The New Design of IOTC's Database

by the IOTC Secretariat

Abstract

The original IOTC database was developed in Microsoft Access, a desktop DBMS. As the size and number of simultaneous users connected to the database grew, it became obvious that it was necessary to shift the database to a Client/Server system, capable of attending high-load requests from concurrent users over a network. In addition, during the normal use of the database several limitations of the original design were detected, particularly in terms of framework rigidity. To address these problems, a new design based on the Client/Server paradigm was implemented.
This document describes the new design, and discusses the concepts and objectives on which the new design was based. The new data base design was implemented on Microsoft SQL server and is currently fully functional.

Conceptual Organization of the New IOTC Databases

Conceptually, IOTC's databases can be grouped into five different categories based on the source and practical use of the information they contain (Figure 1):

- Nominal catches.
- Catch and effort.
- SizeFrequency distribution of the catches.
- Fishing craft statistics/Vessel registry.
- Accessory information.

Nominal Catches: Contain the highest level of fishery data aggregation, as provided by member countries and/or inferred from scientific or official publications and/or any other reliable source. Nominal catches are usually aggregated by country, fishing ground or geographic area (frequently a grid), time (usually in year units), fishing gear and species. Strata in this category tend to be relatively homogeneous, because most of the data is obtained through liaison officers of member countries on a pre-agreed minimum level of stratification. However, because the levels of aggregation may be conditioned to particular internal policies of each country, the stratification level of nominal catches may be subject to changes, in particular the time and area strata.

Catch and Effort : Contain the lowest level of fishery data aggregation, as provided by member countries and/or inferred from scientific or official publications and/or any other reliable source. Catch and effort data are usually aggregated by country, fishing ground or geographic area (ideally a 1x1 degree grid), time (ideally in month units), fishing gear and species. Although IOTC's target is to collect information at the aggregated levels just indicated, confidentiality restrictions, the internal data-collection mechanisms of the involved countries, and/or lack of information some times result on catch and effort data provided at different levels of aggregation. Ideally the catch and effort information is provided as raw data, but a portion of the holdings in the database reaches IOTC as data raised (extrapolated) to the Nominal Catches.

Size frequency distribution of the catches: Contain tallies of fish per size or weight intervals. The stratification parameters and level of aggregation are similar to those in the catch and effort category. Ideally Size/Weight frequency information is provided as raw data, but a portion of the holdings in the database reaches IOTC as data raised (extrapolated) to the Nominal Catches.

Fishing Craft Statistics/Vessel Registry: Contain tallies of boats aggregated at the same levels than nominal catches, and by boat type, class and preservation system. This category will be eventually replaced by the Vessel Registry database, but currently they are being handled as two separate databases.

Accessory information: Composed of a large group of tables with coding information (e.g. species codes, country codes, etc.). These tables are used with a double purpose; to enforce data integrity, and as look-up tables for report generation and data-entry/editing facilitation.

When seen from the point of view of flow of data, IOTC's databases are organized in two main layers (Figure 2):

Raw Data: In this layer data is handled and stored with the same stratification parameters and level of aggregation it had when it was received from the original source. Raw data goes through several processes designed to document and ensure the quality of the data:

1) **Data Reception and Registration of Data Submission**: Raw data is received in many different formats and media (e.g. Excel spreadsheets, ASCII files, e-mail, postal mail, etc.). All data submitted to IOTC is registered in the Data Submissions Database and the originals stored according to the media.
2) **Data-entry/Codification**: Data that is not submitted in digital format is key-punched and coded into digital files. Data submitted in digital format may be re-arranged and re-formatted during this phase, to make it easier to verify and document. These processes are done using different tools that range from custom-made programs (*e.g.* ILDDE, a program to enter and report data on longline landings and samplings), to off-the-shelf applications (*e.g.* Excel).
3) **Data Verification, Correction and Documentation**: During this phase, the original data is revised for inconsistencies and omissions. Corrections may be performed (usually in consultation with the data source). Additional information documenting any changes, inconsistencies and corrections may also be added to the original data in this phase. At this point, data is stored and backed-up in an loosely organized database.
4) **Reorganization**: This process involves re-arranging the data so it can be easily added to IOTC's database. It is important to note that during this phase, the parameters and levels of aggregation of the original submission are never changed.
5) **Incorporation in IOTC's main database**: Once data has been re-organized, it is added to the main database. Reorganization and incorporation of data is usually done automatically with a series of custom-made programs and procedures that ensure data integrity.

Analyzed (Processed) Data: This layer contains information that has been processed to produce homogenized, aggregated and summary statistics. This layer is implemented as a series of pre-written queries, views, stored procedures and intermediate tables, and is still under development. The layer can also be subdivided into two sub-layers:

1) **Standard database polling and homogenization procedures**: Composed of some basic tools that provide the capability of polling the database for available information for

given set of strata, as well as queries and views that present raw data into flat tabular formats.

2) **Customized procedures**: Developed to generate summaries and aggregated data statistics, and which may use raw-data directly or rely on the standard polling and homogenization procedures.

Main objectives of the new database design

The development of the new database design was based on some principles and objectives that were considered particularly important for a fisheries database. This section briefly describes these principles.

Separation or raw and processed data

Flexibility to accommodate diverse information and facility of use of the stored information are frequently opposite paradigms in database designs. A database that is flexible enough to store information in many forms and shapes can be extremely hard to use when the need to analyze the data arises. On the other hand, providing data in a consolidated, homogeneous form, suitable for research and analysis, frequently involves aggregation and transformations of the original data, with concomitant losses of information. In some cases, the consolidation and homogenization processes may also introduce additional assumptions that were not present in the original data, limiting the usefulness of the database as research tool. A good database design should minimize assumptions introduced in the data as artifacts of data-entry, standardization and normalization, and in those cases where assumptions are introduced, ensure they are well documented and understood by the groups using the database for research and analysis.

There are several factors that make the design of IOTC's databases particularly complicated:

1) IOTC receives information from many and diverse sources. Each source has its own particular scope of technologies for data-collection and processing, as well as internal restrictions and limitations on the quality and detail of the data that can be reported to the Commission. Standardized data-reporting forms and agreements minimize the source-related variability of the type and shape of the information. However, the differences are still large enough that the Commission has to implement individual data-entry and processing procedures for each source.
2) The Commission is not only a data-compilation, but also a data processing and analysis center.
3) There is a need to be able to track any and all changes that may be applied to data to make them conformant to the database design. The overhead and documentation information necessary to perform this function is relatively high.
4) The spawn of the research work to be done on the final data is very wide and in some cases still unforeseen.

The two-layer design explained in the previous section addresses fairly well the four constraining factors described above, it also provides the necessary flexibility to accommodate evolving needs in the areas of research and data analysis.

It is also important to note that with this two-layer design, most of the decisions regarding the methods and procedures used to homogenize, aggregate and standardize the data are removed from the database. Instead these decisions take place at the analyzed-data layer and are completely controlled by the scientists undertaking the analysis.

The main disadvantage the design is that the raw-data layer is still a relatively complicated, so developing the procedures to produce information for the processed-data layer is far from trivial. Additionally, because the processed-data layer is mostly provided by the result of queries and views that are run in real-time, the design is conductive to high rates of CPU activity.

Flexibility and Adaptability

The flexibility and adaptability of the database design had to be considered at two levels:

1) The same type of data (*e.g.* catches, effort, etc.) can be provided in different formats, different levels of stratification and different units.
2) The prospect of having to integrate new and unforeseen data into the database design in the near future

The first issue was addressed by incorporating a design that normalizes stratification parameters (*i.e.* country, fishing grounds, time, etc.) independently of the data (*e.g.* catches for a given species). This approach is explained in more detail later in this section. The second issue was addressed by segregating units of information into logical groups of tables. The tables inside a logical group may be related to each other, but are independent of the tables that belong to another logical group. The logical groups closely match those described in the section that describes the conceptual organization of the database. The link between logical groups is established only at the first level of the stratification parameters (Figure 3). Separating the entities that define the relationship between logical groups of data from the data itself provides the flexibility necessary to incorporate new data, or new relationships, without having to perform extensive modifications in the database structure. For example, integrating oceanographical information into the database would be a matter of adding the tables in the new information and establishing the link to the appropriate combination of stratification parameters.

Multiple estimates of the same data

Another important objective was to provide a design that would allow for multiple estimates of the same piece of information. These may appear in two forms; multiple estimates of the same data provided by different sources, and multiple estimates of the same data provided in different units (for example, the same effort measured in days at sea and in number of sets/day). Some unit conversions (*e.g.* from kg to tons) are linear, so repeated estimates can be neglected without loosing information. In other cases however, the conversions are not linear or the information necessary to effect the conversions is not available. In particular, non-linear conversions usually involve adding assumptions and constraints to adapt existing data to a theoretical model. A typical example of this situation is effort submitted in fishing days and standardized effort. Failing to incorporate the two estimates in the database would represent loss of information.

The design addresses both issues by incorporating the source of the information and the units in which the information is reported as second-level and third-level stratification parameters (Figure 4).

Heterogeneous Space and Time Strata

IOTC receives information with different levels of stratification. For example, some countries report monthly catch and effort data aggregated over a 1x1 square-degree grid, others report their data aggregated over a 5x5 square-degree grid. Some countries report punctual data or data aggregated over irregular spatial polygons or time intervals. Because of internal restrictions and regulations, variations in the level of data aggregation may also occur inside a single data source. For example, a country may report data aggregated over 1x1 square-degrees one year and over

5x5 square-degrees the next year. There are two possible approaches to handle information with this level of heterogeneity:

1) Standardize the database to a fixed set of spatial and temporal strata, and force the received data into this standard scheme. This has been the traditional approach used in fisheries databases. It may involve a fair amount of data-preprocessing, extra and interpolations, data substitutions, etc. all of which introduce assumptions and constraints in the information stored in the database. The main disadvantage of this approach is that the introduced assumptions become part of the database, as well as any analysis, report and statistics derived from it. If the need to change or relax these assumptions arises, the only solution is to go back to the original data and recreate the database. The main advantage is that databases using this scheme are usually easier to use for end-users.
2) Provide a database scheme that allows for the heterogeneity of the stratification parameters, deferring the decision-making processes involved in aggregating and homogenizing data to the analysis phase. The major disadvantage of this approach is that the database becomes harder to use by end-users. The main advantages are that data analysts are always aware of all the constraints and assumptions introduced in the data, and they can change them if necessary, without having to modify the information in the database.

The first approach would violate the constraint of maintaining a clear separation between raw and processed data, while still keeping both sets of data available. Because of this, designing a database capable of handling heterogeneous stratification parameters, while still maintaining some level of usability, became a priority.

To provide a suitable database framework with the flexibility to accept information with heterogeneous stratification, the stratification parameters were separated from the data into a independent entities. The analysis of the current stratification schemes indicated the existence of at least three well-defined levels of stratification parameters:

1) **Level one stratification parameters**: These are common to all the logical groups of the database and include three factors: Country, geograhical area or fishing grounds (referred in this document with the generic name of geographic features) and time-interval.
2) **Level two stratification parameters**: These are stratification factors that are used by individual logical groups and that further classify the data already stratified by level-one parameters. They include the following factors per logical group:
   a. Catch and effort: fishing gear, boat type, boat class and source of information.
   b. Nominal catches: Gear, species, catch units and source of information.
   c. Size frequency: Gear, species, school (or set) type, source of information, and type of measurement.
   d. Fishing craft statistics: Gear, boat type, boat class, preservation method and source of information (these factors will change when the Vessel Registry is implemented).
3) **Level three stratification parameters**: These are classification factors specific to the type of data included in a single table. They stratify information at a higher resolution than level-two parameters. For example, catch information in a given level-two stratum in the catch and effort database, is further stratified by species, set type and catch units.

Although there is an inherent hierarchy in the levels of stratification of the design (*i.e.* classification parameters at level two are always subclasses of stratification parameters at level one), the hierarchies can be relaxed by introducing codes that indicate that a stratum does not include classification based on a given parameter. For example, the catch of a given species could be stratified on a country-geographical feature-time interval-gear stratum identified as Seychelles-Area10001-(10/10/91-10/11/91)-Unclassified to indicate the stratum has no classification information at the level of fishing gear. For such a case, the data analyst can ignore the stratum completely, or choose any of many different methods to distribute the unclassified catch over other gears for the same Country-geographical area-time interval stratum. The advantage of this system is that it is possible to work with different combinations of classification parameters without loosing the information that exists in unclassified strata.

Data Integrity

As is the case for most databases, specifying the mechanisms to ensure the integrity of the data is an important part of the design's objectives. These mechanisms are responsible for avoiding orphaned records in tables, fields with codes that have not been defined, etc. In addition, the database also implements a mechanism to time-stamp record changes that will help to track errors in the data. Data integrity verification and enforcement was implemented using standard database procedures, including foreign keys, table constraints and record-level triggers. The benefits of such self-verification mechanisms became obvious while the information from the old database was being transferred the new design; an number of irregularities in the data were immediately flagged and corrected.

Current Status and Conclusions

The design briefly described in this document has already been implemented and is currently used by IOTC's staff for their routine data edition and analyses. Figures 5 to 7 show abbreviated versions of some of the logical groups in the design, with many of the less-important tables removed for clarity. The design has proven to be flexible enough to allow changes without large restructuring of the database and to accommodate the heterogeneity of submitted data, while at the same time maintains a level of complexity that is accessible to most users with some basic knowledge of database querying techniques.