Improving the core IOTC data management processes

PREPARED BY: IOTC SECRETARIAT[1], 18TH MAY 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 data-poor stocks (Neritic tunas) is greatly reduced with respect to the current processes. For more complex stocks (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 team of the Commission) to be able to respond to data updates in a quick and convenient way, 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 look like (and mostly behave) as an integrated system, this approach was not completely pursued since the beginning, 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 than they should have been.

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.

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 produce the outputs it was designed for, at the same time struggling to remain easy to maintain and extend.

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

This redesign process – very demanding in terms of efforts – is still ongoing, although it has 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, indeed, are 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 wherever they are, as long as they have network connectivity and a web browser available. 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.

Current state of the art

Data are reported to the Secretariat (from each CPCs) at different times of the year, depending on the type of datasets, or are collected from other sources (published data, national institutions etc.) based on their availability.

Data came in multiple formats, most of the times as custom Excel files that need to be transformed into the target format (still an Excel file, see the previous data submission forms) 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 do provide contrasting information.

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

Finalized data can serve multiple purposes:

- It can be made available through the Online Query Panel (confidential records are published only 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:

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

From a technological point of view, all the above processes are implemented as a mix of different solutions including multiple Relational Databases (Microsoft SQL Server and MySQL Server), Access databases with embedded forms, macros and VBA scripts, Excel templates for the production of summaries and the dissemination of dataset-specific records, R scripts for the production of graphs and reports.

All the components of this technology stack are linked together by means of shared data sources (ODBC sources that should be defined on each and every desktop PC that is running any of the above steps) and shared files.

Furthermore, they require that users performing any of these operations 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 have:

- 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.

As anticipated, the main IOTC database is responsible for the storage of the 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.

Furthermore, none of the management steps described above does implement any access control logic: any user that has physical access to the involved databases has potentially the rights to perform any kind of operation on them. Also, for this reason, destructive operations on the data (be it pinpointed record modifications or bulk upload of reported data, as well as the production of synthesized data following the execution of any of the processes above) are not accounted, and it is practically impossible to identify who did what on a specific dataset.

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.

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.

Nevertheless, the 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:

- 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[2] 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 (users have limited capabilities: they can perform only the operations they're entitled to);
- Inclusion of geospatial features as natively provided by the RDBMS of choice;
- Data extraction could be made available to external users as well (as long as the data is public: otherwise, specific credentials could be granted to scientists / data analysts);
- 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 incorporated 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
The initial step for the successful implementation of the new data management processes was 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 'AG') 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 RDBMS, it was now possible to incorporate the proper geographical boundaries for each area,

---

[2] REST stands for *REpresentational State Transfer*, which is a coordinated set of components, connectors, and data elements within a distributed hypermedia system – See: https://en.wikipedia.org/wiki/Representational_state_transfer

including most of the irregular ones. As a consequence the overlapping fraction between two different fishing grounds can be calculated and used to proportionally distribute quantities when reallocating data sets (Reported Catch and Efforts / Reported Size-Frequencies).

The notion of '*alternative effort*' was also 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 originally adopted to be able to tell which of the different effort units should be considered as primary.

This was particularly evident when producing the reallocated catch-and-effort dataset, as efforts not referring to the main unit (by gear) were simply discarded. Now, with this new approach, no effort record is discarded anymore (as long as it refers to either the primary or the alternative effort for a given gear) and this might lead to changes in the produced effort values within the disseminated datasets.

Data aliasing

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

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

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

for different types of barracudas that now, under the new '*Species*' codelist, are all equivalent to the aggregate:

- **AG32** – *Barracudas*

This means that it will not be always possible to perfectly match records when comparing datasets produced by the current process with the same datasets produced by the new one.

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

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

that 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*

Features overview

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

As anticipated, the application is accessible as a remote web application: therefore, all that is needed is a web browser (we recommend using recent, W3C 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 needed to avoid excessive scrolling through the application pages.
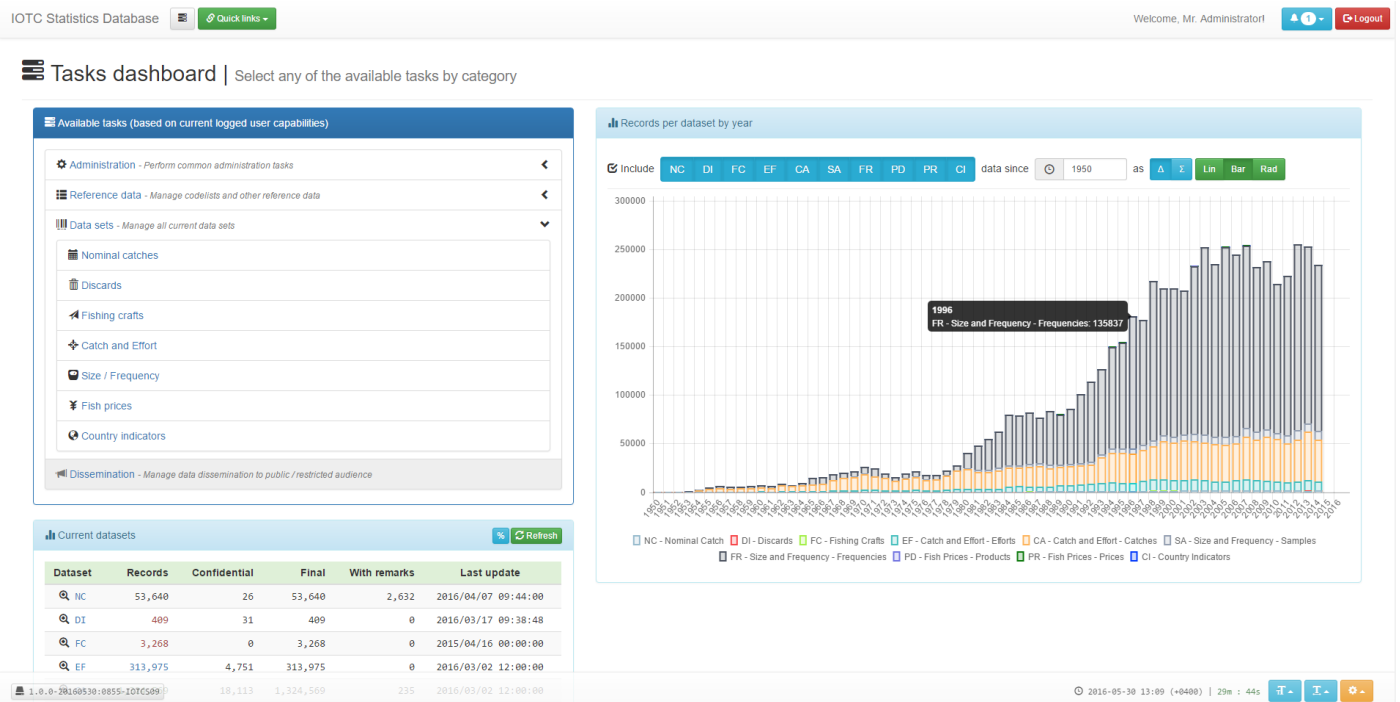
Dashboard

Figure 1. The overall dashboard as available to an administrative user

Figure 1 shows how the dashboard looks like for an administrative user (all grants enabled). On the top left section, you can see the list of available tasks grouped by category. To its right, a graph showing the size (in terms of number of records) of each dataset is also available. It can display data as a line chart, a bar graph (shown) or as a radar graph. Data can be filtered by dataset and by starting year.

The overall dataset details are also displayed in the bottom-left panel, reporting – by dataset – the total number of records alongside the number of records marked as final or confidential and the number of records with remarks set. The date of last update is also shown.

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

Figure 2 and 3 show the expanded details of the tasks list as available for an administrative user (Figure 2) and for a Nominal Catch and Reference Data manager user (Figure 3).

As you can see, the list of tasks available to the non-administrator user is limited to Reference Data and Nominal Catch management. This will prevent such user from performing administrative operations and changes to other data sets he / she is not directly responsible for.
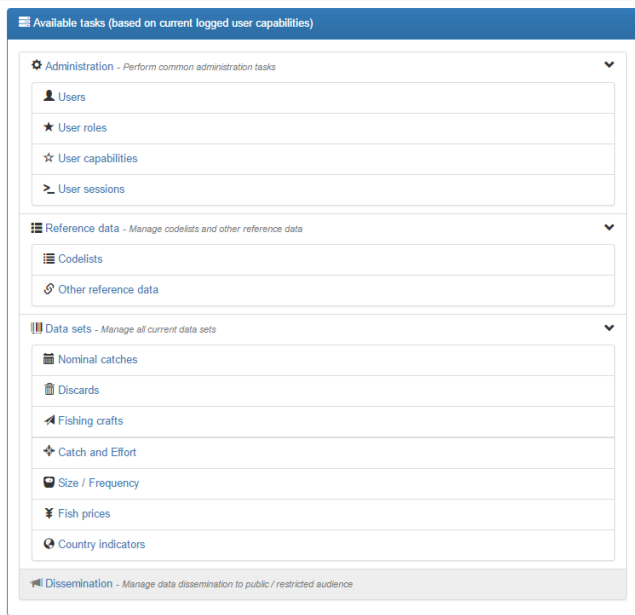
Figure 2. Tasks dashboard (admin user)
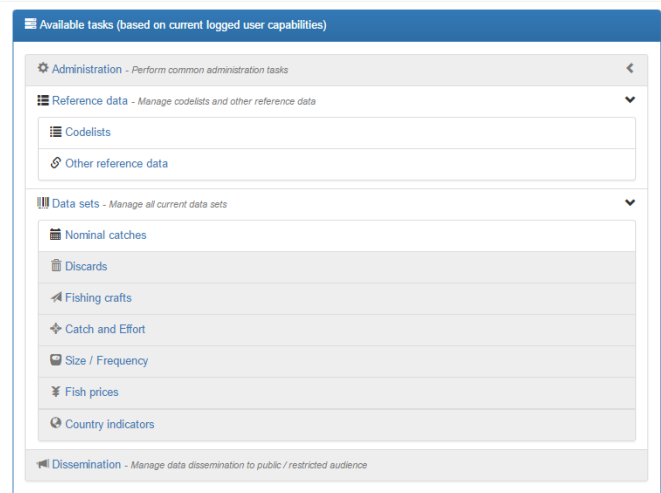
Figure 3. Tasks dashboard (NC manager user)

The overall list of available tasks is the following:

- **Administration**
  - **Users** – list / create / update / enable / disable system users and change their roles
  - **User roles** – list / create / update / delete user roles
  - **User capabilities** – list / create / update / delete user capabilities
  - **User sessions** – list / terminate user sessions
- **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
  - **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
  - **Fish prices –** manage fish prices records / upload fish prices data
  - **Country Indicators –** manage country indicator records / upload fish prices data

Actual available tasks will differ – by user – based on his / her grants.

User management

Within this section (accessible only to users with the proper administrative grants), new users, roles and capabilities can be created. It is also possible to change capabilities by role and update user grants and details.

Figure 4. Listing the available system users with their roles / capabilities

Figure 5. Listing and editing roles for a specific user

Figures 4 and 5 show the current interface to list and edit system users, as well as editing a specific user's roleset. (these operations are only available to administrative users).

Codelists and other reference data management

These tasks will allow entitled users to perform changes to existing codelists or other types of reference data (including mappings, aggregations and configuration tables). They're two of the crucial operations that can be performed within the system, as modifications to either codelists or reference data will 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).

The major difference between codelists and 'other' reference data is that the first are referenced throughout all the 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.

Figure 6. Listing all the available codelists

Figure 7. Listing the content of a specific codelist (the 'Species' codelist, in this example)

Figure 8. Editing a codelist entry

The codelist and other reference data management tasks allow changing (updating / deleting) entries from within a codelist as well as creating new entries.

When updating or creating a new codelist entry, the system will actively support the user into ensuring that all mandatory fields are set, and will provide 'search as you type' facilities to identify related reference data (see the 'Species official' field in Figure 8) that the codelist might include.

Dataset management

The dataset management tasks encompass all the operations that might be performed on the type of data currently managed by the system. Functionalities common to all datasets do include:

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

Furthermore – depending on the type of dataset – some additional operations are also available.

Figure 9 and 10 show the interfaces used to filter the content of a given dataset and edit any of its record. They're consistent across all dataset, therefore we will limit to present their appearance only for Nominal Catches.

Figure 9. Filtering the content of a dataset (Nominal Catches)

Figure 10. Editing a dataset record (Nominal Catches)

As was the case when editing a codelist entry, the system supports users into ensuring – when editing or creating a new dataset record – that all mandatory fields are set, and will present filterable dropdown list to allow selection of any referenced codelist (when required by the specific dataset).

Uploading new data 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[3].

Although the adoption of these forms is not mandatory for CPCs, we strongly encourage that all data is submitted to the Secretariat by means of any of these. This will ensure that the sanity checks and error fixing procedures available within the system could be effectively used to perform a first quality assessment of the uploaded data as soon as they are received.

Figure 11 shows the user interface presented to users after a valid XLS file has been uploaded to the system. At this stage, the data assistant is supposed to select a proper source and quality level for the uploaded data, as well as the confidentiality level to assign to the uploaded records.

The system will automatically display the metadata as available within the uploaded file (submitter name / organization / contact details etc.) and will also present a preview of the uploaded data (the first ten rows).

---

[3] http://www.iotc.org/data/requested-statistics-and-submission-forms
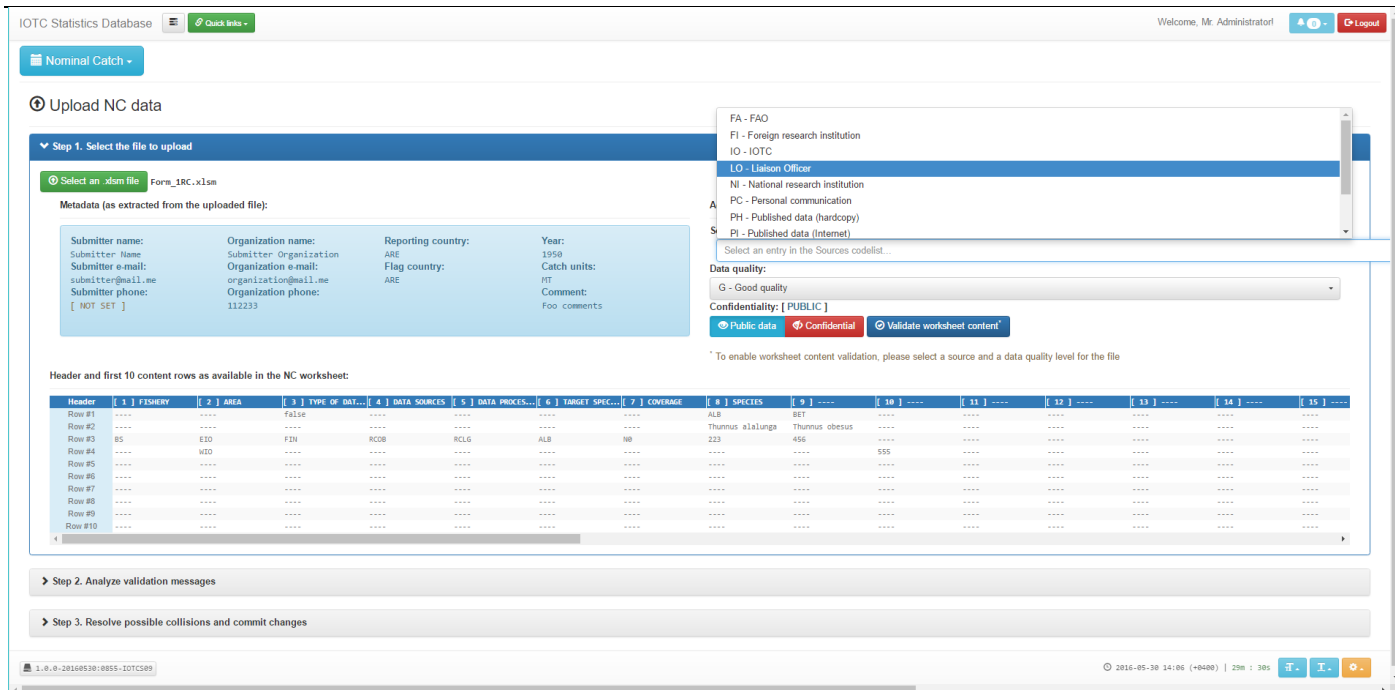
Figure 11. Adding metadata and showing a preview of the uploaded data

Once the 'source' and 'quality' fields have been filled, the data are ready to be validated.

The currently applied validation rules will check that all mandatory data are provided, that there is no gap in the records and that all codelist references are correctly set.

If any record is breaking one of these rules, the upload process stops and presents the user with the list of faults, some of which might require updating the content of the involved codelists (e.g. fishing grounds) while other will require updating the actual file sent for upload.

Figure 12 shows an example of the error reports received once data is validated.

In this specific case the system has identified that within the uploaded form there are five fields that were supposed to be mandatory and that have not been filled. It's up to the user deciding whether these can be properly filled or it's the case of getting back to the original data submitter and request for clarification.
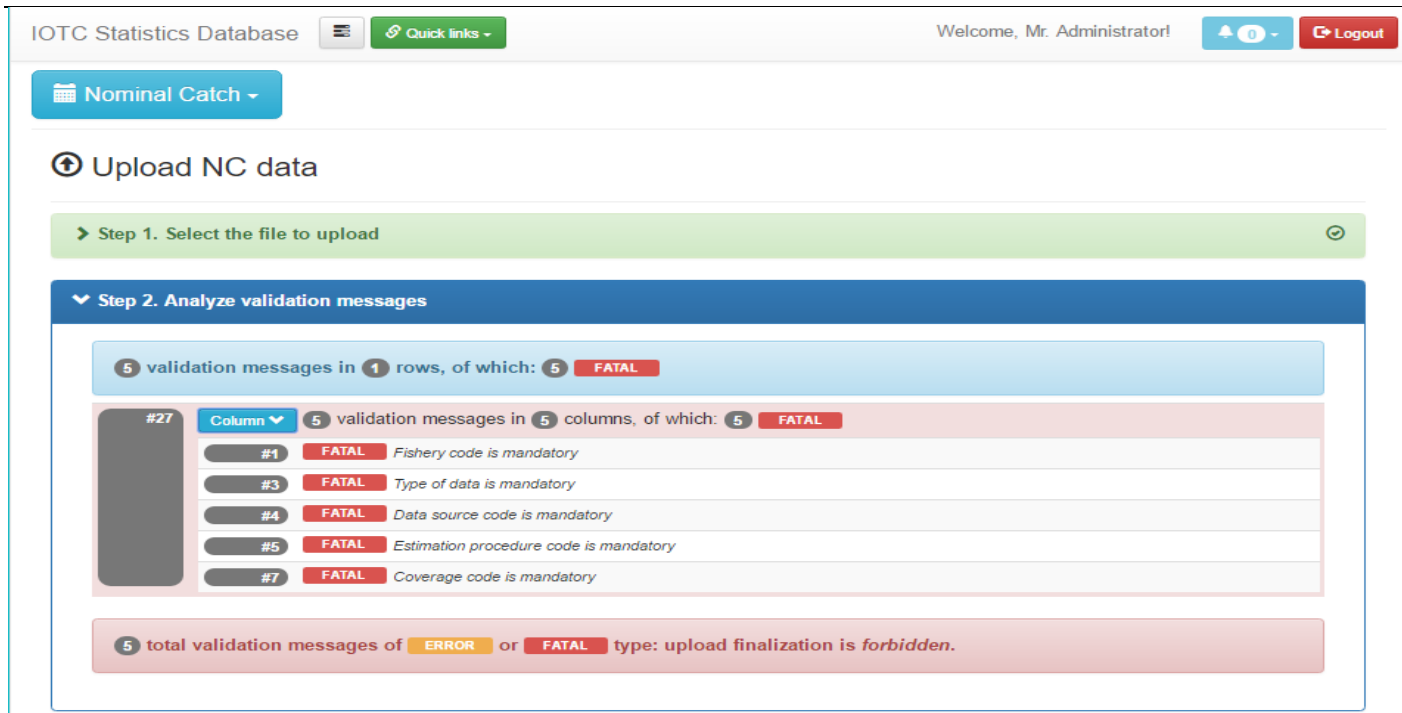
Figure 12. Sanity check results for the uploaded data

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: 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.

Figure 13 shows an example of the collision reports received once data is successfully validated. In this case, there has been an update on the catch quantity and confidentiality status of an existing record. The user can decide whether this updates should be committed or not within the system.
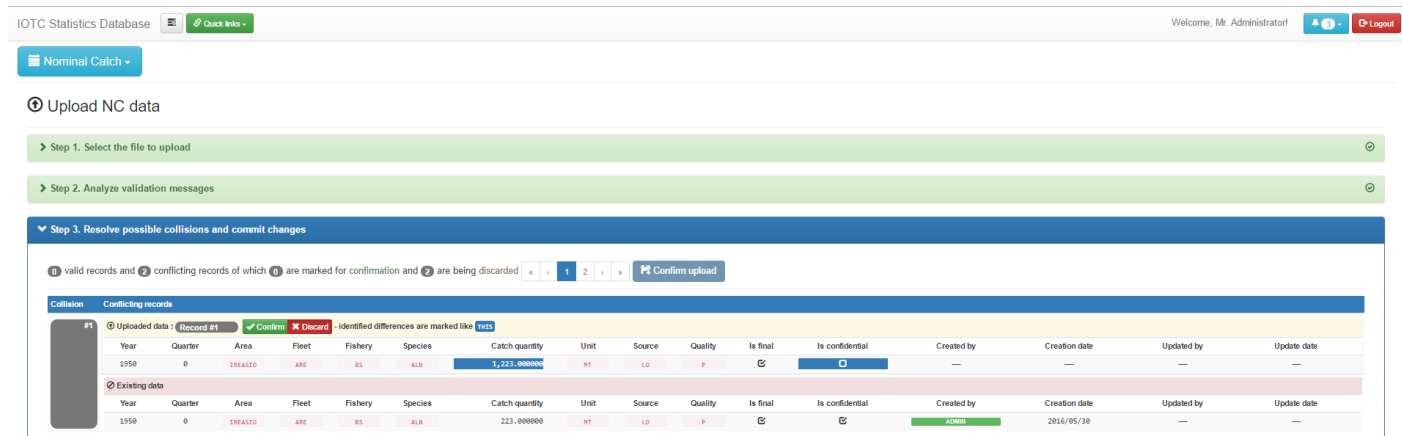
Figure 13. Resolving data collisions prior to upload finalization

Nominal catches

Beside the common management processes already described for all the datasets, entitled users are enabled to perform additional sets of operations on the Nominal Catch 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 nominal catch records) as well as produce reports showing trends and composition of the current records within the dataset.

Additionally, the system can also produce Nominal Catch disaggregation upon request, as a preliminary step for the generation of the datasets required for the stock assessment of all involved species.

An example of the type of reports that can be generated for Nominal Catches is shown in Figure 14.
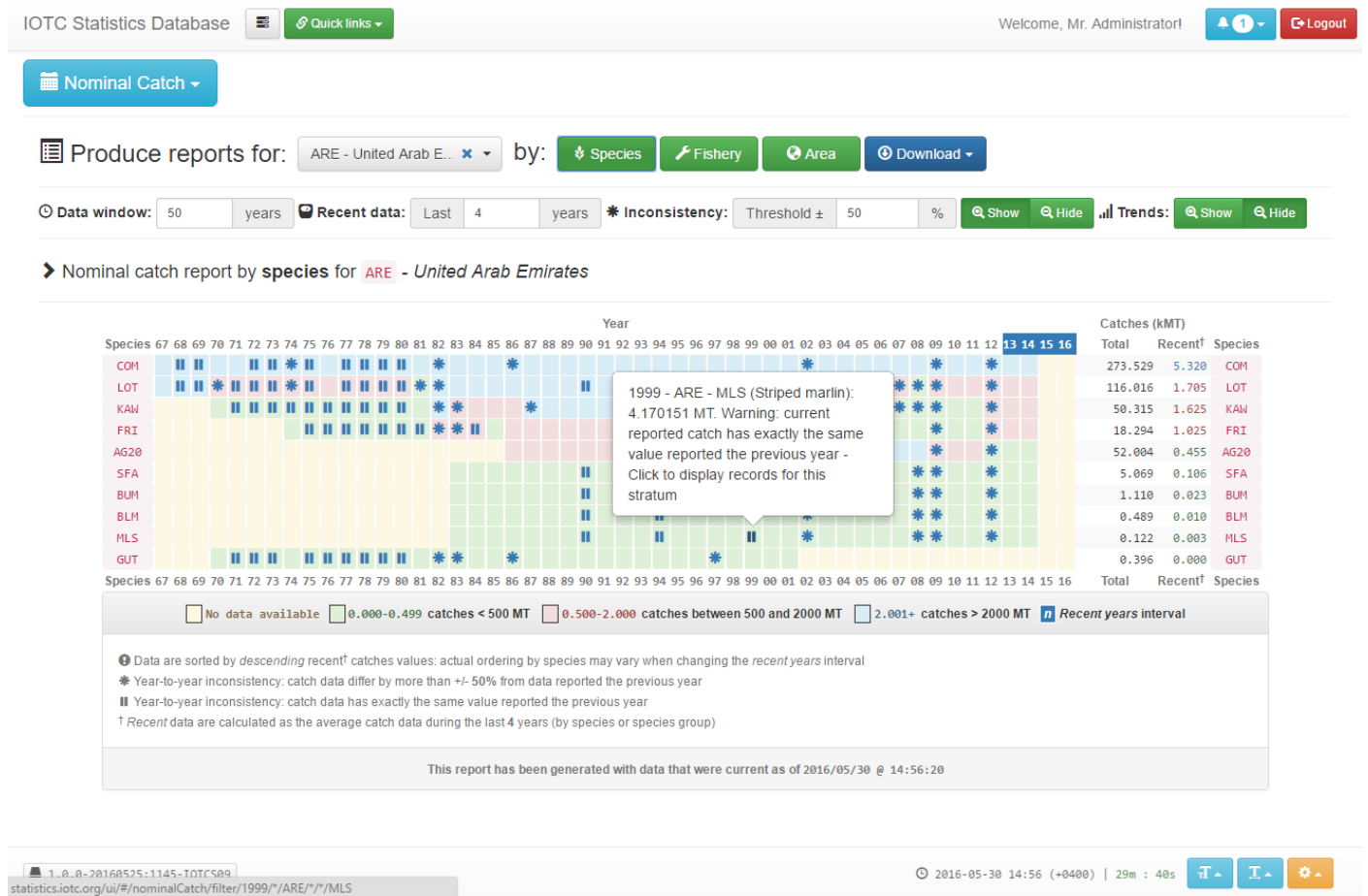
Figure 14. Nominal Catch reports by fleet and species

These type of reports will show catch amounts (for a given fleet) during a specific user selectable timeframe. The cell colors tell the range of catches for each species / year, whereas the icons superimposed to the cell will help identifying relevant discrepancies between data reported between two consecutive years as well as repetition patterns of some kind. In Figure 14, as an example, is shown that catches of Frigate tuna for United Arab Emirates were constantly reported with the same value (100 MT) from 1975 to 1981.

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 images or as raw CSV data for later analysis.

Another type of summary report currently available for Nominal Catches is shown in Figure 15.
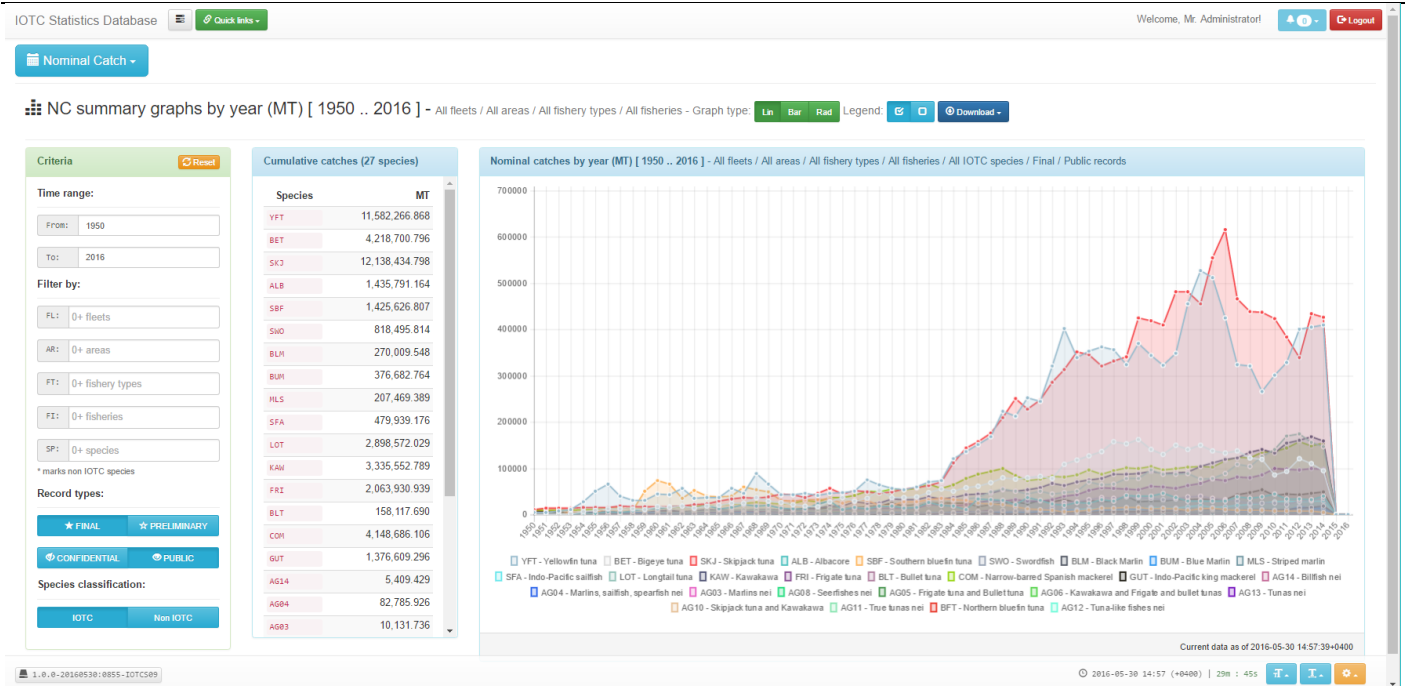
Figure 15. Nominal Catch summary reports and graphs

This type of summary report is providing 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, a cumulative bar chart or a radar chart.
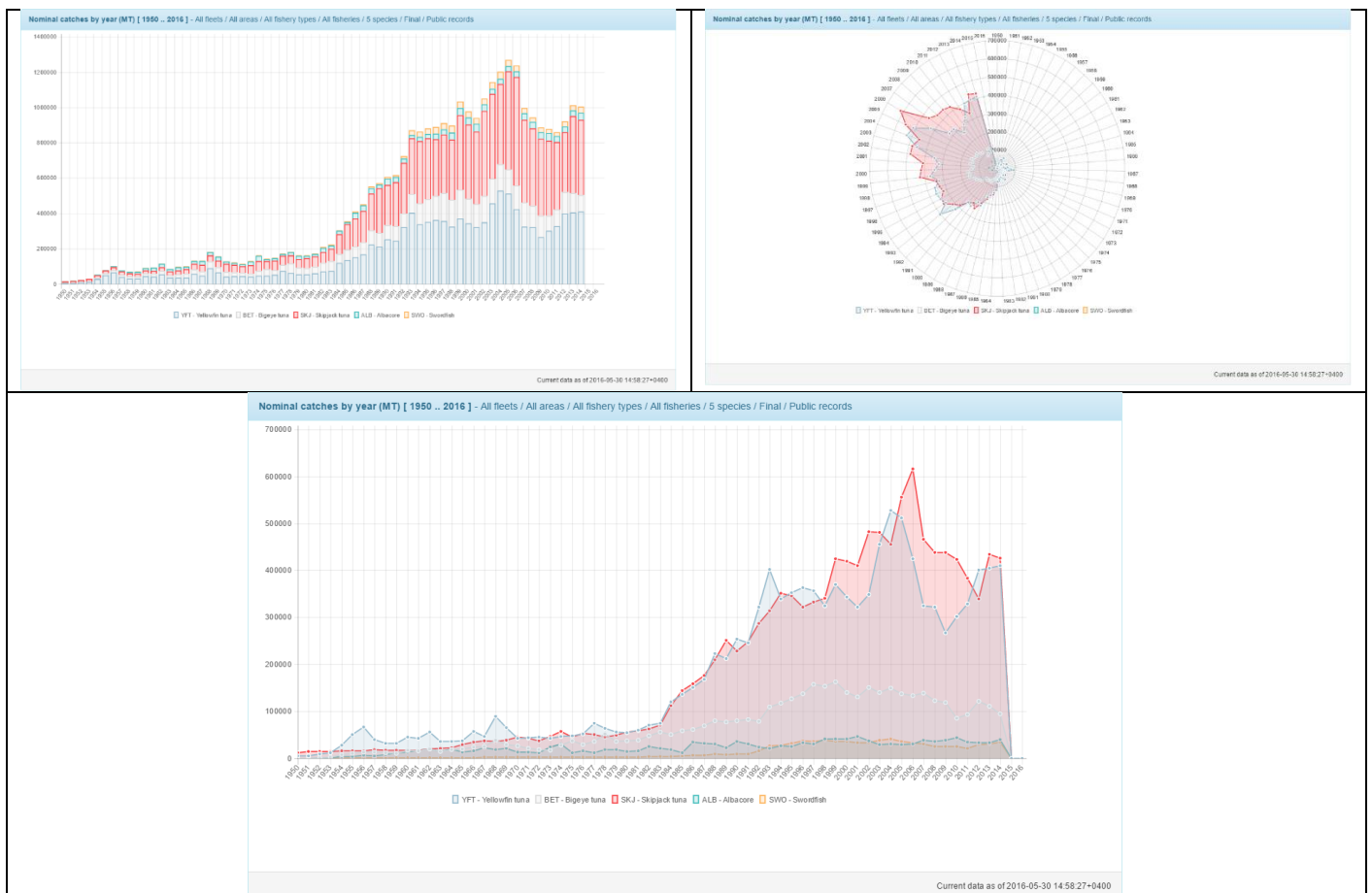
Figure 16. The different types of graphs that can be generated for Nominal Catch summary reports

The selected data subset can then be exported as a CSV file or graphs downloaded as images.

The last Nominal Catch specific feature is related to the Nominal Catch Disaggregation process. So far, within the legacy system, it is implemented through a specific Access database that takes care of linking to the existing dataset and apply the envisaged disaggregation procedures to produce a dataset which does not refer to neither gear nor species aggregates.

The current implementation of the Nominal Catch process is accessible through the user interface in Figure 17.

Figure 17. The Nominal Catch Disaggregation process console

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. Furthermore, disaggregation results are stored within the system and can be 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, as it should be most of the times), the system will process the records and – based on the current disaggregation procedures – will provide as output the overall results of its execution.

Under some circumstances, it might happen that some records cannot be automatically disaggregated: in that case, the user is supposed to provide his / her estimates about how to assign catches by single gear / species.

Figure 18. Disaggregation results (showing the number of records disaggregated by procedure)

Figure 18 shows an example of the preliminary report produced by the disaggregation process. In particular, we see that of the original 338 Nominal Catch records to be disaggregated (limited only to Australia) the system was able to produce 848 distinct (non-aggregated) new records by applying any of the 8 currently available procedures (see appendix A.2 to this document).

These were still not able to disaggregate 20 original records, for which manual intervention is required.

The reason why still some record need to be disaggregated by hand is that most of the current procedures are looking at different time intervals and proxy fleets or gears (see appendix A.1 to this document) 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.
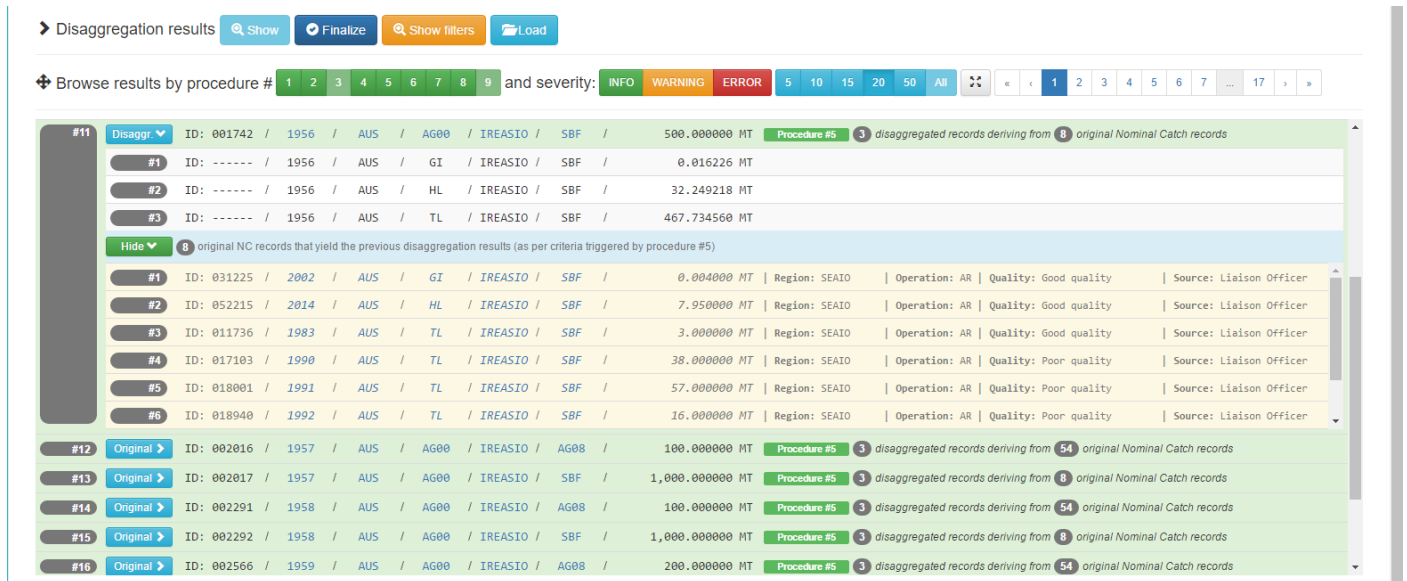
Figure 19. Disaggregation outputs

Figure 19 shows a subset of the disaggregation outputs for results related to the above mentioned example (disaggregation of Australian Nominal Catches only). In particular, we can see the three disaggregated records that will be used to replace the single aggregated record referring to 1956 / AUS / AG00 (Unidentified gear) / IREASIO (Eastern Indian Ocean) / SBF (Southern Bluefin tuna).

This record – that was originally referring to an aggregation of gears - has been broken down to three non-aggregated records by using eight other Nominal Catch records identified by procedure #5 (Same fleet, same type of operation, same region, same IOTC area, any fishing year).

Thanks to these additional eight records, the system was able to split the original catches (500 MT) into three different quantities related to the same strata and to one among Gillnet, Handline and Trolling as gears. The original catches are proportionally assigned based on how catches for the same gear contribute to the total within the subset of proxy records identified by the triggered disaggregation procedure.
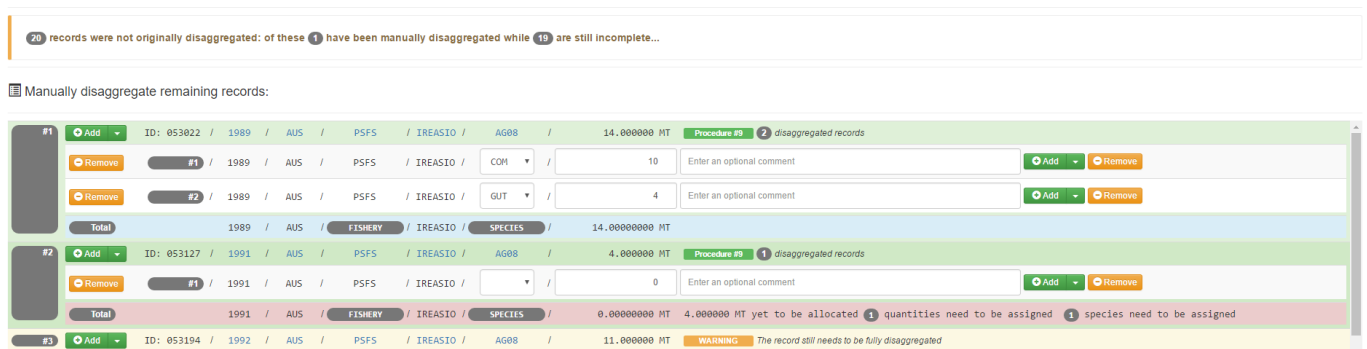
Figure 20. Manual disaggregation of Nominal Catch records

Figure 20 shows how the system can support users into providing disaggregation estimates for records that could not be automatically identified.

Depending on the record referring to either a gear or a species aggregation (even both), users will be presented with the original record and the required user interface controls to add specific disaggregation values for any of the gear or species involved.

The system will also check that the values inserted by the user for all the manually disaggregated records do not exceed the original catch value. At the same time, the disaggregation can be finalized (that is, stored within the system) only when all the records requiring manual disaggregation are updated correctly.

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 and one fishery type (Longlines, Purse seines and bait boats, Other coastal gears) and then decide the level of reallocation that they want to get.

In particular, data can be reallocated (by any of the species identified as relevant for the selected fishery type) on their original fishing ground (as reported) as well as on 1x1 or 5x5 grids. When selecting any of the two regular grid resolution, 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.
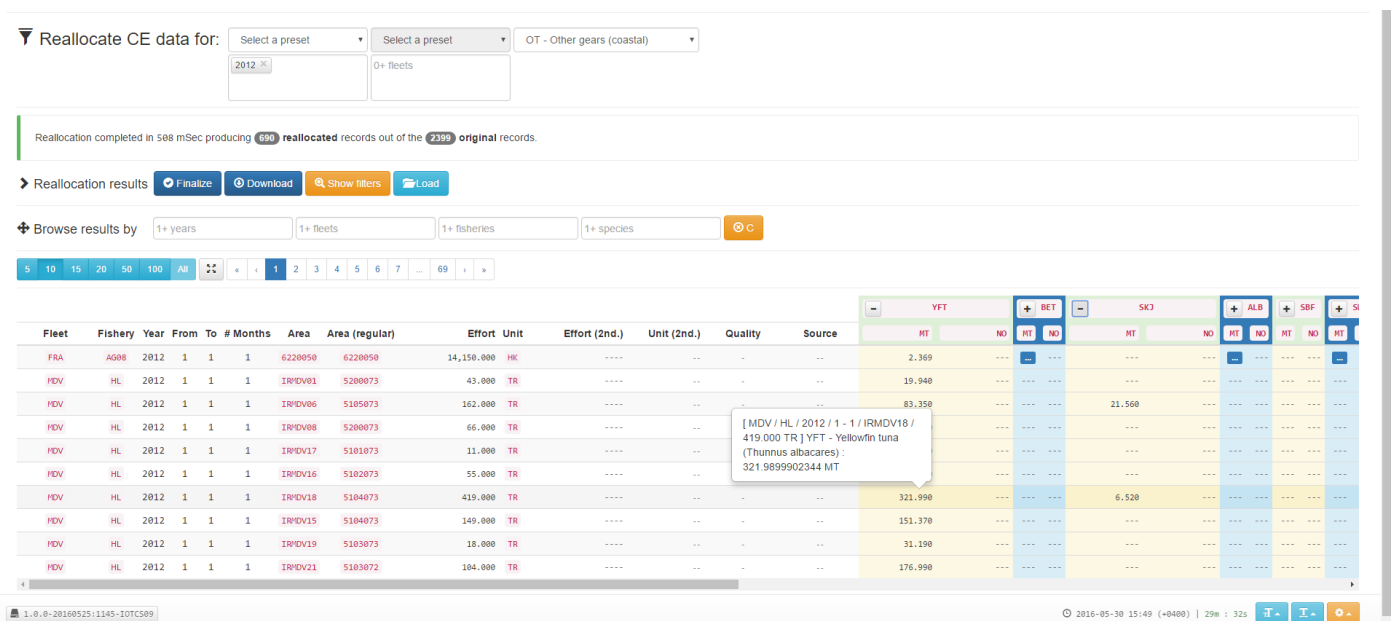
Figure 21. Catch and Effort reallocation results (Other gears – coastal, 2012)

Another set of additional features available for the Catch and Effort dataset is capable of producing geospatial representation of reported catches / efforts / CPUE values for any given subset of the data.

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

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.

Figure 22 shows an example of the geographical distribution of YFT catches in 2014, with grids having a resolution of 1x1 degrees. In this specific figure, reported catches are plotted on a logarithmic scale and limited to final and public (i.e. non-confidential) records.

This feature (also available for reported efforts and CPUE) is extremely useful to identify seasonality and common patterns within the reported data. In particular, Figures 23 a-b-c shows the evolution of reported Spanish YFT catches during January / April and October (totals for each year).
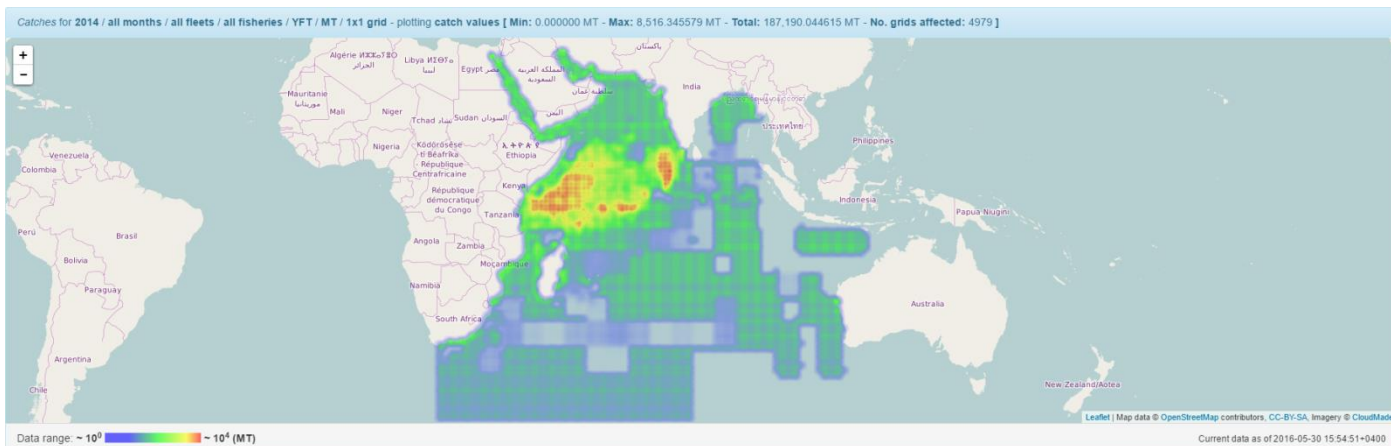
Figure 22. YFT reported catches for 2014 (MT, logarithmic scale over 1x1 degree grids)
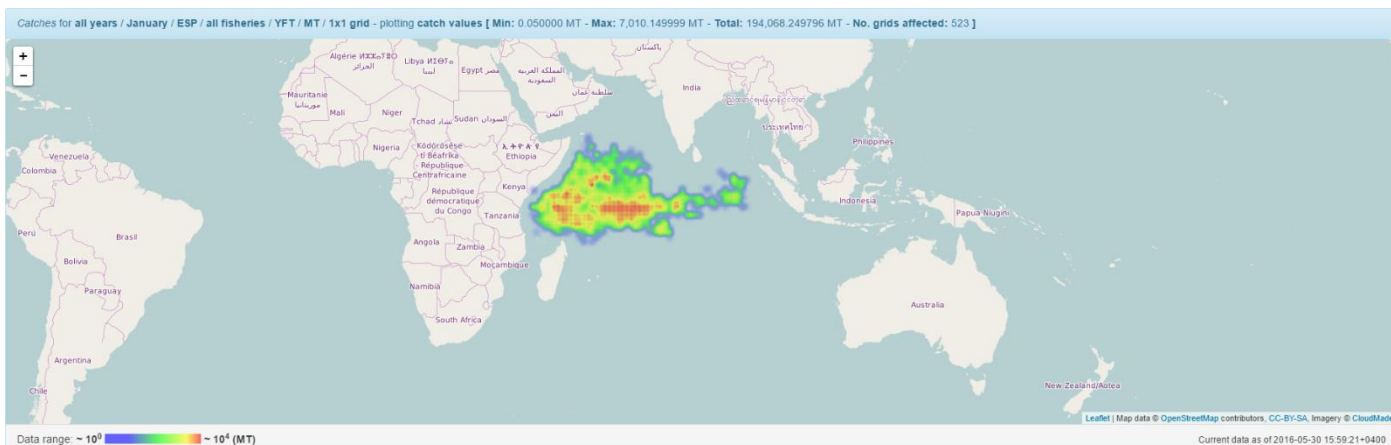
Figure 23a. YFT reported catches for ESP (MT, logarithmic scale over 1x1 degree grids) during January
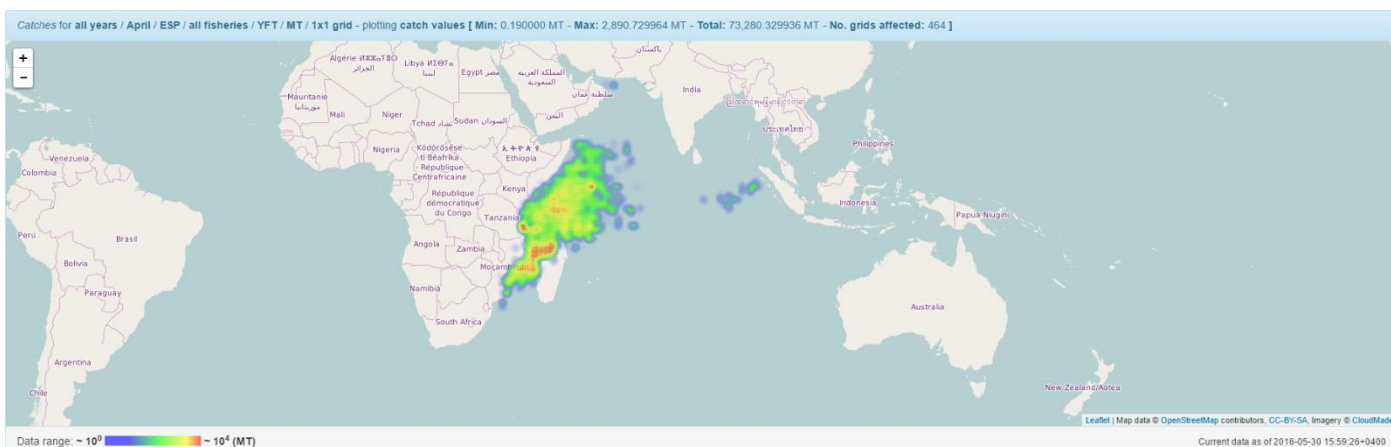
Figure 23b. YFT reported catches for ESP (MT, logarithmic scale over 1x1 degree grids) during April
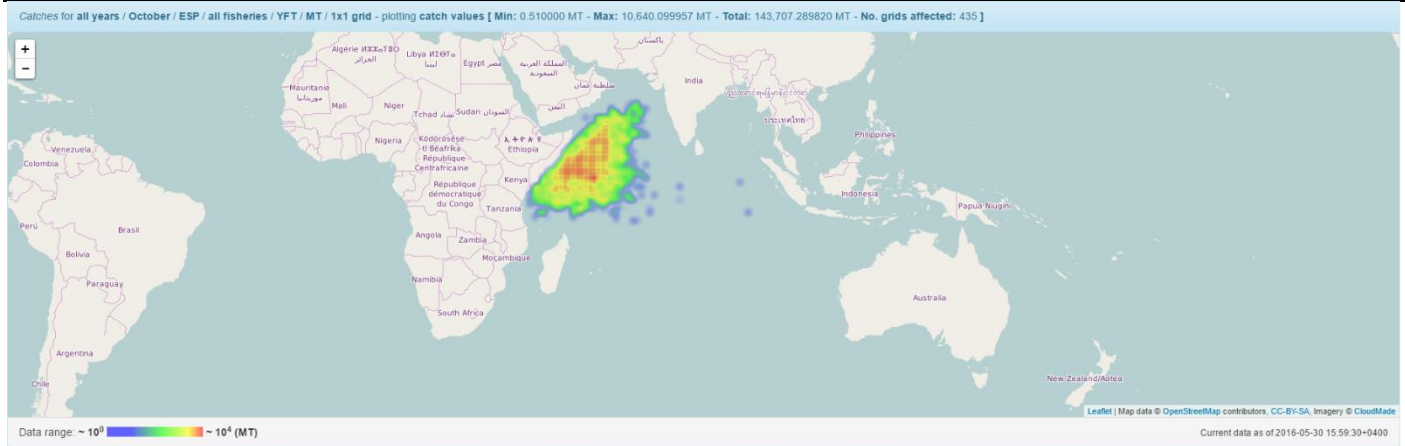
Figure 23c. YFT reported catches for ESP (MT, logarithmic scale over 1x1 degree grids) during October

These heatmaps can be produced in almost real time: it is extremely interesting to use this feature to actually 'animate' the reported catch / effort distributions as they evolve during different period of the year or across years.

Data can be downloaded as images or CSV files.

Size-frequency

Beside the common management processes already described for all the datasets, entitled users are enabled to 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 groups (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.

Additionally, users can identify strata for which (due to the reported length units for a given species) there is no length – weight conversion equation available. Also, the number of fishes per size bin can be proportionally colored in order to show where the highest concentration of samples is for each specific strata.

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

Users can filter the data by year, month, fleet, fishery and species, and eventually produce heatmaps at different level of resolution.

Also, data can be plotted either by number of raw samples or by reported sample numbers. Furthermore, it is possible to limit the geospatial plot to any custom area by simply providing the vertex coordinates in a WKT-like format.

Figure 25a shows an example of the geographical distribution of SKJ reported raw samples for any available year, with grids having a resolution of 1x1 degrees. In this specific figure, reported raw samples are plotted on a logarithmic scale and limited to final and public (i.e. non-confidential) records.

Figure 25b shows an example of the geographical distribution of SKJ reported sampled numbers for any available year, with grids having a resolution of 1x1 degrees. In this specific figure, reported sampled numbers are plotted on a logarithmic scale and limited to final and public (i.e. non-confidential) records.

This feature, exactly as anticipated for Catch and Effort records, could be extremely useful to identify seasonality and common patterns within the reported data.
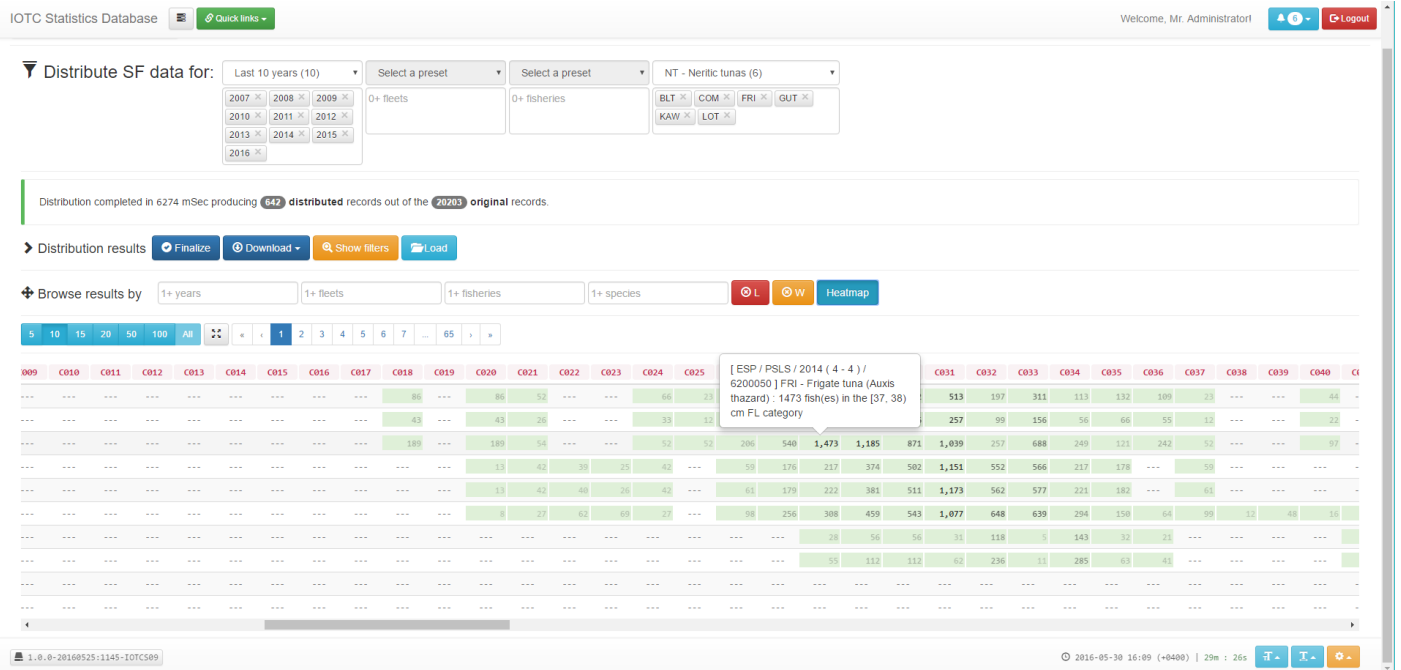
Figure 24. Neritic Tunas reported length distribution for 2007-2016
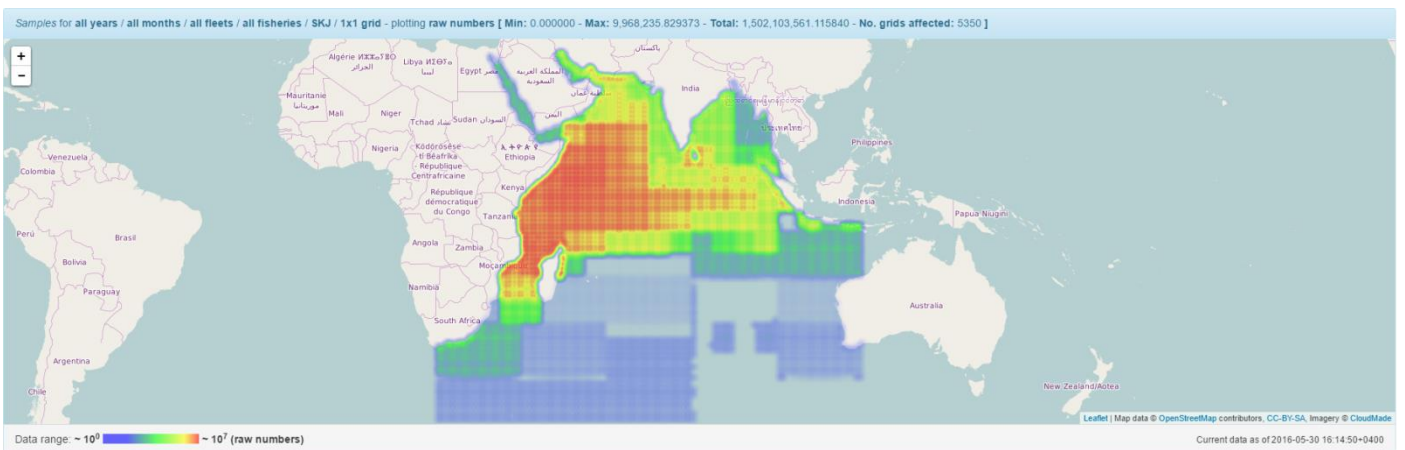
Figure 25a. Reported raw samples distribution for SKJ (logarithmic scale over 1x1 degree grids)
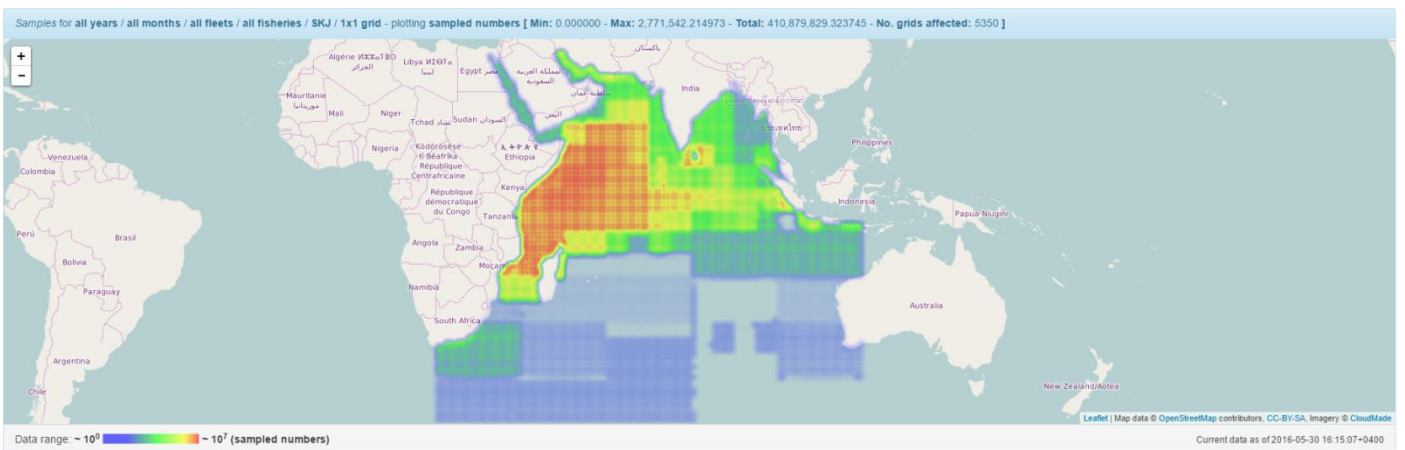
Figure 25b. Reported sampled numbers distribution for SKJ (logarithmic scale over 1x1 degree grids)

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.

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

Of course, the Secretariat will ensure (and this pre-requisite is already perfectly satisfied by the key concepts behind the access control mechanism in place within the new system) that only data that can be safely disseminated is therefore made publicly available.

Also, the possibility of sharing data and processes in a formal and automated way (by means of the described REST services) will open up endless integration possibilities for scientists worldwide.

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 should be used to break down aggregated ones. The current disaggregation procedures are looking at any number of records from proxy fleet / gears and the results – when available – is used to proportionally assign catches for aggregated records.

Currently, we do not set a 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.

Also, the geospatial plots of the reported catches can be extended to display similar results with respect to *raised catches* (and not only *reported* catches) produced as a preliminary artifact prior to the preparation of stock assessment files (currently, this is still a work in progress). Once available, the results of such geospatial plots extended to raised catches will be extremely useful to assess and fine tune the catch raising process itself.

Accessing the application

The latest implementation of the Integrated Data Management console is currently available at the following URL:

http://statistics.iotc.org

Users can get read only, public access to the console by providing GUEST as both username and password.

The data made available to guest users through the application is just a snapshot of the current legacy database (as of June 1st 2016) in read-only mode and limited to public and final records, in order to fulfill the recommendation set forth by resolution 12/02.

We suggest using a fairly recent version of either Google Chrome™ or Mozilla Firefox™ to access the application, with an optimal screen resolution of 1920 x 1080 pixels (this because of the high density of information that can be found in most of the pages).

Furthermore, if not just because of the introduced changes in species, fleets and gears classifications (as well as in gear / species aggregations) we advise against using the data that might be extracted from the application for any scientific or official purpose.

Although the content of the new database is a 99.9% accurate representation of the content of the current database, we cannot guarantee that the data and the process' results might be fit – at this stage – for purposes other than just testing and evaluating the overall behaviour of the integrated system.

Appendix A

Appendix A.1 – Example of proxy fleet / gears currently used for nominal catch disaggregation

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
IDN	IDN	BS	IREASIO	EASIO	ART
IDN	IDN	CN	IREASIO	EASIO	ART

IDN	IDN	DDSEI	IREASIO	EASIO	ART
IDN	IDN	HOOK	IREASIO	EASIO	ART
IDN	IDN	RIN	IREASIO	EASIO	ART
IDN	IDN	TRAW	IREASIO	EASIO	ART
IDN	IDN	DSEI	IREASIO	INDON	ART
IDN	IDN	FN	IREASIO	INDON	ART
IDN	IDN	GHLI	IREASIO	INDON	ART
IDN	IDN	LIFT	IREASIO	INDON	ART
IDN	IDN	LLCO	IREASIO	INDON	ART
IDN	IDN	LLF	IREASIO	INDON	ART
IDN	IDN	PSS	IREASIO	INDON	ART
IDN	IDN	LL	IREASIO	EASIO	IND
IDN	IDN	BB	IREASIO	INDON	ART
IDN	IDN	GILL	IREASIO	INDON	ART
IDN	IDN	HAND	IREASIO	INDON	ART
IDN	IDN	LIGB	IREASIO	INDON	ART
IDN	IDN	TROL	IREASIO	INDON	ART
IDN	IDN	UNCL	IREASIO	INDON	ART
IDN	IDN	FLL	IREASIO	EASIO	IND
IDN	IOTC	LL	IREASIO	EASIO	IND
IDN	IOTC	LL	IRWESIO	EASIO	IND
IDN	KEN	FLL	IRWESIO	WESIO	IND
IDN	LKA	FLL	IREASIO	EASIO	IND
IDN	MDV	FLL	IRWESIO	WESIO	IND
IDN	MUS	FLL	IRWESIO	WESIO	IND
IDN	MYS	FLL	IREASIO	EASIO	IND
IDN	SGP	FLL	IREASIO	EASIO	IND
IDN	SYC	FLL	IRWESIO	SEYCH	IND
IDN	THA	FLL	IREASIO	EASIO	IND
IDN	YEM	FLL	IRWESIO	WESIO	IND
JPN	JPN	GILL	IREASIO	EASIO	IND
JPN	JPN	LL	IREASIO	EASIO	IND
JPN	JPN	PS	IREASIO	EASIO	IND
JPN	JPN	UNCL	IREASIO	EASIO	IND
JPN	JPN	LL	IRWESIO	WESIO	IND
JPN	JPN	PS	IRWESIO	WESIO	IND
KOR	KOR	PS	IREASIO	EASIO	IND
KOR	KOR	PS	IRWESIO	WESIO	IND
KOR	KOR	BB	IREASIO	EASIO	IND
KOR	KOR	LL	IREASIO	EASIO	IND
KOR	KOR	BB	IRWESIO	WESIO	IND
KOR	KOR	LL	IRWESIO	WESIO	IND
LKA	LKA	BS	IREASIO	SRILA	ART
LKA	LKA	GL	IREASIO	SRILA	ART
LKA	LKA	HATR	IREASIO	SRILA	ART
LKA	LKA	LG	IREASIO	SRILA	ART
LKA	LKA	LLCO	IREASIO	SRILA	ART
LKA	LKA	PS	IREASIO	EASIO	IND
LKA	LKA	FLL	IREASIO	SRILA	IND

LKA	LKA	G/L	IRWESIO	SRILA	ART
LKA	LKA	BB	IREASIO	MALDI	ART
LKA	LKA	G/L	IREASIO	SRILA	ART
LKA	LKA	GILL	IREASIO	SRILA	ART
LKA	LKA	GIOF	IREASIO	SRILA	ART
LKA	LKA	HAND	IREASIO	SRILA	ART
LKA	LKA	RIN	IREASIO	SRILA	ART
LKA	LKA	TROL	IREASIO	SRILA	ART
LKA	LKA	UNCL	IREASIO	SRILA	ART
MDV	MDV	LL	IREASIO	WESIO	IND
MDV	MDV	BB	IRWESIO	MALDI	ART
MDV	MDV	BBM	IRWESIO	MALDI	ART
MDV	MDV	BBN	IRWESIO	MALDI	ART
MDV	MDV	FN	IRWESIO	MALDI	ART
MDV	MDV	HAND	IRWESIO	MALDI	ART
MDV	MDV	LLCO	IRWESIO	MALDI	ART
MDV	MDV	TROL	IRWESIO	MALDI	ART
MDV	MDV	TROLM	IRWESIO	MALDI	ART
MDV	MDV	TROLN	IRWESIO	MALDI	ART
MDV	MDV	UNCL	IRWESIO	MALDI	ART
MDV	MDV	FLL	IRWESIO	WESIO	IND
MUS	MUS	LL	IREASIO	EASIO	IND
MUS	MUS	PS	IREASIO	WESIO	IND
MUS	MUS	HATR	IRWESIO	SWEIO	ART
MUS	MUS	SPOR	IRWESIO	SWEIO	ART
MUS	MUS	TROL	IRWESIO	SWEIO	ART
MUS	MUS	ELL	IRWESIO	SWEIO	IND
MUS	MUS	LL	IRWESIO	SWEIO	IND
MUS	MUS	PS	IRWESIO	WESIO	IND
NEI	BLZ	FLL	IREASIO	EASIO	IND
NEI	BLZ	FLL	IRWESIO	WESIO	IND
NEI	IDN	FLL	IREASIO	EASIO	IND
NEI	IND	LL	IREASIO	EASIO	IND
NEI	IND	LL	IRWESIO	WESIO	IND
NEI	IOTC	ELL	IREASIO	EASIO	IND
NEI	IOTC	FLL	IREASIO	EASIO	IND
NEI	IOTC	LL	IREASIO	EASIO	IND
NEI	IOTC	ELL	IRWESIO	WESIO	IND
NEI	IOTC	FLL	IRWESIO	WESIO	IND
NEI	IOTC	LL	IRWESIO	WESIO	IND
NEI	IPTP	LL	IRWESIO	WESIO	IND
NEI	IPTP	PS	IRWESIO	WESIO	IND
NEI	IRN	PS	IREASIO	EASIO	IND
NEI	IRN	PS	IRWESIO	WESIO	IND
NEI	SGP	FLL	IRWESIO	WESIO	IND
NEI	SGP	FLL	IREASIO	EASIO	IND
NEI	SUN	PS	IREASIO	EASIO	IND
NEI	SUN	PS	IRWESIO	WESIO	IND
NEI	THA	FLL	IREASIO	EASIO	IND

SUN	SUN	PSS	IRWESIO	REDSE	ART
SUN	SUN	LL	IREASIO	EASIO	IND
SUN	SUN	PS	IREASIO	EASIO	IND
SUN	SUN	GILL	IRWESIO	WESIO	IND
SUN	SUN	LL	IRWESIO	WESIO	IND
SUN	SUN	PS	IRWESIO	WESIO	IND
SUN	SUN	UNCL	IRWESIO	WESIO	IND
SYC	SYC	LL	IREASIO	EASIO	IND
SYC	SYC	PS	IREASIO	EASIO	IND
SYC	SYC	GILL	IRWESIO	SEYCH	ART
SYC	SYC	HAND	IRWESIO	SEYCH	ART
SYC	SYC	TROL	IRWESIO	SEYCH	ART
SYC	SYC	UNCL	IRWESIO	SEYCH	ART
SYC	SYC	ELL	IRWESIO	SEYCH	IND
SYC	SYC	LL	IRWESIO	WESIO	IND
SYC	SYC	PS	IRWESIO	WESIO	IND
THA	IOTC	LL	IREASIO	EASIO	IND
THA	IOTC	LL	IRWESIO	WESIO	IND
THA	SUN	PS	IREASIO	EASIO	IND
THA	SUN	PS	IRWESIO	WESIO	IND
THA	THA	PSS	IREASIO	ANDAS	ART
THA	THA	FLL	IREASIO	EASIO	IND
THA	THA	PS	IRWESIO	WESIO	IND
THA	THA	GILL	IREASIO	ANDAS	ART
THA	THA	HAND	IREASIO	ANDAS	ART
THA	THA	TRAW	IREASIO	ANDAS	ART
THA	THA	TROL	IREASIO	ANDAS	ART
THA	THA	UNCL	IREASIO	ANDAS	ART
THA	THA	LL	IREASIO	EASIO	IND
THA	THA	PS	IREASIO	EASIO	IND
THA	THA	LL	IRWESIO	WESIO	IND
TWN	IDN	FLL	IREASIO	EASIO	IND
TWN	LKA	FLL	IREASIO	EASIO	IND
TWN	MDV	FLL	IRWESIO	WESIO	IND
TWN	MYS	FLL	IREASIO	EASIO	IND
TWN	NEI	FLL	IRWESIO	WESIO	IND
TWN	SGP	FLL	IREASIO	EASIO	IND
TWN	SYC	FLL	IRWESIO	WESIO	IND
TWN	THA	FLL	IREASIO	EASIO	IND
TWN	TWN	FLL	IREASIO	EASIO	IND
TWN	TWN	FLL	IRWESIO	WESIO	IND
TWN	TWN	GILL	IREASIO	EASIO	IND
TWN	TWN	LL	IREASIO	EASIO	IND
TWN	TWN	GILL	IRWESIO	WESIO	IND
TWN	TWN	LL	IRWESIO	WESIO	IND
ZAF	ZAF	GILL	IRWESIO	SWEIO	ART
ZAF	ZAF	HAND	IRWESIO	SWEIO	ART
ZAF	ZAF	HOOK	IRWESIO	SWEIO	ART
ZAF	ZAF	SPOR	IRWESIO	SWEIO	ART

ZAF	ZAF	UNCL	IRWESIO	SWEIO	ART
ZAF	ZAF	BB	IRWESIO	SWEIO	IND
ZAF	ZAF	ELL	IRWESIO	SWEIO	IND
ZAF	ZAF	SLL	IRWESIO	SWEIO	IND
...	...	...	...	...	...

This is just an excerpt of the actual table modeling the proxy fleet / gears used for nominal catch disaggregation purposes (in red, we list the gears that actually area an aggregation of multiple gears).

When a Nominal Catch record containing either a gear aggregation or a species aggregation is encountered, its current fleet / gear / fishing area (Eastern / Western Indian Ocean) are used to identify an entry in the above table.

If no entry can be identified, the Nominal Catch Disaggregation process stops and prompts for user intervention: in this case, users can decide to add the missing data within the table above and the process can continue.

Entries identified within this table are used to determine a likely region and type of operation that eventually will be used by the disaggregation procedures to identify – when possible – alternative, non-aggregated Nominal Catch records for the substitution (see the definition of procedures 5, 6, 7, 8).

The substitution process is indeed the process of identifying and using alternate, non-aggregated records to determine the proportion (by gear and / or species) of the original Nominal Catch values to be assigned to all gears / species belonging to the aggregation found within the record being processed.

For this reason, at the end of the disaggregation process, the total Nominal Catch quantities (by fleet, year and fishing area) remain unchanged.

Appendix A.2 – Nominal catch disaggregation procedures

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
9	Manual disaggregation				

This table lists the 9 (8 automated plus 1 manual) substitution procedures that are applied – for each nominal catch record to be disaggregated – in the same order as they appear here.

The disaggregation process for a given nominal catch record halts as soon as one of these 8 procedures can identify one or more records for disaggregation purposes.

In order to find the non-aggregated records to use for the substitution, each procedure filters the nominal catch dataset based on its configuration and on the content of the proxy table (as reported in Appendix A.1).

Then, non-aggregated records are further filtered based on their gear and / or species belonging to one of the aggregates found in the original Nominal Catch record.

Procedure #9 (manual disaggregation) is triggered whenever the system cannot identify any meaningful record within the criteria set by any of the other eight procedures: users will be prompted to assign a reasonable proportion of the

original Nominal Catches to one or multiple records referring to any subset of the original gear / species aggregation for the strata.

As anticipated, there's no lower threshold in terms of number of non-aggregated records to be identified in order to consider the produced disaggregated results as reasonable (i.e. of good quality and with as less uncertainty as possible): we are currently in the process of evaluating the introduction of such a threshold, as well as of any additional refinement technique that we might found as fit for the purpose.