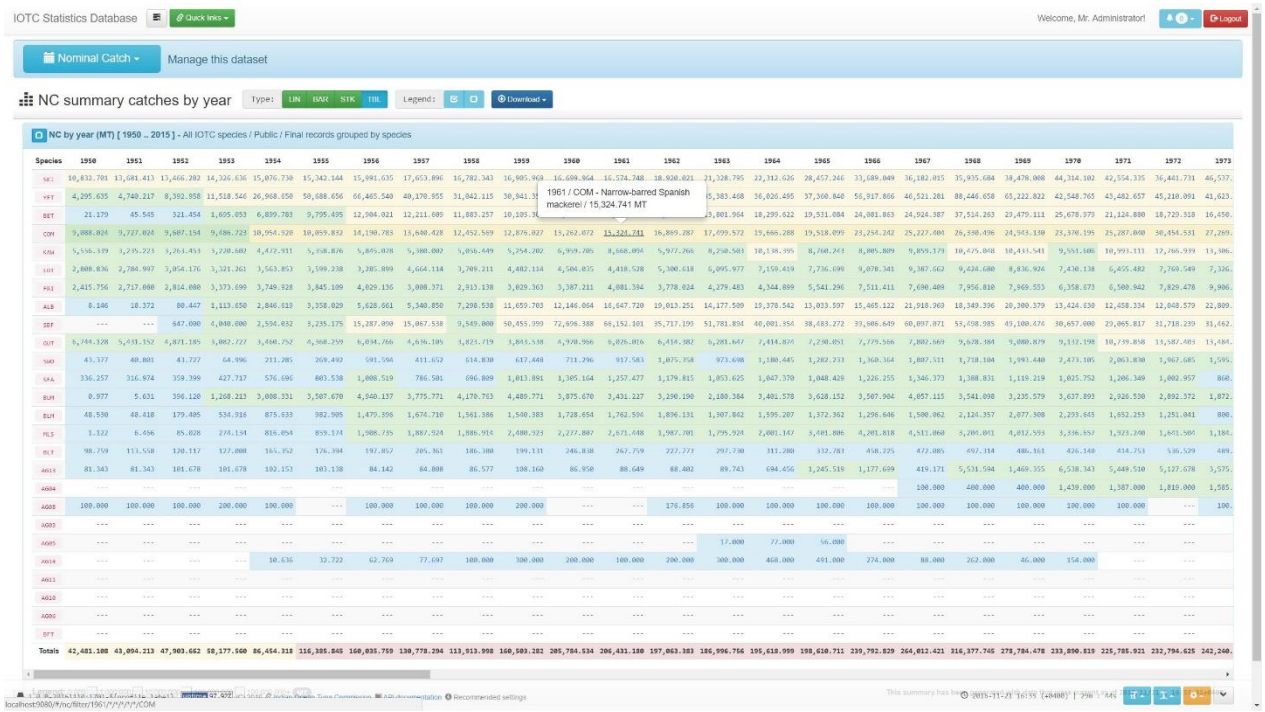

Figure 12d. Nominal Catch summary report in tabular format

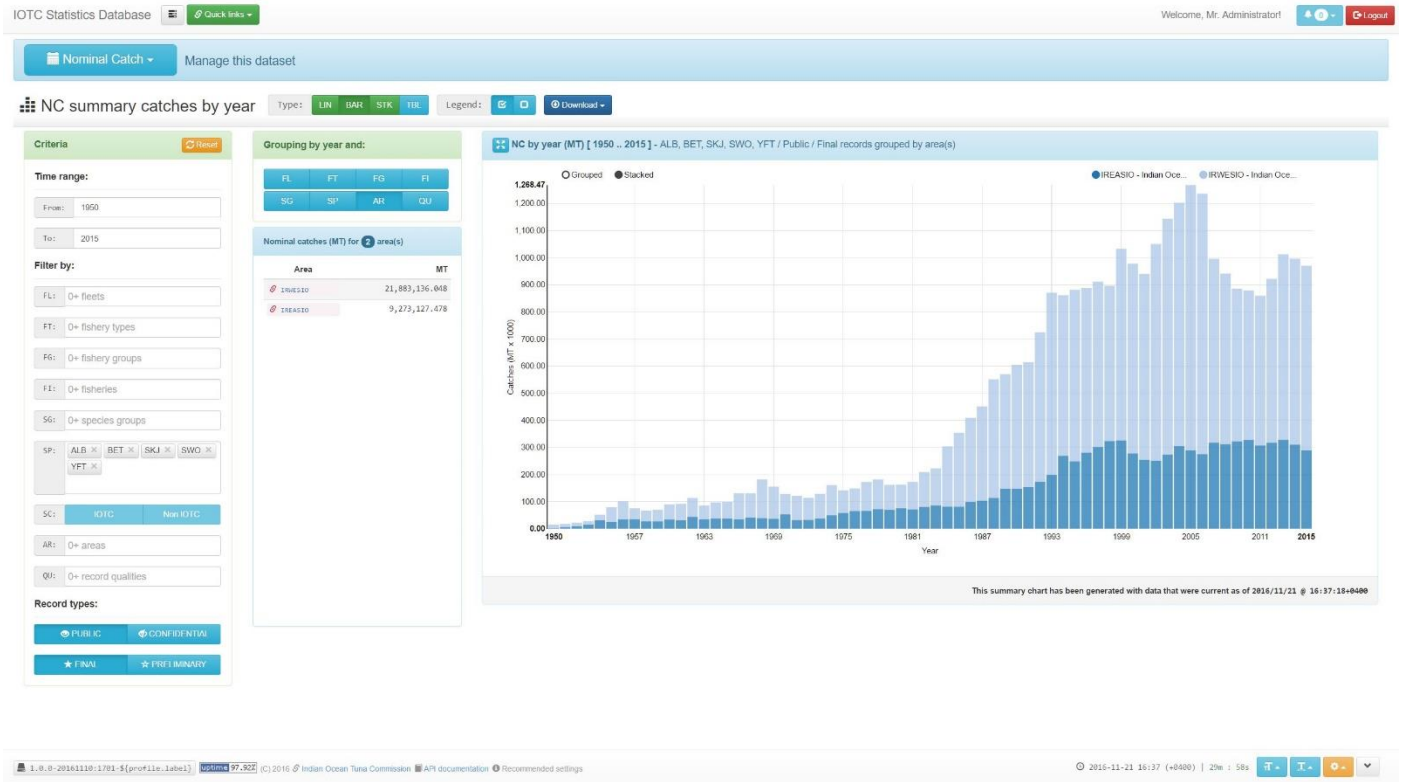

Figure 12e. Nominal Catch summary report as a stacked bar-chart with data grouped by area

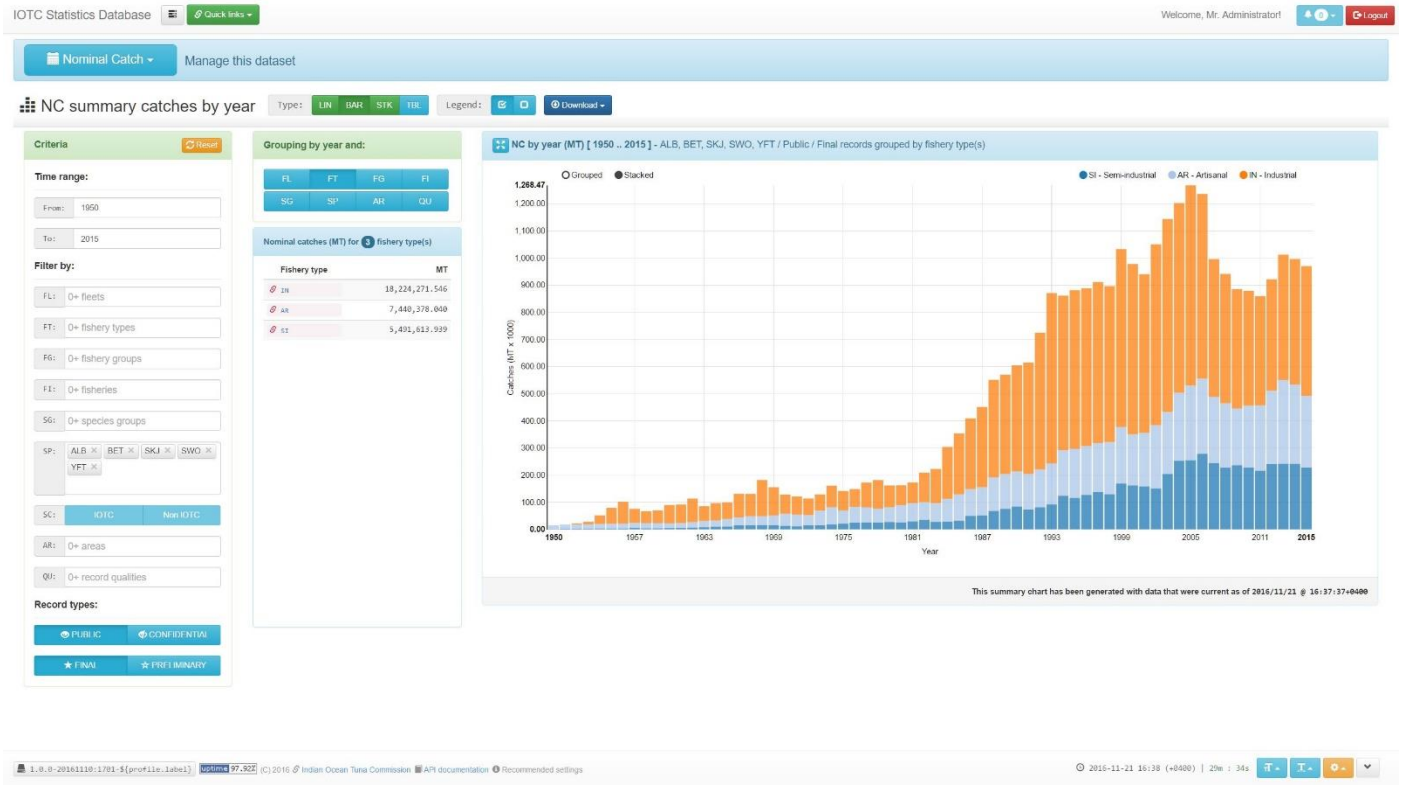

Figure 12f. Nominal Catch summary report as a stacked bar-chart with data grouped by fishery type

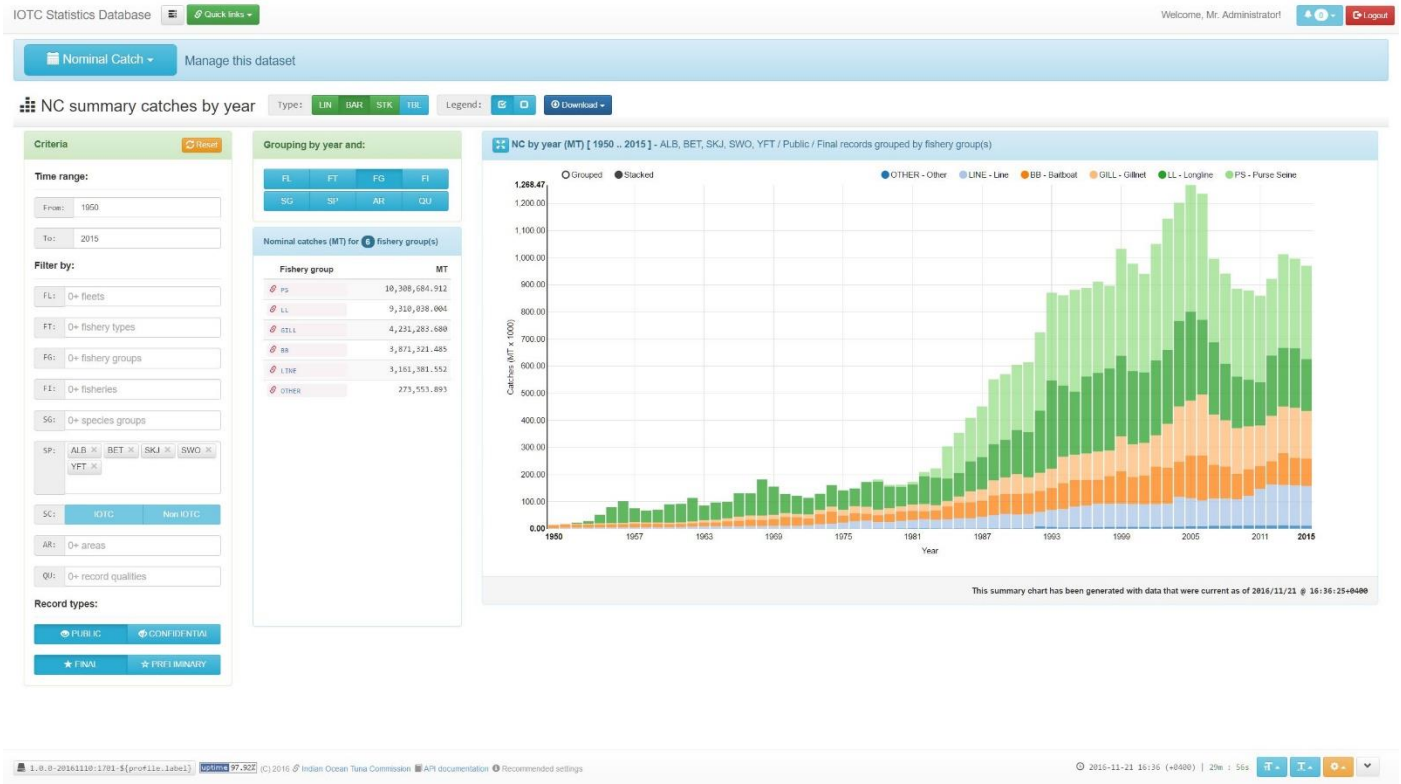

Figure 12g. Nominal Catch summary report as a stacked bar-chart with data grouped by fishery group

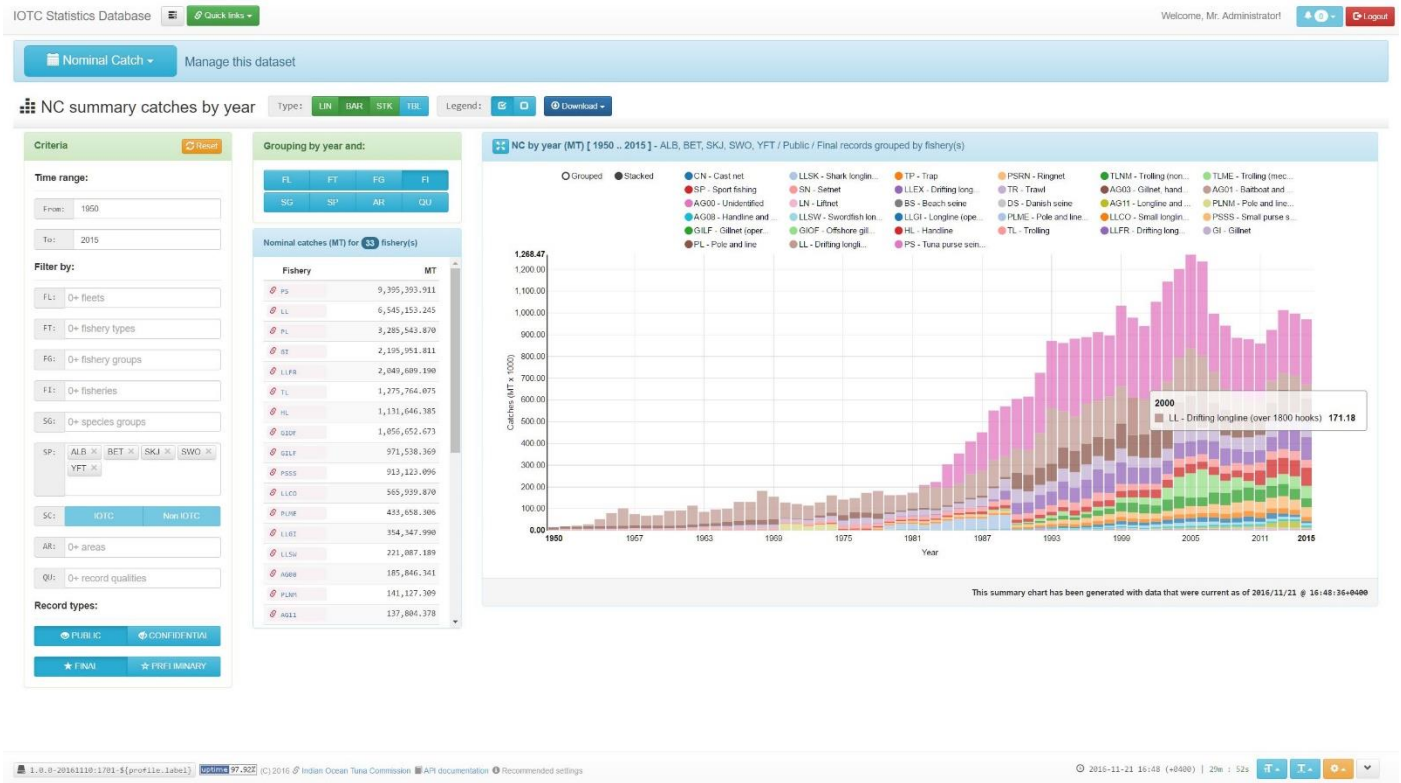

Figure 12h. Nominal Catch summary report as a stacked bar-chart with data grouped by fishery

Figure 12i. Nominal Catch summary report as a stacked bar-chart with data grouped by fleet

Figure 12j. Nominal Catch summary report as a stacked bar-chart with data grouped by species group

Figure 12k. Nominal Catch summary report as a stacked bar-chart with data grouped by record quality

The selected data subset can then be exported as a CSV file and charts can be downloaded as images for the inclusion in documents and reports. The format of the CSV will depend on the type of grouping applied to the original data.

The last Nominal Catch-specific feature currently available is the ***Nominal Catch Disaggregation***: this is the process of breaking down Nominal Catch records that refer to either a species or a gear aggregate into their component (by species or gear) and is the preliminary step for the preparation of the basic datasets required for the stock assessment of all working-parties' species.

It works by applying substitution rules – to all records involved – that look for other records by *proxy* fleet / gear / area of operation / time interval from which proportions by species or gear can be derived.

So far, within the legacy system, the disaggregation process is implemented as a specific Access database that links to the existing Nominal Catch dataset and applies the disaggregation procedures to produce a fully disaggregated dataset (i.e. whose records do all refer to single, non-aggregated gears or species).

A first improvement over the existing Nominal Catch Disaggregation process is that users are now able to select just a subset of the Nominal Catch records to be disaggregated, and apply the same disaggregation procedures to non-IOTC species as well.

Additionally, disaggregation results can now be stored within the database (as 'Disaggregation runs') and re-loaded at any time, to provide further insights into historical results.

Once the user has selected a subset of the Nominal Catch to disaggregate (or possibly the entire dataset for IOTC-species, as it is required most of the times), the system will process the records and – based on the current disaggregation procedures – provide as output the disaggregated results.

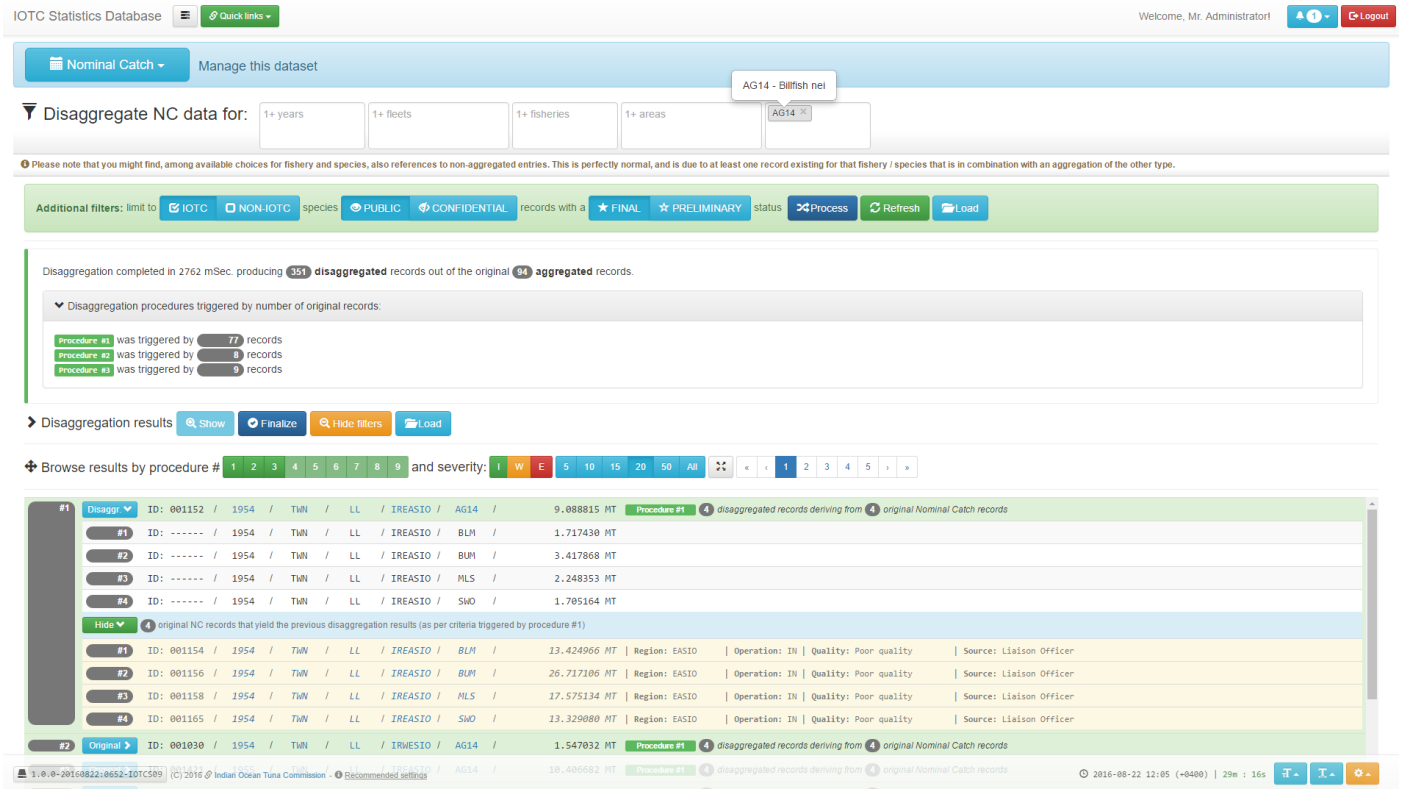

Figure 13. Disaggregation of 'AG14 - *Billfish nei*' Nominal Catch records

Under some circumstances, it might happen that a record cannot be automatically disaggregated by using different records from proxy fleets / gears / areas / time intervals: in this case, the user is supposed to provide his / her estimates about how to assign catches by single gear / species.

The reason why still some records might need to be disaggregated *manually* is that all of the current procedures are looking at different time intervals and proxy fleets / gears to identify records to be used for the disaggregation. Under some circumstances – and with the current procedure configurations – not all records can identify alternate records to be used for the purpose (data is particularly poor or sparse for the involved gears / species combination).

Figure 13 shows an example of the preliminary report produced by the disaggregation process. In particular, we see that of the original 94 '***AG14 – Billfish nei***' Nominal Catch records to be disaggregated the system was able to produce 351 distinct (non-aggregated) new records by applying 3 out of the 8 currently available disaggregation procedures.

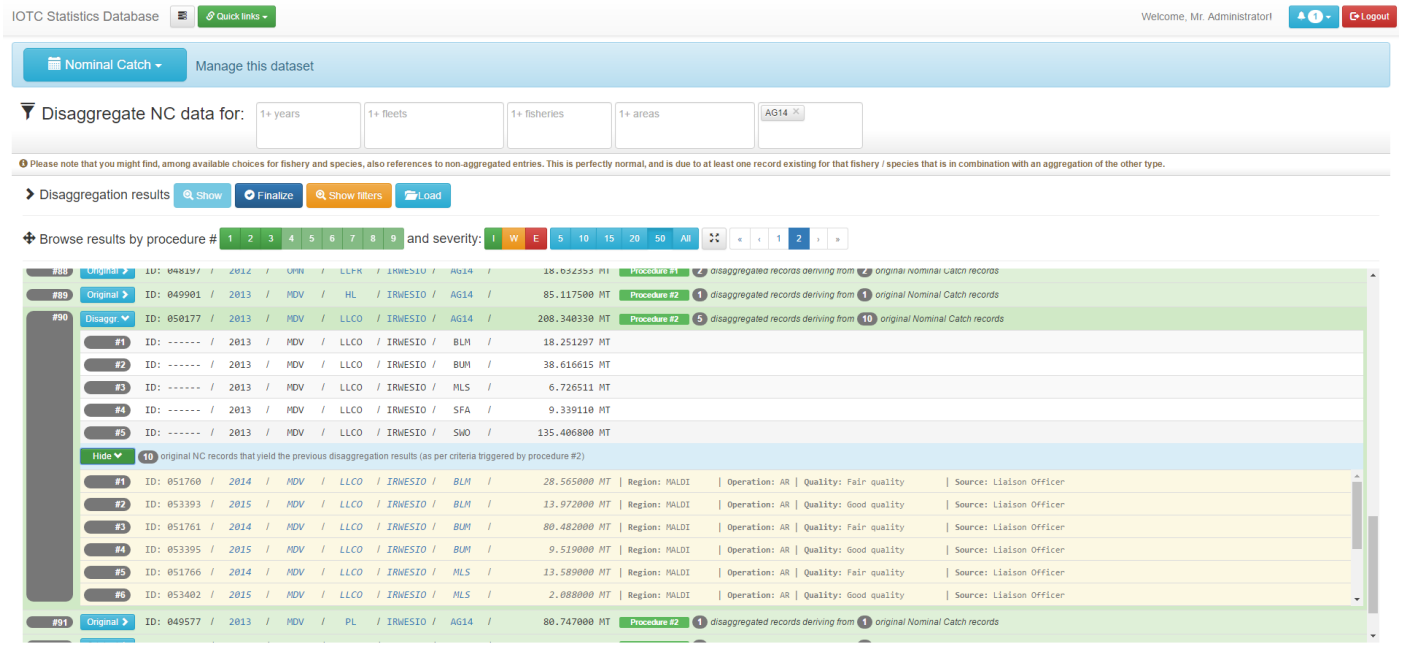

Figure 14. Example of Nominal Catch disaggregation results

Figure 14 shows a subset of the disaggregation outputs for results related to the above mentioned example (disaggregation of '*AG14 – Billfish nei*' catches only). In particular, we can see the five disaggregated records that will be used to replace the single aggregated record referring to **`2013 / MDV / LLCO / IRWESIO (Western Indian Ocean) / AG14 (Billfish nei)`**.

This record – that was originally referring to an aggregation of species - has been broken down to **five** non-aggregated records by using **ten** proxy Nominal Catch records identified by **procedure #5** (*Same fleet, same type of operation, same region, same IOTC area, up to 5 years before or after the fishing year*).

Thanks to these ten proxy records, the system can split the original catches (~ 208.34 MT) into five different, disaggregated catch records referring to the same strata and to one among **Black marlin**, **Blue marlin**, **Striped marlin**, **Indo-pacific sailfish** and **Swordfish** as species. The original catches are proportionally assigned based on how catches for the same species contribute to the total within the subset of proxy records identified by the triggered disaggregation procedure.

In case a record should be manually disaggregated, users will be presented with the original record and the required controls to add specific disaggregation values for any of the gear or species involved.

The disaggregation results can be finalized (that is, stored within the system as a *disaggregation run*) only when all the records requiring manual disaggregation have been correctly updated.

Disaggregation results can also be downloaded (either immediately, after they've been produced, or by recalling the content of a previous disaggregation run) as a CSV file that has the same format of the expected Nominal Catch disaggregated dataset produced prior to each working party.

For a formal definition of the disaggregation process, please refer to Appendix A2.

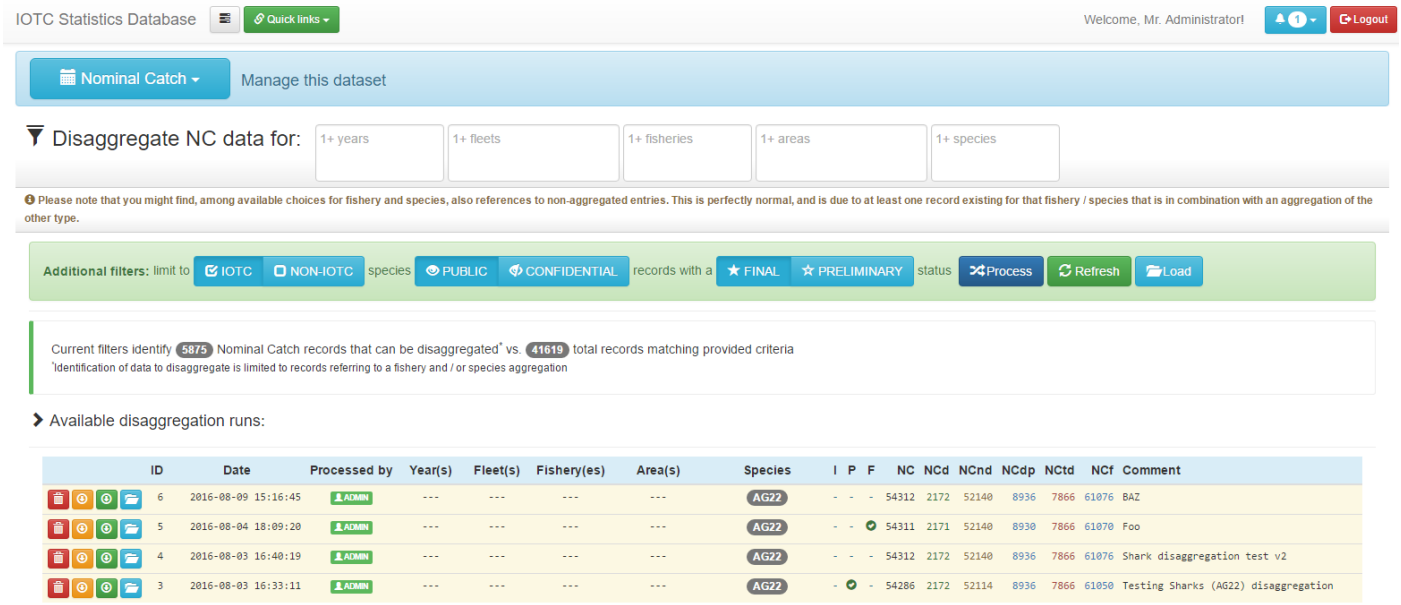

Figure 15. Listing currently available *disaggregation runs*

Catch and effort

Beside the common management processes already described for all the datasets, entitled users are enabled to perform additional sets of operations on the Catch and Effort records. These operations are mostly meant to support the data assistant into identifying possible inconsistencies within the data (prior or following the bulk upload of new Catch and Effort records) as well as provide the ground for the production of the datasets required by each working party.

The first type of additional operation that can be performed on Catch and Effort records is their **reallocation in space and / or time**.

Users can select a time frame, a set of fleets (both optional) and one fishery type (*Longlines*, *Purse seines and bait boats*, *Other coastal gears*) and then decide the level of temporal (by month vs. no reallocation) and spatial (original fishing grounds vs. 1°x1° vs. 5°x5° grids) reallocation.

If spatial reallocation is applied (that is, users select to reallocate records to either 1°x1° or 5°x5° grids) the system will take advantage of the geospatial features available within the RDBMS and proportionally allocate catches and efforts from the original grid to any of the overlapping regular grids by resolution.

The outcome of the reallocation process can be either stored (for later reference) or downloaded as a CSV file in the same format currently expected for dissemination prior to each Working Party.

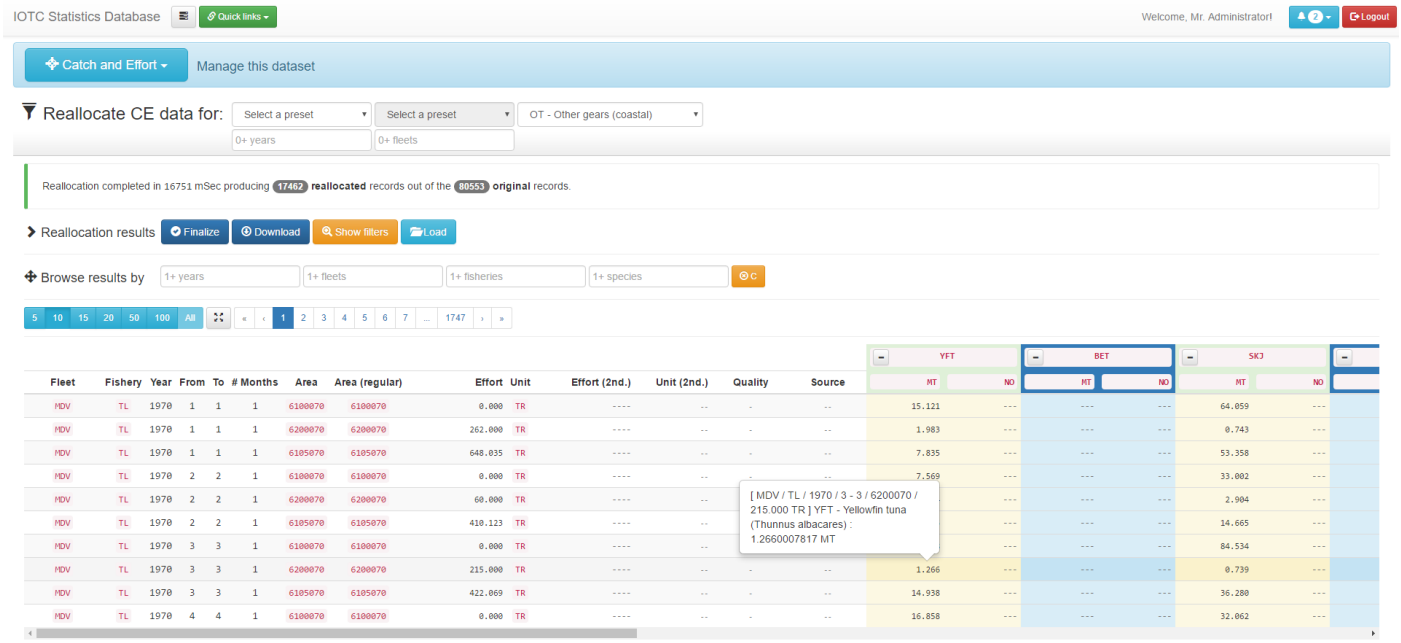

Figure 16. Spatial / temporal reallocation of catches and efforts for *Other gears (coastal)*

In a similar way to what already shown for Nominal Catches, summaries for reported catches / efforts and CPUE can also be calculated and displayed as different types of charts according to a specific grouping criteria (see Figure 17).

Filtered quantities can also be limited to specific areas provided either as a WKT text describing the boundaries of interest or by selecting multiple geometries already available within the system. In both cases, the user-selectable intersection type (*contained* / *proportionally overlapping* / *overlapping*) will have an impact on the returned quantities.

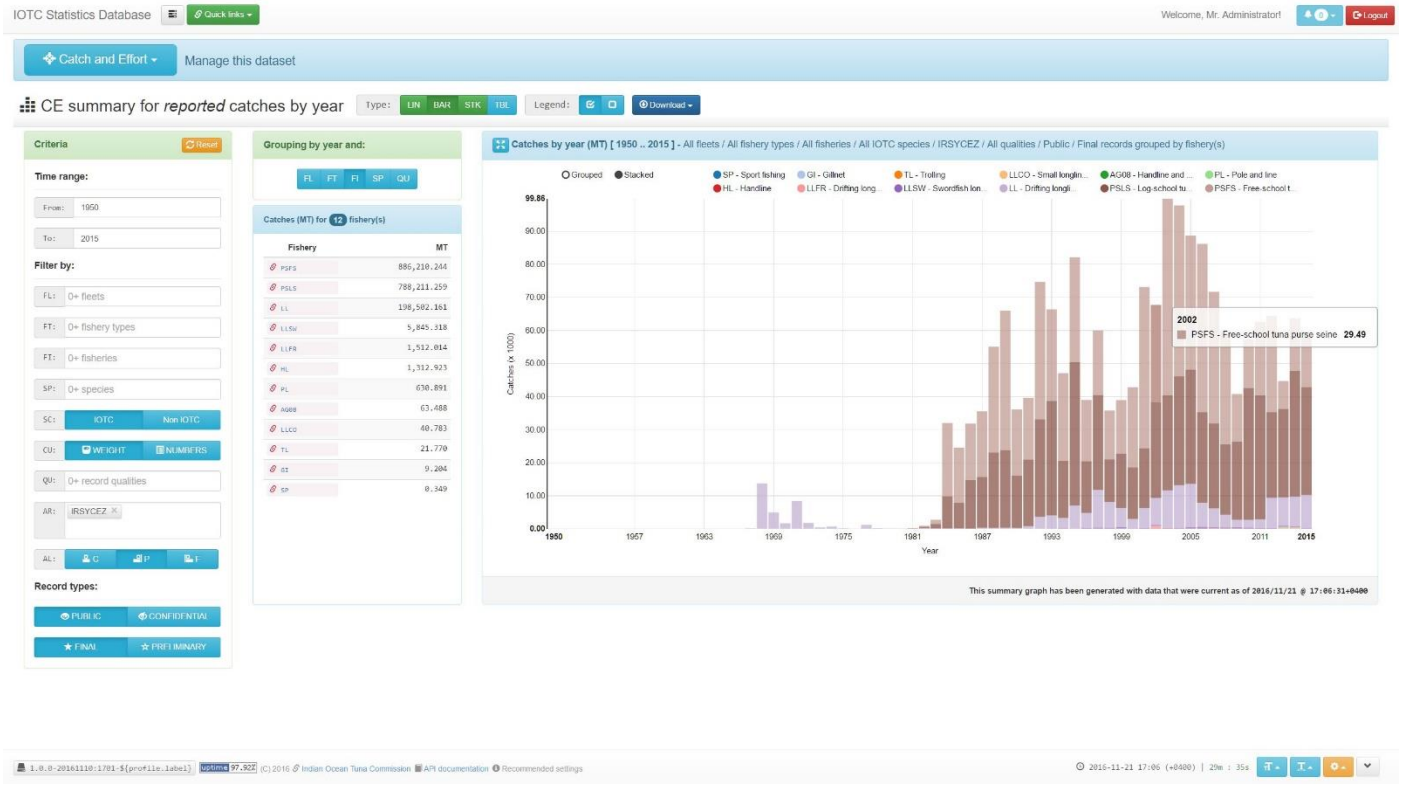

Figure 17. Reported catches summary for all IOTC species within SYC EEZ, grouped by fishery

Another additional feature available for the Catch and Effort dataset is the production of geospatial plots – including animation of data over time – for reported catches / efforts / CPUE and any given data subset.

Users can filter the data by year, month, fleet, fishery and produce *heatmaps* at different level of resolution and with different tiles as background.

Also, data can be plotted either by available number of records per grid or by overall quantities (catches, efforts or CPUE). Furthermore, it is possible to limit the geospatial plot to any custom area by simply providing the vertex coordinates in a WKT-like format and superimpose (and intersect) layers for common IO areas, including high seas and all EEZs.

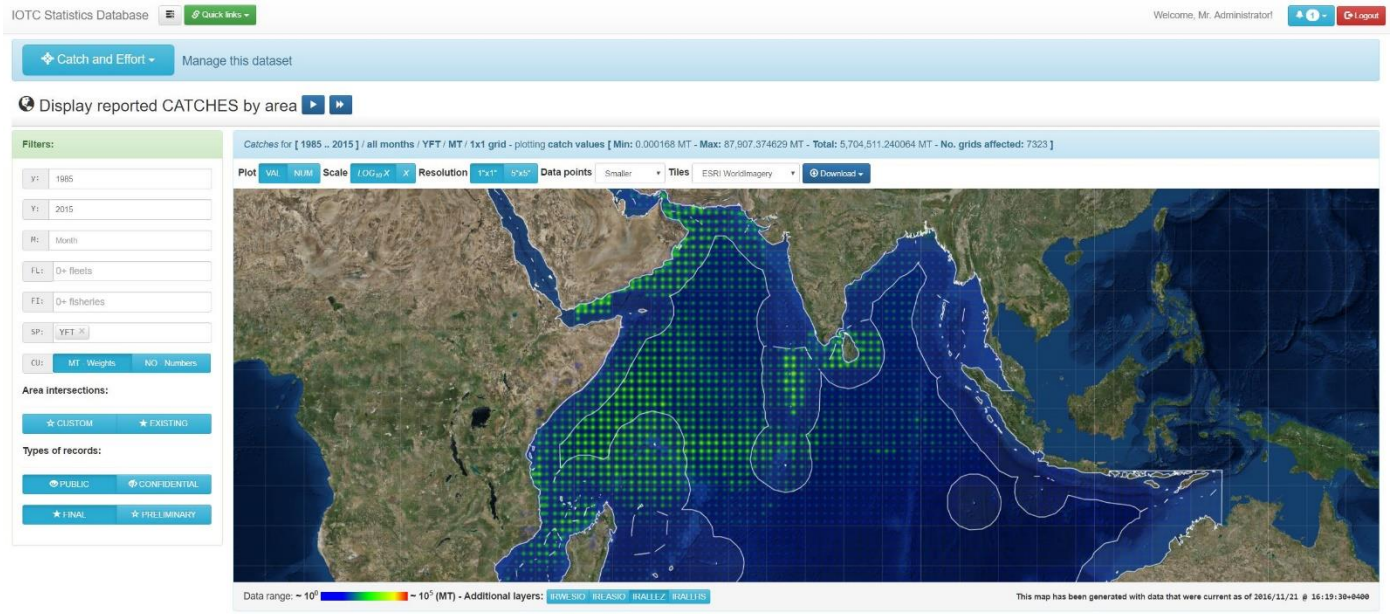

Figure 18. Plotting YFT catches for 1985 – 2015 on 1°x1° grids

Figure 18 shows an example of the geographical distribution of YFT catches in 1985-2015, with grids having a resolution of 1°x1° degrees. In this specific figure, reported catches (limited to final and public records) are plotted using a logarithmic scale. The IO EEZ layer is superimposed to the plot.

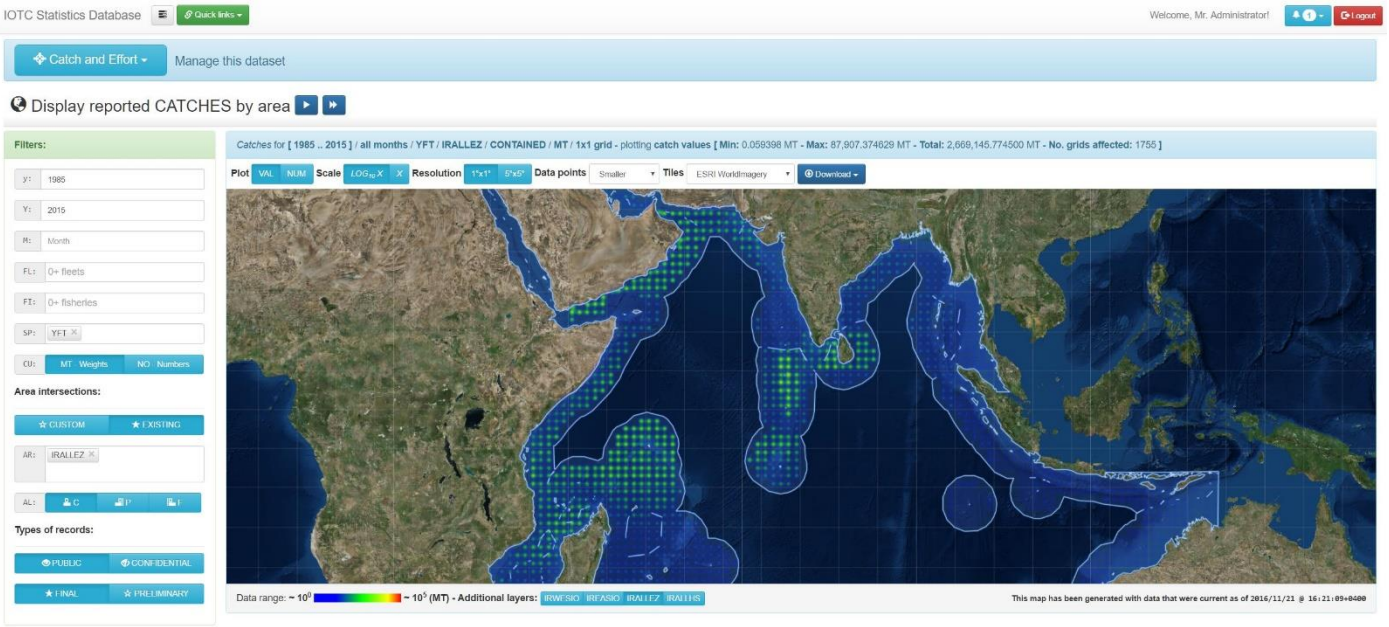

Figure 19. Plotting YFT catches for 1985 – 2015 on 1°x1° grids intersecting with all EEZ areas

Outputs in Figure 19 are based on exactly the same filtering criteria as Figure 18, with catches intersected (and retained, in the final dataset) with all IO EEZ areas.

Data-to-area intersections can be calculated either by providing a WKT string describing the boundaries of a custom area or by selecting one or more of the currently available area geometries. In both cases, the type of intersection can be selected among:

- Contained: only grids that are fully contained within the provided boundaries will be considered;
- Proportionally overlapping: all grids that are overlapping (even partially) with the provided boundaries will be considered and the proportion of catches to keep will be calculated based on the extent of the relative overlapping;
- Overlapping: all grids that are overlapping (even partially) with the provided boundaries will be considered and catches for all these grids will be kept as they are (no proportional allocation applied)

Therefore, given a set of filtering criteria and geospatial boundaries, the relationship between computed catches by intersection type is as follow:

$$Catch_{contained} \leq Catch_{proportional} \leq Catch_{overlapping}$$

When setting a time-frame for the display of reported catches, the system will allow users to calculate and present two type of animations: on a month-by-month basis (for the entire timeframe) or on a year-by-year basis.

This feature – which is also available for reported efforts and CPUE - is extremely useful to identify seasonality and common patterns within the reported data as is the possibility of limiting the displayed data to a given month over the entire timeframe (see Figures 20a – 20b – 20c – 20d).

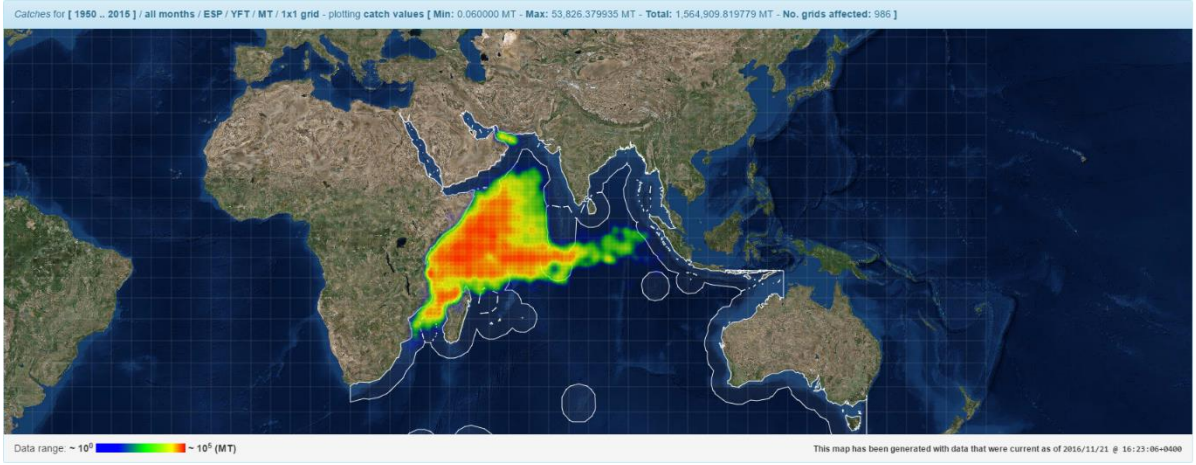

Figure 20a. YFT reported catches for ESP (entire time-series, MT, logarithmic scale over 1°x1° degree grids)

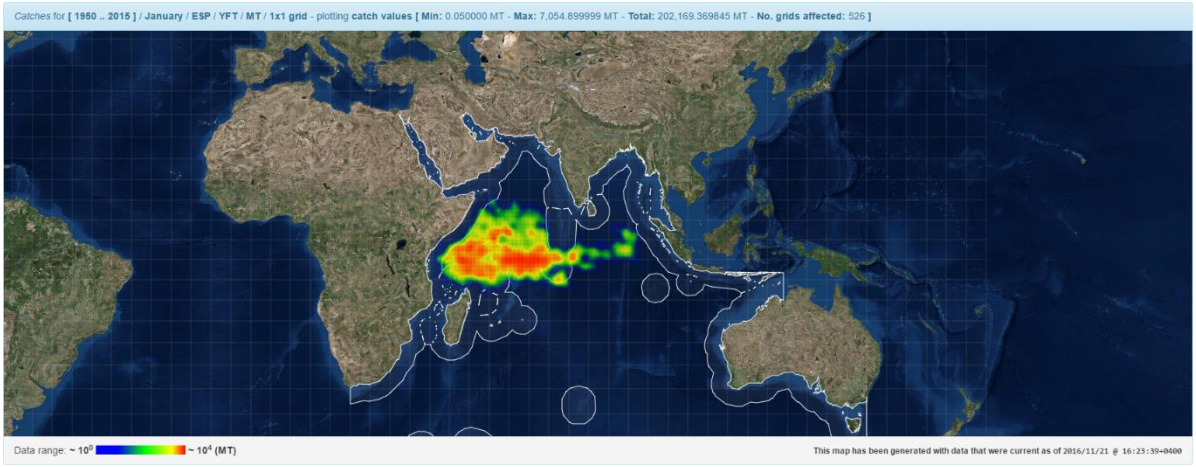

Figure 20b. YFT reported catches for ESP (MT, logarithmic scale over 1°x1° degree grids) during January (all years)

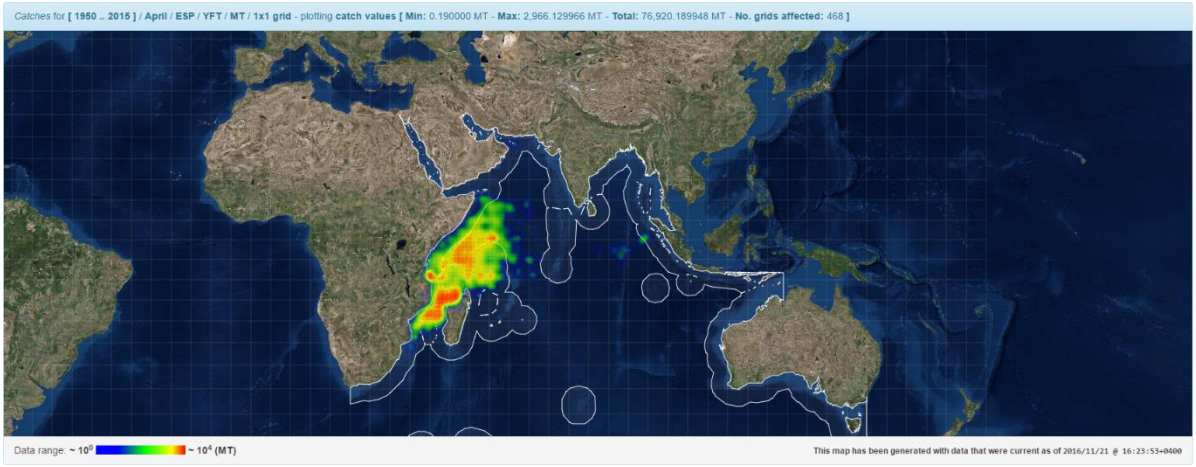

Figure 20c. YFT reported catches for ESP (MT, logarithmic scale over 1°x1° degree grids) during April (all years)

Figure 20d. YFT reported catches for ESP (MT, logarithmic scale over 1°x1° degree grids) during October

These plots can be produced with close-to-realtime response times: this allows using the feature to actually 'animate' the reported catch / effort / CPUE distributions and show how they evolve during different period of the year (or across years).

Data can be downloaded as images (to be embedded within reports and other documents) or as CSV files for further analysis.

Size-frequency

Beside the common management processes already described for all the datasets, entitled users can perform additional sets of operations on the Size and Frequency records. These operations are mostly meant to support the data assistant into identifying possible inconsistencies within the data (prior or following the bulk upload of new Size and Frequency records) as well as provide the ground for the production of the datasets required by each working party.

The first type of additional operation that can be performed on Size and Frequency records is the **conversion of non-standard length or weight units to standard ones**, followed by the **conversion of lengths to weights** (based on the available conversion equations) and the **redistribution of samples across size bins**. Users can select a time frame, a set of fleets, a set of gears and one or more species currently grouped by Working Party species (Billfishes / Neritic Tunas / Temperate Tunas / Tropical Tunas).

The outcome of the redistribution process can be either stored (for later reference) or downloaded as a CSV file in the same format currently expected for dissemination prior to each Working Party.

Users can identify and browse strata for which - due to the reported length units for a given species - there is no length – weight conversion equation available and apply a different color shade (*heatmap*) to the number of fishes per size bin, in order to show where the highest concentration of samples for each specific strata lies.

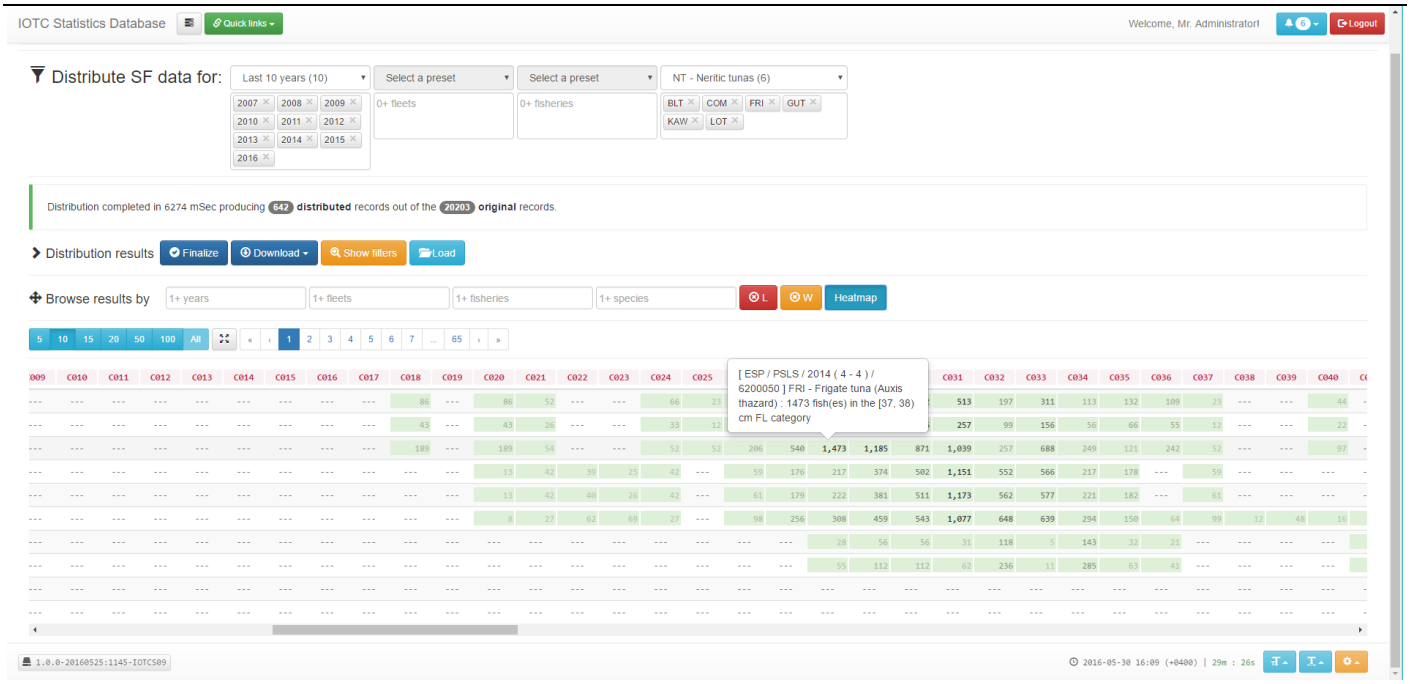

Figure 21. Swordfish reported (and converted to standard) length distribution

In terms of reported Size and Frequency data, users can also **display the raw and reported samples size distribution (by species and measurement type)**, eventually filtering the results by fleet, fishery and timeframe.

Produced results can be displayed as charts of three types (stacked bar chart, stream chart and expanded chart) and exported either as PNG images (for inclusion within other reports and summaries) or as CSV files.

Displayed data is reported in terms of raw number of samples per length class and *estimated* samples number per length class, based on the information provided by the original data submitter.

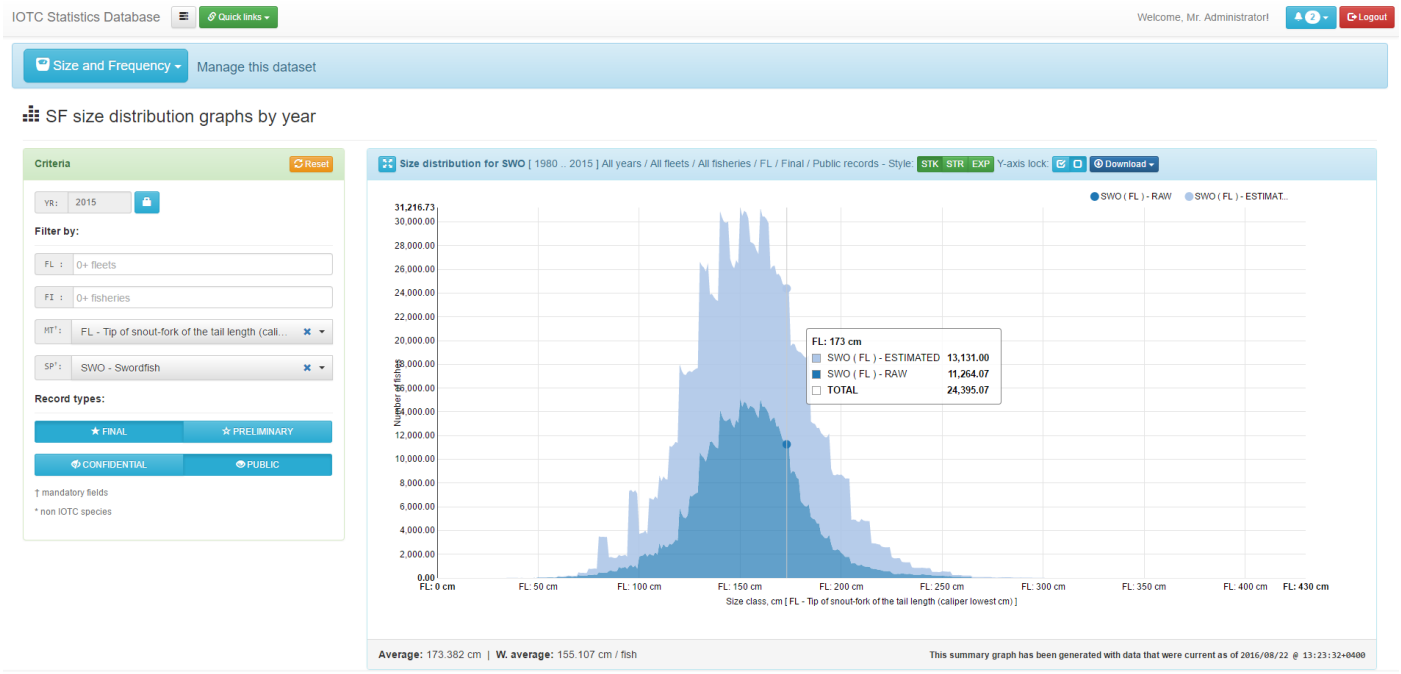

Figure 22a. Swordfish size distribution (FL – Fork length) as a stacked chart

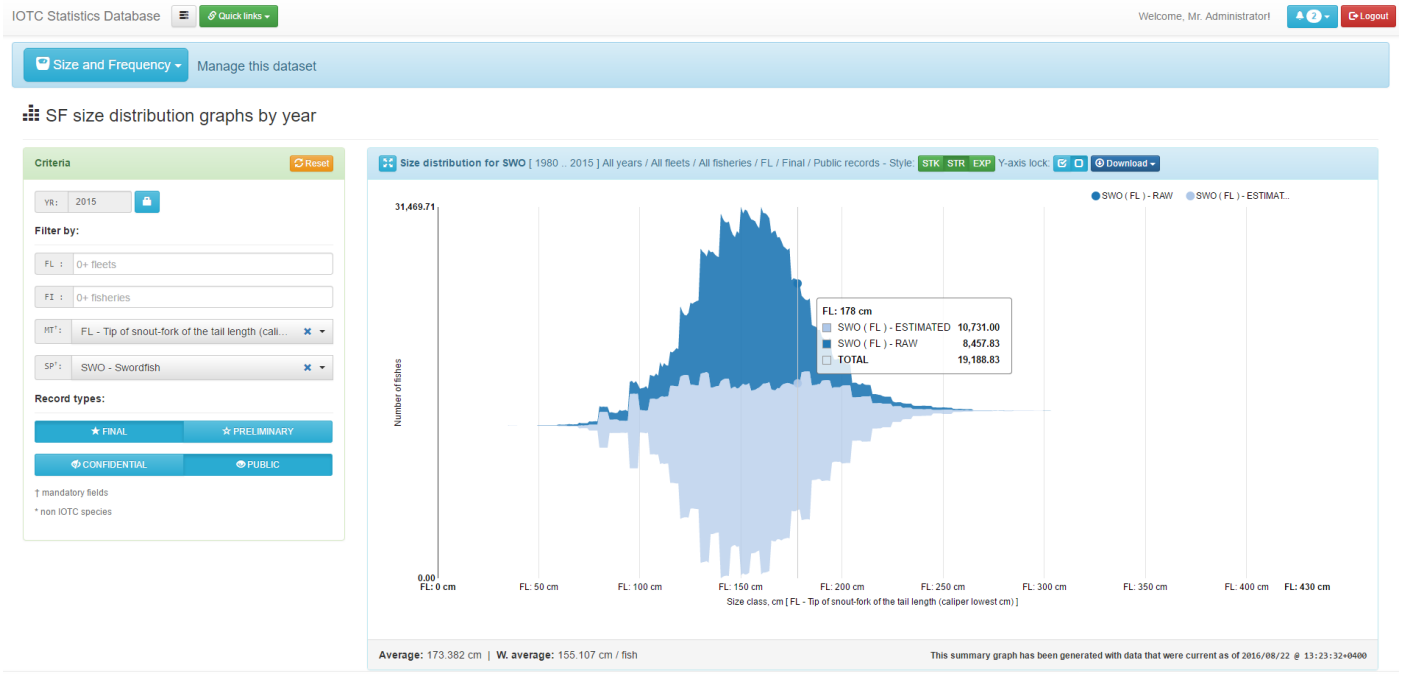

Figure 22b. Swordfish size distribution (FL – Fork length) as a stream chart

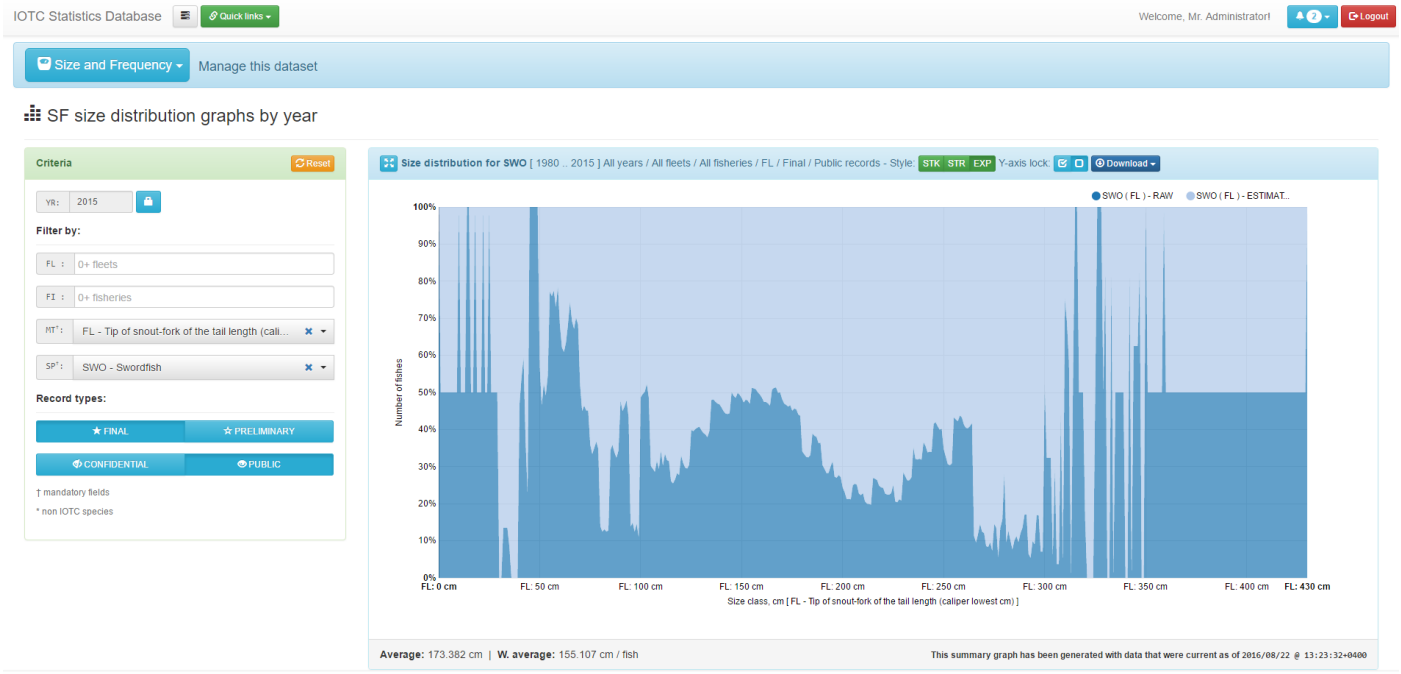

Figure 22c. Swordfish size distribution (FL – Fork length) as an expanded chart

Another set of additional features available for the Size and Frequency dataset is the **production of geospatial plots of sampled numbers / raw samples** for any given subset of the data.

Similar to what already described for the Catch-and-Effort dataset, users can filter the data by year, month, fleet, fishery and species, and eventually produce static or animated *heatmaps* at different level of resolution.

In this case, data can be plotted either by number of raw samples or by reported sample numbers and it is also possible to limit the geospatial plot to any custom area by simply providing the vertex coordinates in a WKT-like format or intersect the results with well-known IO areas.

Figure 23a shows an example of the geographical distribution of Skipjack (SKJ) **reported raw samples** for any available year, with grids having a resolution of 1°x1° degrees. In this specific figure, reported raw samples (limited to final and public records) are plotted using a logarithmic scale.

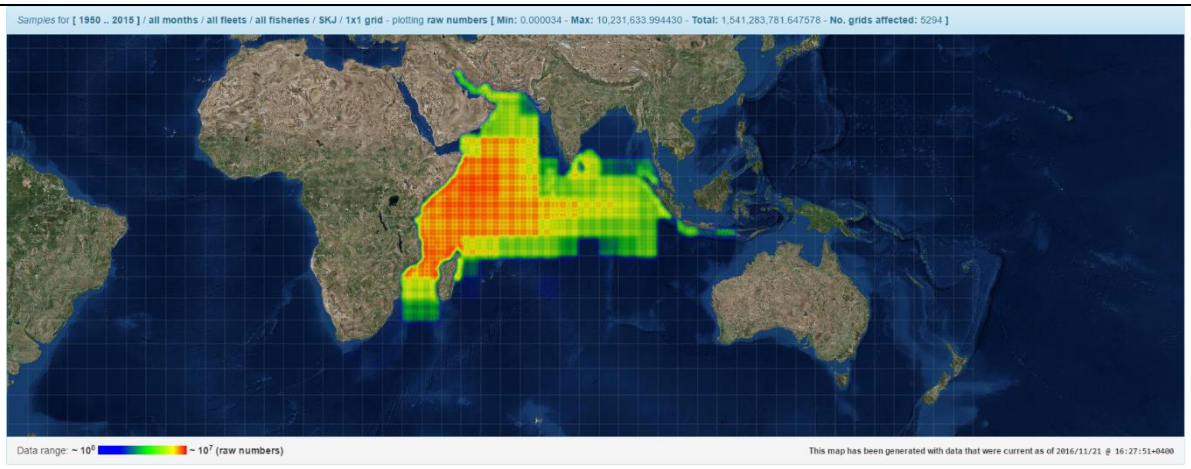

Figure 23a. Reported raw samples distribution for SKJ (logarithmic scale over 1°x1° degree grids)

Figure 23b shows an example of the geographical distribution of Skipjack (SKJ) **reported sampled numbers** for any available year, with grids having a resolution of 1°x1° degrees. In this specific figure, reported sampled numbers (limited to final and public records) are plotted using a logarithmic scale.

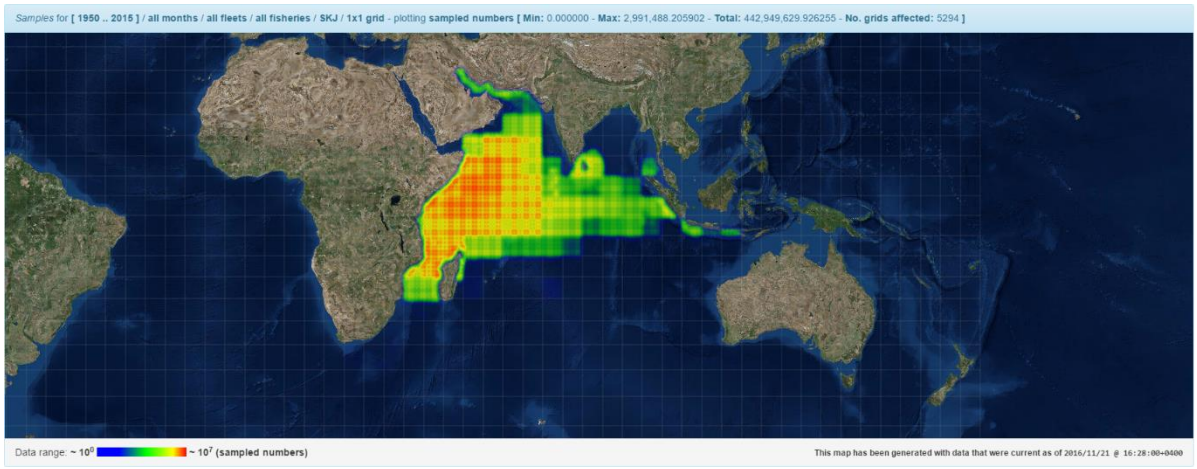

Figure 23b. Reported sampled numbers distribution for SKJ (logarithmic scale over 1°x1° degree grids)

Fishing crafts

No specific operations are available for the 'Fishing crafts' dataset.

Discards

No specific operations are available for the 'Discards' dataset.

Country indicators

No specific operations are available for the 'Country indicators' dataset.

Fish prices

No specific operations are available for the 'Fish prices' dataset.

Other tools

The new IOTC data management system provides also a few other tools for the analysis and management of the geospatial information currently stored within the database.

These provide the following features:

- Filter and display the geospatial details for all areas (of any type) and download their definition as WKT text

- Analyze and display the geospatial details for areas (of any type) by the fraction of Indian Ocean covered

The first type of functionality (see Figure 24) is useful to assess the placement and shape of the geospatial areas used by the system and export their WKT definitions (when needed). Geospatial areas can be displayed either in their raw form (as regular grids, when is the case) or in their *intersected* form, taking into account the current geospatial definition of the entire Indian Ocean.

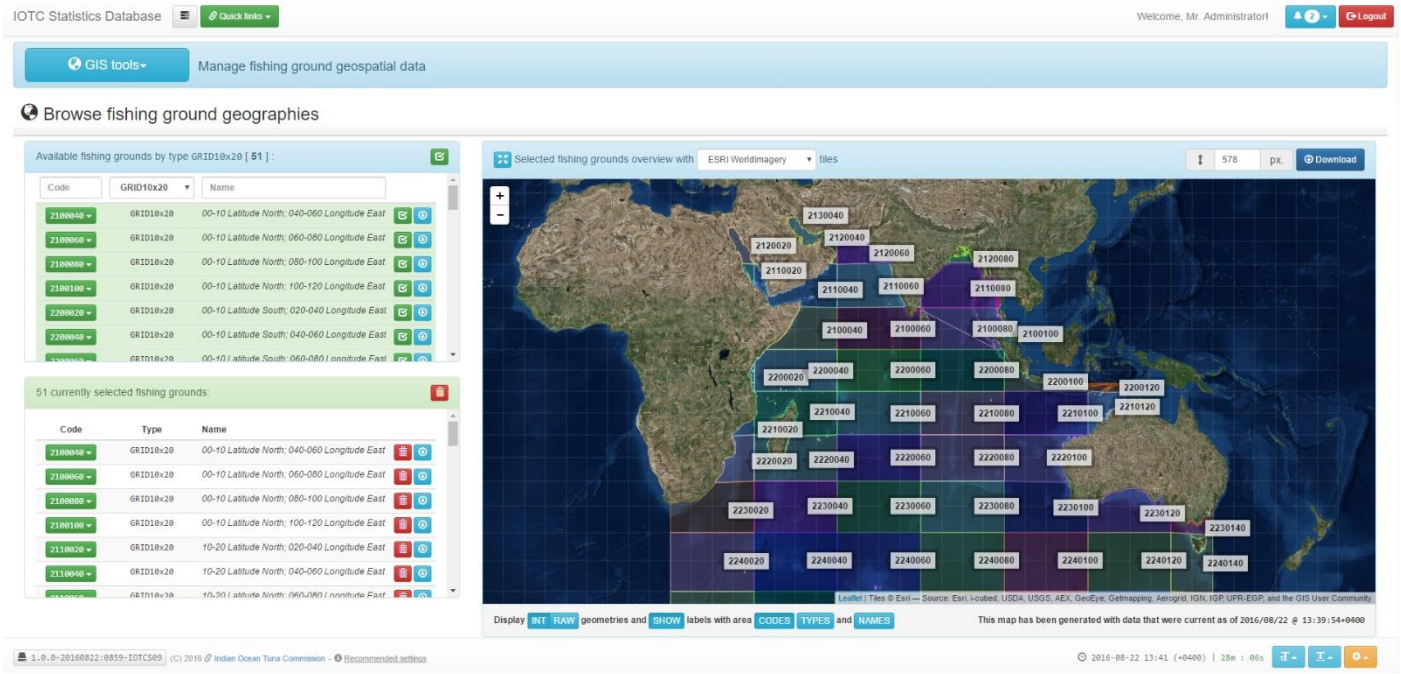

Figure 24. Displaying 10x20 grids and their intersection with the Indian Ocean

The second type of functionality (see Figure 25) is particularly useful to support the internal data-curation process, as it allows identifying those areas which – although part of the available grids – do not actually intersect with the current geospatial definition of the entire Indian Ocean.

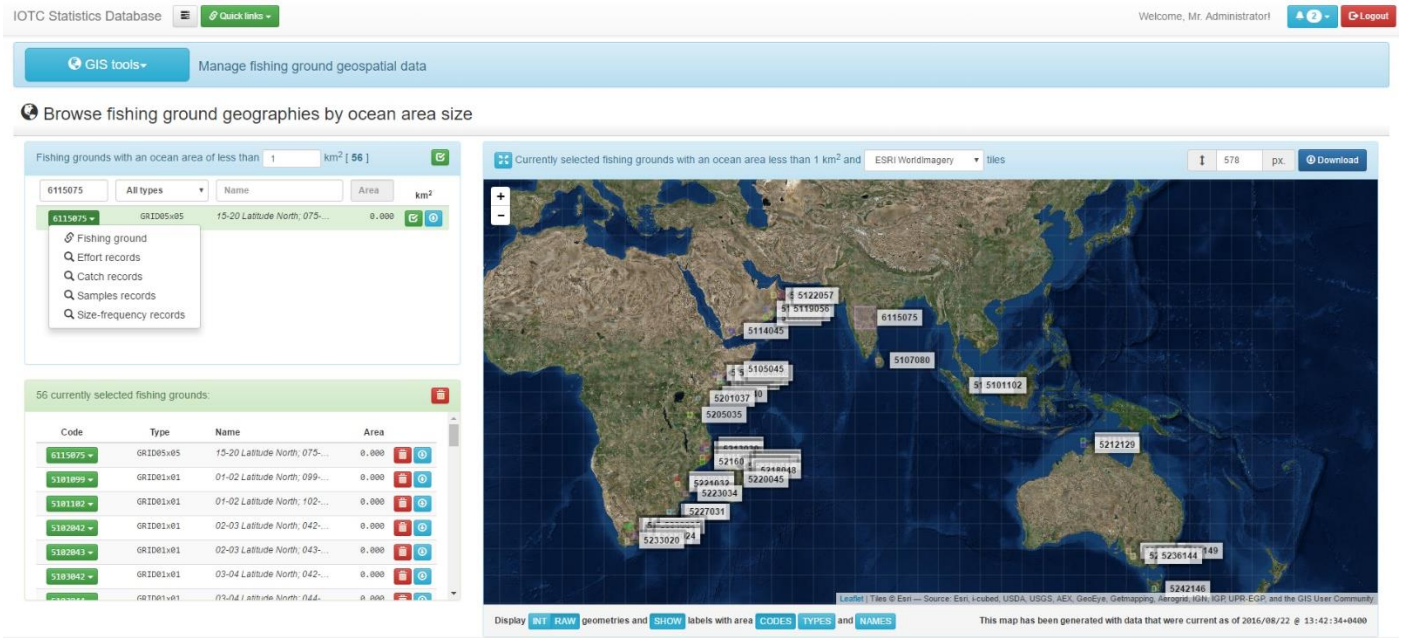

Figure 25. Displaying all grids / areas having an intersection with the Indian Ocean that is less than 1Km^2

Both functionalities allow identifying all records (by dataset, including catch and effort and size-frequency) that do refer to any of the displayed areas.

Future developments

One of the core concepts behind the adoption of the new data management system is that some of the features it provides to internal users could also be made available to a broader audience.

We envisage that the reported catch and effort reallocation, the reported size and frequency redistribution as well as the summary and geospatial plots that could be built on top of the currently available datasets can be extremely useful features for the scientific community.

The Secretariat will ensure (and this is already enforced by the adoption of the built-in access control mechanism in place within all layers of the new system) that only data that can be safely disseminated is made publicly available, whereas external users with special grants could get access to more fine-grained datasets and functionalities.

Also, the possibility of sharing data and processes in a formal and automated way (by means of the REST services APIs – see the Appendix A1 for some basic examples) will open up endless integration possibilities for scientists worldwide.

Feedbacks / improvements

Some of the processes implemented so far should definitely benefit from feedback coming from the community.

For the time being, the **Nominal Catch Disaggregation process** is using proxy fleets / gears to identify non-aggregated records that could break down aggregated ones.

The current disaggregation procedures can reference any number of records from proxy fleet / gears and the outputs – when available – are used to proportionally assign catches for aggregated records.

Currently, the system does not set any lower threshold in terms of the minimum number of identified records that the process needs in order to consider a disaggregation output as '*reasonably accurate*': this threshold might depend on the species / fisheries involved and – at the same time – on the specific disaggregation procedure that is triggered.

When implemented, such a threshold might provide additional information to scientists in term of the uncertainty related to the produced disaggregated Nominal Catch records.

This is only an example of the feedback that the community could provide to the Secretariat. We expect – as soon as the new system and its remote APIs are made available to a broader audience – that the suggestions and requests coming from the scientific community will greatly contribute to increase the usefulness and efficacy of the data management processes as a whole. This will have, at the same time, the positive consequence of enabling users to have a more in-depth and formal understanding of the entire IOTC data management chain.

Appendix

A1. Programmatically accessing IOTC data services via remote APIs

APIs (*Application Programming Interfaces*) specify the definition - in terms of business processes performed, data exchange protocol and access point - required to invoke services on a remote system.

In this case, the new IOTC data management system is inherently built with facilities to expose all of its business processes through REST APIs ([IOTC-2016-WPDCS12-26_Rev1]).

For this to be effective, data-consumers (i.e. the users that want to incorporate *live* data coming from the IOTC data management system) need to be assigned an API-key that will uniquely identify their business requests and limit the features / functionalities they can have access to.

Results from two simple examples of the new IOTC data management processes integration within R follow: in this case, two remote services (one for the production of aggregated nominal catch data by species and another for the production of reallocated catch-and-effort data by grid size) are invoked and the returned data is used – within R itself – to plot sample graphs and charts.

The returned data is – at any time – the *live* content of the IOTC database: this means that a scientist that wants to produce charts and graphs from the available IOTC dataset does not need to wait until these datasets are publicly disseminated through the IOTC website.

As long as users have an API key that provides valid access to the dataset or operation of interest, they could use this API key (in a similar way to what displayed in the R examples) to access information through the IOTC data management system or execute data manipulation processes on the IOTC server and retrieve the results.

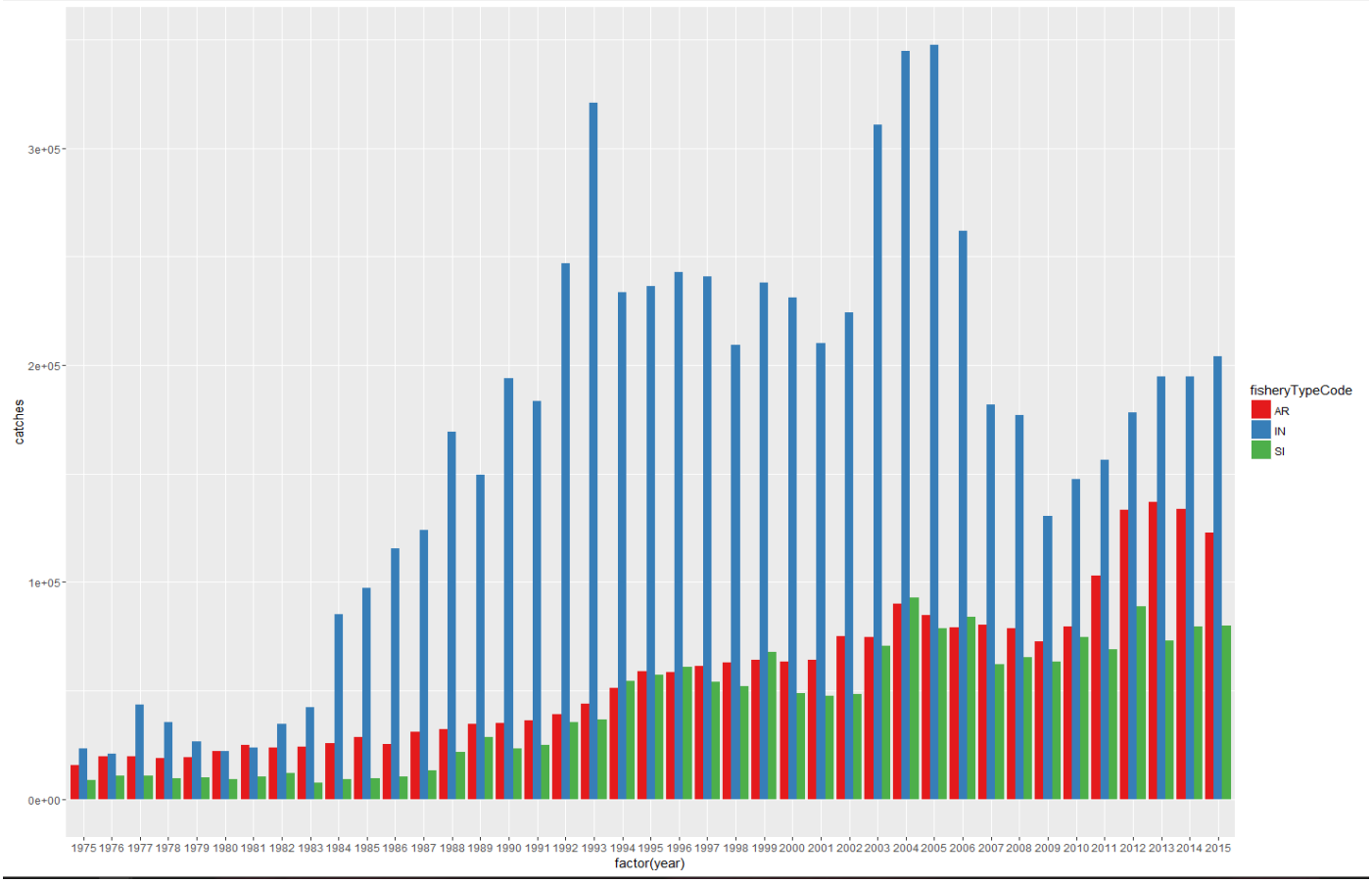

Figure A1.1. YFT nominal catches in the timeframe between 1970 and 2015 grouped by year and fishery type

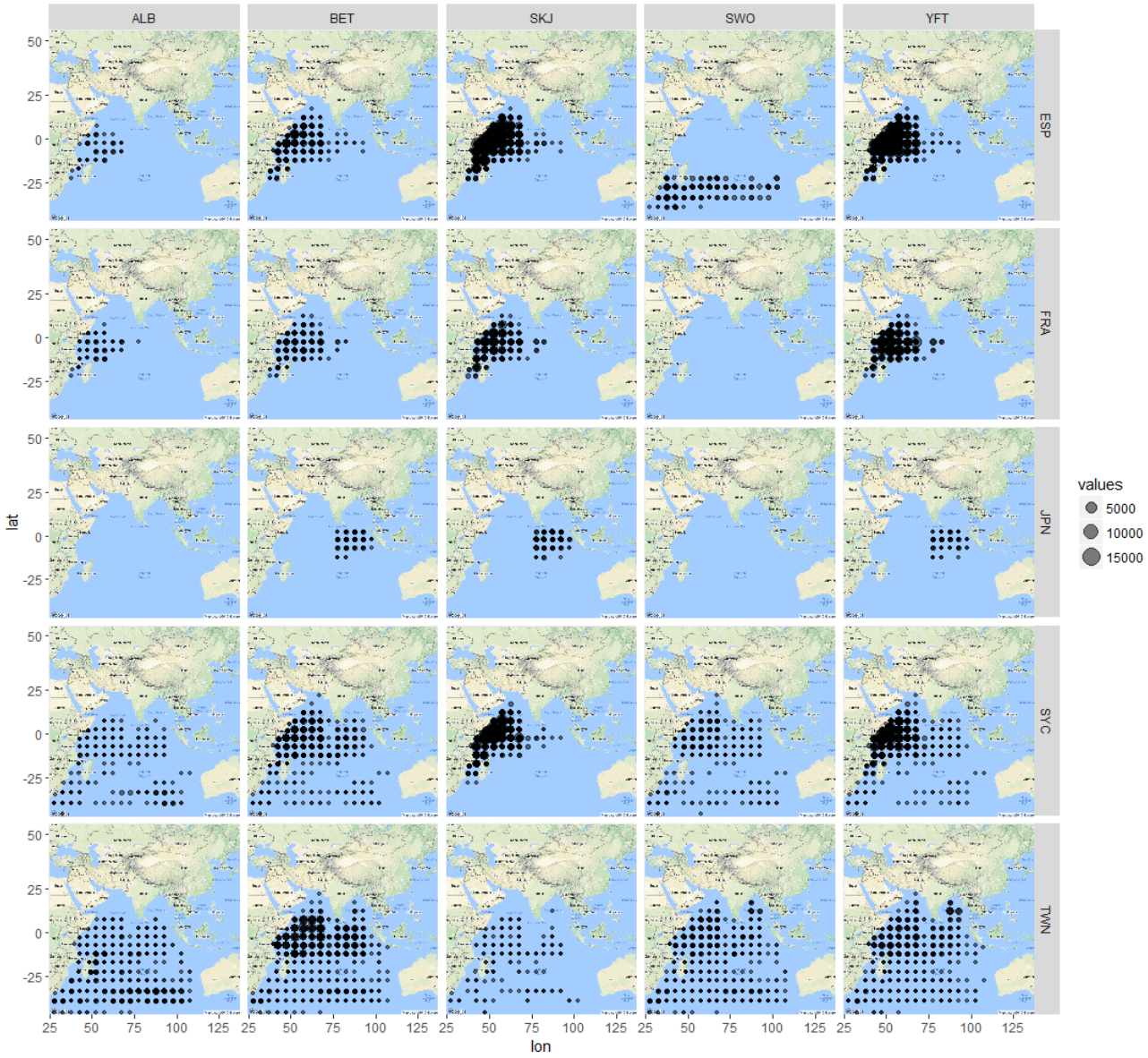

Figure A1.2. Catches in MT from the main 5 species, plotted over 5x5 degrees grid and faceted by species and fleet, for the years between 2010 and 2015 and ESP, FRA, JPN, SYC and TWN longliners or purse-seines

A2. The Nominal Catch disaggregation process

The ***Nominal Catch disaggregation*** is the process of breaking down all records – from the Nominal Catch dataset – that do refer to either a gear or a species aggregate (or both) in order to produce a dataset that contains only disaggregated records (i.e. referring only to single species and gears).

Its outputs are crucial for the achievement of valid stock assessments results because, being fully disaggregated, they help scientists in providing more accurate estimates and advices.

It is a non-linear yet repeatable process that can also be used to reconstruct Nominal Catch time series for all those combination of species and gears whose information is *sparse* in time (especially for past decades, when data was collected from hard copies and other 'indirect' sources).

The key concept behind the process is that catch quantities from every aggregated record can be proportionally assigned to a different combination of species or gears (in the list of entries belonging to the species / gears aggregates being processed).

In order to identify these proportions, the process applies a sequence of multiple disaggregation procedures that can identify relevant *proxy records* from within the original Nominal Catch dataset.

Once these proxy records are identified, the proportions of catches by species or gears are used to assign the original catch quantity to a combination of (potentially) multiple, *synthetic* records that are the output of the disaggregation process itself.

The procedures used to identify the proxy records do filter the original dataset by *fleet*, *type of operation* (Artisanal / Semi-Industrial / Industrial), *region*, *area* and *timeframe*: they rely on a specific configuration table that assigns – to each and every current combination of fleet / gear / area for which at least one Nominal Catch record exists – a *region* of most-likely operation.

Currently, the process adopts eight different disaggregation procedures that are complemented by a ninth procedure (*manual disaggregation*) triggered when no proxy record can be identified by any of the others.

The high-level data flow is as follows:

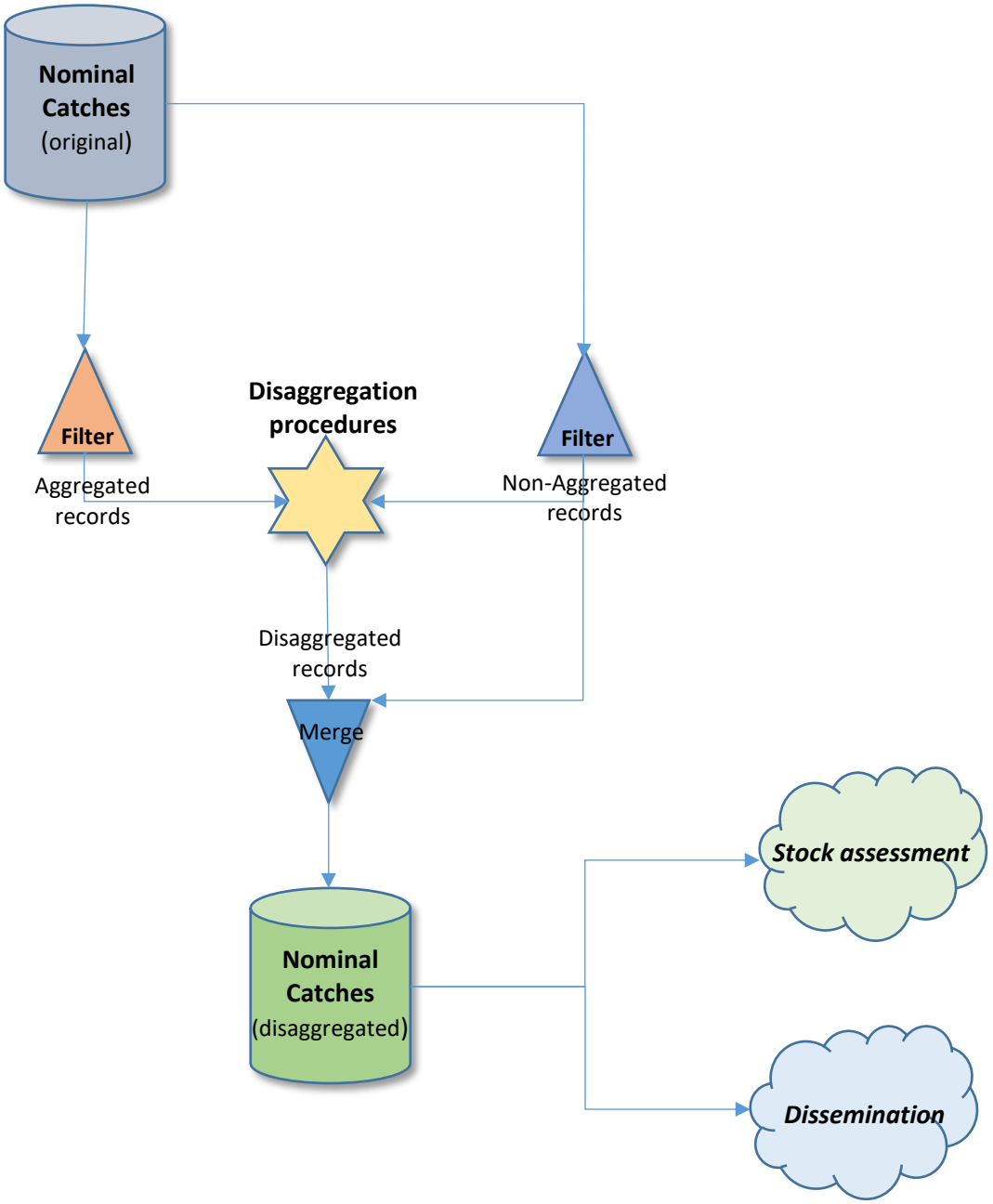

Figure A2.1. The Nominal Catch Disaggregation process

Its pseudo-code implementation would be similar to the following:

```
var DisaggregationProcedures := { P1, P2, ..., P8 };

var NC_Original := { NC1, NC2, ..., NCn };

var NC_NonAggregated := { };
var NC_Aggregated := { };

var NC_Proxies := { };

var NC_Disaggregated := { };

for each NC in NC_Original:
      if !NC.isAggregated
            NC_Disaggregated.add(NC);
            NC_NonAggregated.add(NC);
      end if;
end for;

for each NC in NC_Original:
      if NC.isAggregated
            loop: for each Proc in DisaggregationProcedures:
                  NC_Proxies := Proc.apply(NC, NC_NonAggregated);
                  if NC_Proxies != { }
                        NC_Disaggregated.addAll(NC.breakdown(NC_Proxies));
                        break loop;
                  end if;
            end for;

            //Manual breakdown is required if none of the procedures identifies
            //any proxy record to use for the disaggregation of current record
            NC_Disaggregated.addAll(NC.manualBreakdown);
      end if;
end for;

return NC_Disaggregated;
```

Listing A2.2. Pseudo-code for the Nominal Catch Disaggregation process

Where:

- *P1, …, P8* are the eight currently available Disaggregation Procedures;
- *NC1, …, NCn* is the input Nominal Catch dataset;
- *procedure*.**apply**(<*NC record*>, <*non aggregated NC records*>) returns the proxy records (according to the current disaggregation *procedure*) for the aggregated record <*NC record*> being processed, as these are identified within the full set of <*non aggregated NC records*>;
- *record*.**breakdown**(<*NC proxies*>) breaks down the original, aggregated *record* into multiple disaggregated records, whose catch quantities (and species / gears) are proportionally assigned based on the identified <*NC proxies*>;
- *record*.**manualBreakdown** prompts users for their own, manual breakdown of the original aggregated *record*, as none of the disaggregation procedure was able to identify any valid proxy record for it;

The process by itself is quite simple and straightforward: its ability in identifying proper *proxy records* for the proportional breakdown of the original catches lies in the definition of the disaggregation procedures and in the *fleet* / *type of operation* / *region* / *area* mappings.

The currently available Disaggregation Procedures are defined in the following table:

Procedure #	Fleet	Type of operation	Region	Area	Years
1	Same	Same	Same	Same	Same
2	Same	Same	Same	Same	+ / - 5 years

3	*Any*	Same	Same	Same	Same
4	Same	Same	Same	Same	+ / - 10 years
5	Same	Same	*Any*	Same	Same
6	*Any*	Same	*Any*	Same	Same
7	*Any*	Same	*Any*	Same	*Any*
8	*Any*	Same	*Any*	*Any*	*Any*

Table A2.3. Disaggregation procedures definition

When moving from procedure #1 to procedure #8, the constraints for the identification of proxy records become more relaxed.

The first procedure, basically, looks for proxy records for the same fleet / type of operation / region / area and fishing year that do refer to any of the species / gear belonging to the aggregates involved.

Conversely, the last procedure will look for proxy records from any fleet, any region, any area and any fishing year, as long as the type of operations are the same, and that at the same time do refer to any of the species / gear belonging to the aggregates involved.

Intermediate procedures have a behaviour that is a mixture of these two.

Country	Rep. country	Gear	Area	Region	Type of operation
…	…	…	…	…	…
ESP	ESP	LLEX	IREASIO	EASIO	IND
ESP	ESP	LLEX	IRWESIO	WESIO	IND
ESP	ESP	PS	IREASIO	EASIO	IND
ESP	ESP	ELL	IREASIO	SWEIO	IND
ESP	ESP	LL	IREASIO	SWEIO	IND
ESP	ESP	ELL	IRWESIO	SWEIO	IND
ESP	ESP	LL	IRWESIO	SWEIO	IND
ESP	ESP	BB	IRWESIO	WESIO	IND
ESP	ESP	PS	IRWESIO	WESIO	IND
ESP	ESP	SUPP	IRWESIO	WESIO	IND
FRA	**FRA**	**HAND**	**IRWESIO**	**MOZCH**	**ART**
FRA	FRA	TROL	IRWESIO	MOZCH	ART
FRA	FRA	ELL	IRWESIO	SWEIO	IND
FRA	FRA	PS	IREASIO	EASIO	IND
FRA	FRA	PS	IRWESIO	WESIO	IND
FRA	REU	LLCO	IRWESIO	SWEIO	ART
FRA	REU	HAND	IRWESIO	SWEIO	ART
FRA	REU	HATR	IRWESIO	SWEIO	ART
FRA	REU	TROL	IRWESIO	SWEIO	ART
FRA	REU	ELL	IRWESIO	SWEIO	IND
FRAT	FRA	PS	IREASIO	EASIO	IND
FRAT	FRA	HAND	IRWESIO	MOZCH	ART
FRAT	FRA	HATR	IRWESIO	MOZCH	ART
FRAT	FRA	TROL	IRWESIO	MOZCH	ART
FRAT	FRA	ELL	IRWESIO	SWEIO	IND
FRAT	FRA	PS	IRWESIO	WESIO	IND
…	…	…	…	…	…

Table A2.4. A sample of the *fleet* / *type of operation* / *region* / *area* mappings

The *fleet* / *type of operation* / *region* / *area* mappings (see Table A2.4) is used to identify – whenever needed – the region of most likely operation for a combination of fleet (country / reporting country), gear, area and type of operation.

As an example, if an aggregated record for `FRA / HAND / IRWESIO / ART` needs to be broken down into its disaggregated components, the disaggregation procedures might look for proxy records as available for any other fleet that is likely to operate in the Mozambique Channel (**MOZCH**) thus restricting the possible records based on the knowledge of the fishery.

Example of disaggregation results

To give a better understanding of the overall process, let's assume we run the disaggregation for all records referring to the ***AG14 – Billfish nei*** species aggregate.

So far, the IOTC database contains 94 Nominal Catch records referring to that species aggregate. When the process is completed, it produces 351 *disaggregated* records out of the 94 *original* ones.

Looking at the disaggregation results for the following record:

```
1986 / SUN / LL / IRWESIO / AG14 / 6 MT (Nominal Catch)
```

shows that the process was able to apply procedure #2 (same fleet, same type of operation, same region, same area and + / - 5 years of difference with respect to the fishing year) to identify the following proxy records:

```
1983  / SUN  / LL  / IRWESIO  / BUM  /  3.000000 MT
Region: WESIO  |  Operation: IN  |  Quality: Poor quality  |  Source: Liaison Officer

1984  / SUN  / LL  / IRWESIO  / BUM  /  3.000000 MT
Region: WESIO  |  Operation: IN  |  Quality: Poor quality  |  Source: Liaison Officer

1987  / SUN  / LL  / IRWESIO  / SWO  / 37.000000 MT
Region: WESIO  |  Operation: IN  |  Quality: Poor quality  |  Source: Liaison Officer
```

Which refer to **BUM** (Blue Marlin, the first two) and **SWO** (Swordfish, the last one), with total catches for BUM accounting for `6.0 MT` and total catches for SWO accounting for `37.0 MT` (over a total of `43.0 MT`).

Therefore, the original `6.0 MT` for the aggregated record are proportionally assigned to **BLM** for a total of `0.837209 MT ( 6 MT * ( 6 MT / 43 MT ) )` and to **SWO** for a total of `5.162791 MT ( 6 MT * ( 37 MT / 43 MT ) )`.

The sum of the catches assigned to each disaggregated record **is always equal** to the original catches for the aggregated record, as this process is not changing the overall Nominal Catches quantities but just reallocating these to the disaggregated records.

Using the Nominal Catch disaggregation to reconstruct catch time series

Nominal catches for **AG22 - Sharks nei** can be broken down to reconstruct and improve the catch time series for **BSH - Blue shark** and **FAL - Silky shark** (among others).

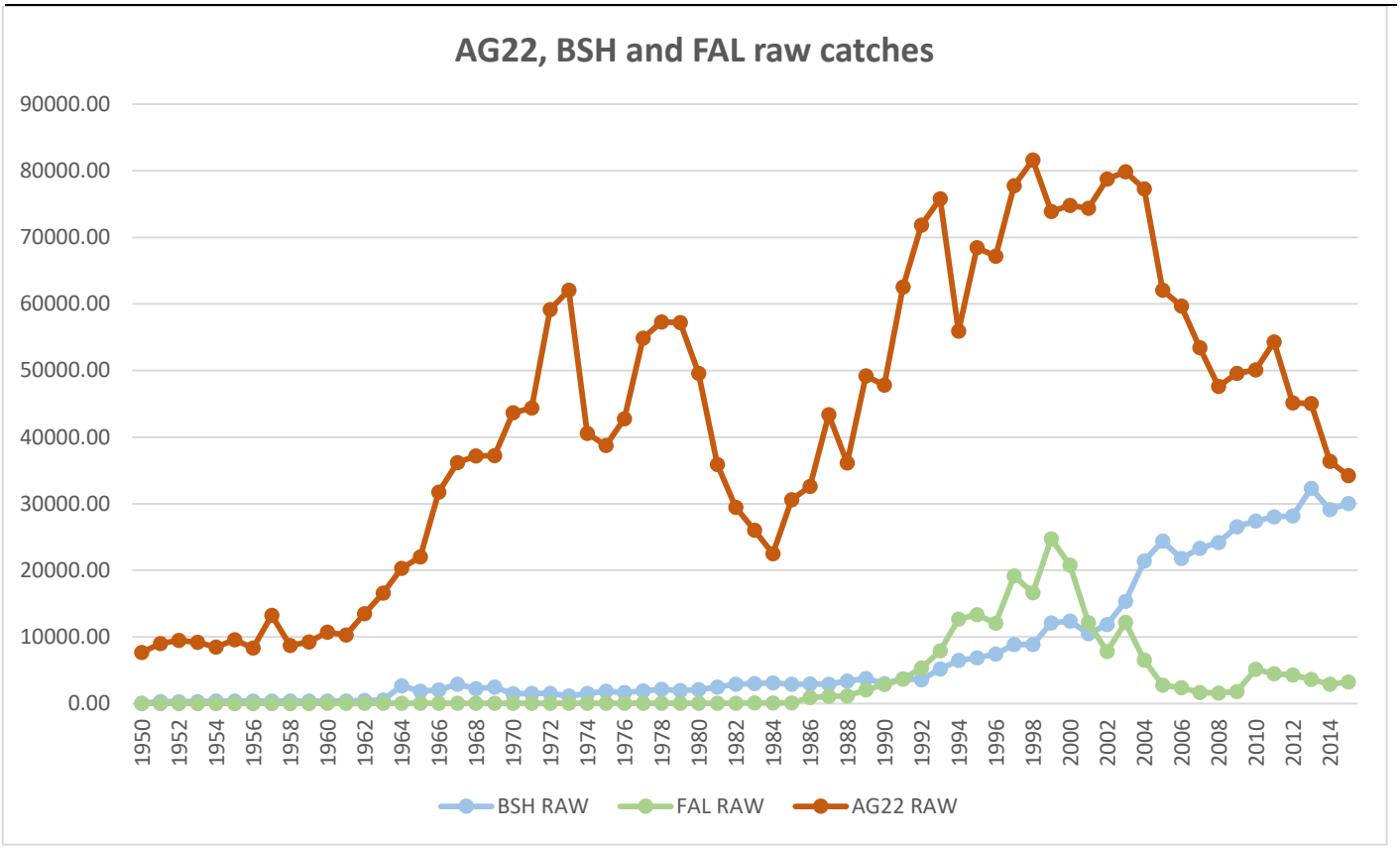

Figure A2.5. Raw nominal catches for AG22, BSH and FAL

Figure A2.5 shows that catches for the **AG22** aggregate (*Sharks nei*) are quite consistent since the '50s, whereas catches for **BSH** (*Blue shark*) and **FAL** (*Silky shark*) are of a certain relevance only from the '90s onwards.

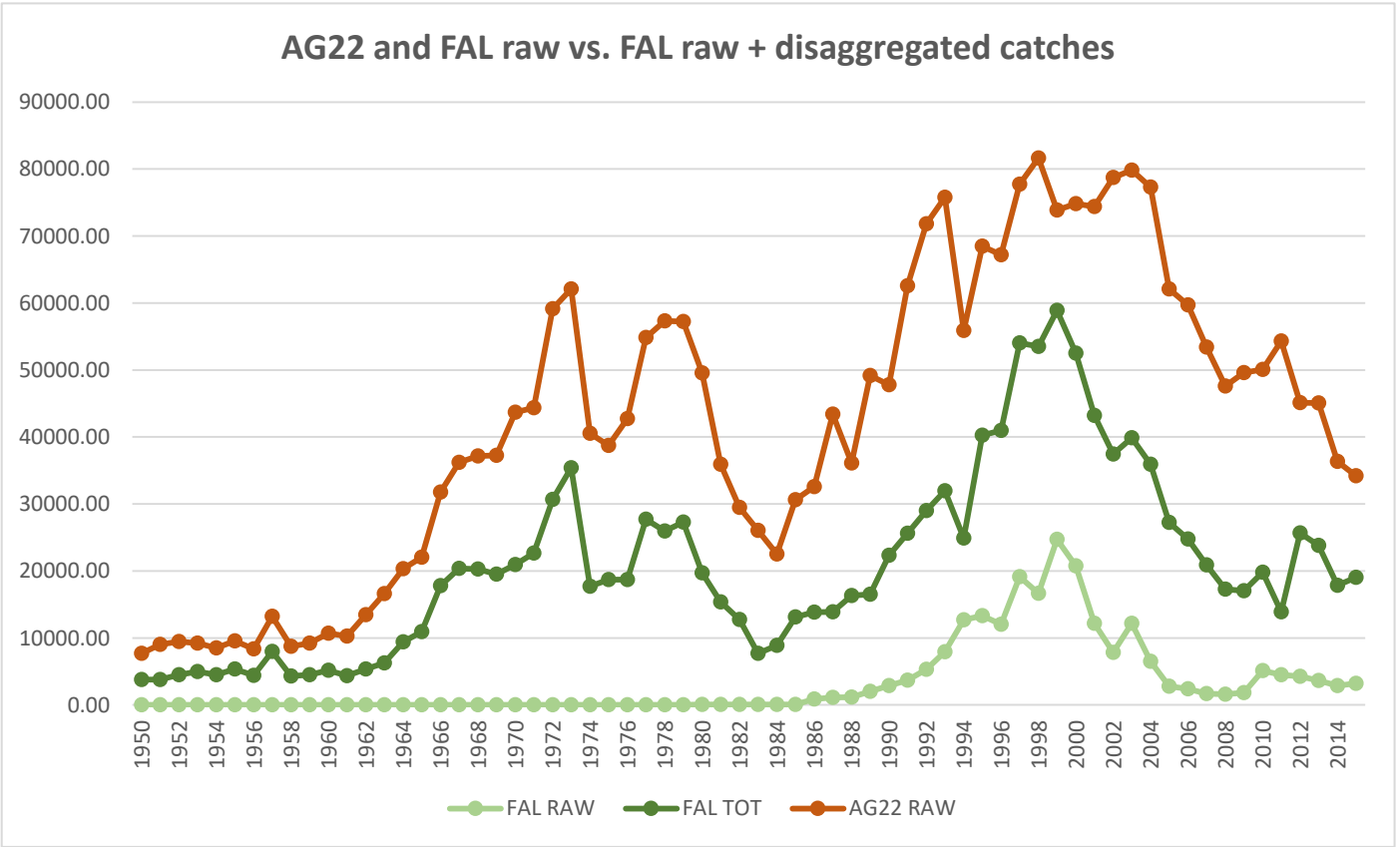

Figure A2.6. AG22 and FAL raw vs. FAL raw + disaggregated catches

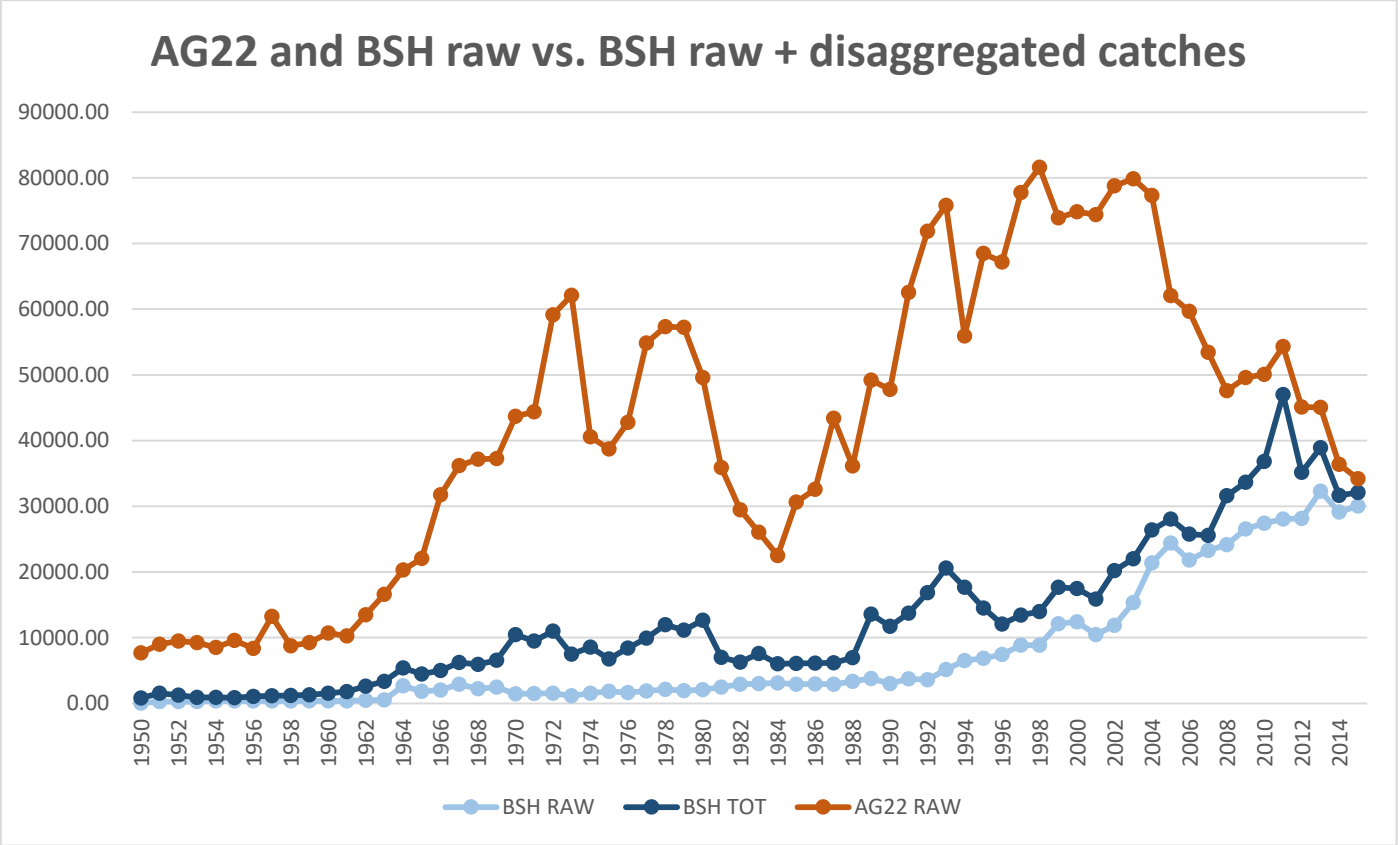

Figure A2.7. AG22 and BSH raw vs. BSH raw + disaggregated catches

Scientists seem to agree that most of the **AG22** catches from the '50s onwards are indeed including mostly catches for **BSH** and **FAL** under the aggregate. For this reason, it would be interesting to apply the disaggregation procedure to aggregated **AG22** catches and evaluate the produced results.

As can be seen in Figure A2.6 and A2.7, most of the **AG22** catches, especially in the '60s and '70s, are assigned to **FAL** (Silky sharks) with **BSH** getting far minor contributions over the entire time series.

These results are explained by the fact that in the IOTC database, most of the **AG22 – Sharks NEI** catches are recorded under *Artisanal* or *Semi-Industrial* gears. Between **FAL - Silky Sharks** and **BSH - Blue Sharks**, is the first that is mostly caught (as can be seen from the raw Nominal Catch data) by means of *Semi-Industrial* gears, whereas the latter is mostly caught with *Artisanal* and *Industrial* gears.

Therefore, based on the way in which the Nominal Catch Disaggregation process works (and based on the currently available disaggregation procedures), the reconstructed time series are perfectly in line with the available data.

It can be easily verified how the process does provide quite a substantially different outcome than either applying a constant factor to convert from **AG22** to **FAL** / **BSH** catches or by applying a moving average that reconstructs catches using the proportions available for both species in the closest *n* years.

References

- IOTC-2016-WPDCS12-26_Rev1: *Data as resources: how to enhance data sharing capabilities between the Secretariat and the scientific community* – IOTC Secretariat (2016), 12th Session of the Working Party on Data Collection and Statistics, Victoria, Seychelles
- REpresentational State Transfer (REST) – R. Thomas (2000), *Architectural Styles and the Design of Network-based Software Architectures* – Chapter 5 (Ph.D.). University of California, Irvine