Data toolbox for fisheries: the case of tuna fisheries

Paul Taconet[*], Emmanuel Chassot[†], Jérôme Guitton[‡], Fabio Fiorellato[§], Enrico Anello[¶], Julien Barde[‖]

SUMMARY

Assessing the status of tuna and tuna-like populations for providing management advice requires the analysis of multiple data sets collected by the contracting parties and cooperating non-contracting parties of Tuna Regional Fisheries Management Organizations (tRFMOs) Conventions. Data on the magnitude and composition of landings, discards, and fishing effort are currently managed at basin scale by the Secretariats of the tRFMOs. Consequently, data formats and reference codes have evolved rather independently despite some links with the FAO Coordinating Working Party on Fishery Statistics. We have developed a global harmonized database for tuna fisheries data by collating the public domain datasets (total catch, monthly-spatially aggregated catch and effort, and catch at size) from IOTC, ICCAT, IATTC and WCPFC. The database currently covers the period 1919-2014 and is freely accessible online along with a set of open source codes to handle the data, i.e. transform the data formats, load the standardized data into the database, extract data and compute a suite of indicators (e.g. global maps of catch). The database along with the codes represent the "toolbox".

In a previous note ([Taconet et al.]), we have described the general methodological approach used for the collation and formatting of the data, as well as the benefits of using standardized data formats and code lists for scientific research and fisheries management. In this note, we present the codes available in the toolbox, the services that they offer and how users can ac-

[*]Institut de Recherche pour le Dveloppement, UMR MARBEC, CRH, Avenue Jean Monnet, BP171, 34203 Ste cedex, FRANCE. ; Paul.Taconet@ird.fr .

[†]IRD, Seychelles Fishing Authority, BP570, Victoria, SEYCHELLES. ; Emmanuel.Chassot@ird.fr .

[‡]Agrocampus Ouest, UMR ESE 0985, 65 Rue de Saint-Brieuc, 35000 Rennes, FRANCE.

[§]IOTC Secretariat, Le Chantier Mall (2nd floor), PO Box 1011, Victoria, SEYCHELLES. ; Fabio.Fiorellato@iotc.org .

[¶]FAO, Viale delle Terme di Caracalla, 00153 Rome, ITALY. ; Enrico.Anello@fao.org .

[‖]Institut de Recherche pour le Dveloppement, UMR MARBEC, CRH, Avenue Jean Monnet, BP171, 34203 Ste cedex, FRANCE. ; Julien.Barde@ird.fr .

cess them through various applications. We argue that these methods are likely to bring visibility to the tuna RFMOs public domain datasets, as well as to facilitate their use. Overall, we showcase how the variety of available services and applications enables to reach different communities of users that might be interested in accessing tuna fisheries data for various uses and purposes.

KEYWORDS: Catch/effort, Fishery statistics, Data services, Online processing

1. Introduction

Assessing the status of tuna and tuna-like populations for providing management advice requires the analysis of multiple data sets collected by the contracting parties and cooperating non-contracting parties (CPCs) of Tuna Regional Fisheries Management Organizations (tRFMOs) Conventions. In absence of fishery-independent data for most tuna fisheries over the world, stock assessment models mostly rely on commercial fisheries data that describe the magnitude and composition of landings, discards, and fishing effort. Such data are collected and processed by the CPCs through logbooks, landings and size-frequency samples and provided to the tRFMOs Secretariats following the rules (i.e. nature, formats, deadline) defined by the Conservation and Management Measures and Resolutions in force with each Commission. Datasets are stored and managed by the Secretariats to provide a holistic view of the tuna and tuna-like exploited populations and fisheries, and prepare the datasets for scientific analyses, including stock assessments. Hence, data formats and reference codes are currently managed at basin scale by each Secretariat. Consequently, they have evolved independently over time despite some common backgrounds and links with the Coordinating Working Party on Fishery Statistics of the Food and Agriculture Organization (FAO).

Building upon past and current projects aiming at providing overviews of the world tuna fisheries, we have developed a global harmonized database for tuna fisheries by collating the public domain datasets available from the International Commission for the Conservation of Atlantic Tunas (ICCAT), the Indian Ocean Tuna Commission (IOTC), the Inter-American Tropical Tuna Commission (IATTC) and the Western-Central Pacific Fisheries Commission (WCPFC). The datasets include total catches, spatially-aggregated catches and efforts, and catch-at-size derived from size-frequency data. Our overarching objectives are to: (i) review the tuna datasets available from each tRFMO and propose coding systems and standard nomenclatures to facilitate their merging for analysis, (ii) give more visibility to the data and more transparency to the processing steps driving to the datasets used as inputs for assessment models, and (iii) provide tools to facilitate data discovery, extraction, processing and visualization to anyone interested in tuna fisheries.

In a previous note ([Taconet et al.]), we have described:

- the general methodological approach used for the collation and formatting of the tuna RFMOs public-domain data,
- the data available in the global database,

- how the database model that we have developed enables to bring transparency to the data that are stored within the database and the processes that lead to the creation of new data,
- the benefits of using standardized data formats among tuna RFMOs for the sustainability of our project as well as for scientific research and fisheries management in general.

In this note, we present the set of open source codes that we have developed to handle the data, from the data reformatting to the analysis. We describe the services that they offer and how users can access them through various applications. We argue that these methods give more visibility and transparency to the data, and facilitate their use. Overall, we show that these services enable to reach different communities of users that might be interested in accessing tuna fisheries data: scientists, Countries or Contracting Parties of the tuna RFMOs, policy makers, Non Governemental Organizations, general public, etc.

2. The data toolbox: presentation, access and use

The data toolbox is the set of codes that have been developed to cover the whole chain of data processing: transformation to a common format, merging, loading into repositories, accessing and extraction, post-processing, computation of indicators. The aim of making this toolbox open and accessible on line is to enable users to reproduce the workflow, from data transformation to indicators computation. As users might be interested in only part of the workfow, the scripts can be used independently one from the others.

These codes, along with the data and web servers to make both data and codes available, provide *services* to the users. The data services are the various uses that can be made with the toolbox out of the data. They concern six main domains:

- **Pre-processing** (i.e. transformation and load) of the data,
- **Storage** of the data,
- **Discovery** of the available data,
- **Access and extraction** of the data,
- **Post-processing** (i.e. creation of own dataset) of the data,
- **Communication/reporting**.

Services are accessible through *applications*, that as such, represent the gateway to the data and the processes. Expected services might be different following the user's profile - nature, technical skills, etc. As examples, scientists might want to download data in a specific formats for further research, while CPCs might be interested in visualizing indicators on-line. Our objective is therefore to enable access to the data and use of the toolbox through a wide range of applications that cover all the services so as to answer to all the potential user's needs for data fisheries or statistics.

In this section, we describe the codes available in the toolbox and the various data services that they cover.
These scripts are available here https://goo.gl/h5GmHc. Work is ongoing to document them.

2.1 *Data pre-processing: transform data format and load them in the repository*

Some R scripts have been developed to pre-process the data:

- Transform all the raw tuna RMFOs datasets from their original data structure definition to the harmonized data structure definition (ICCAT doc),
- Load into the database or in given web repositories (see section Data storage) a dataset or a set of datasets whose data structure definition has been transformed to the harmonized one,

2.2 *Data storage*

We have developed an SQL database - called SARDARA - to store the data in a consistent and efficient way,. The database has been implemented with open-source software (PostgreSQL and PostGIS). Through a collaboration between FAO, IRD and technology partners in the context of the BlueBRIDGE project, it has been ported and is currently hosted on the iMarine platform and accessible online.
SARDARA is used as a facility to store and process the data, taking advantage of the benefits of SQL databases management system for data storage, format consistency checking and data processing - including geographical data. Data - raw data, processed data, code lists, mapping between code lists, ancillary data and metadata - are also available in CSV format on a public server to enable access to non-SQL experts.

Both data repositories - the SQL database and public servers to store data in any format - can be used by any user to store new data. Section Data

post-processing describes how users can create their own global or regional tuna fisheries datasets. A new created dataset can either be uploaded on the open database - taking advantage of all the services offered by PostgreSQL and PostGIS database management system in general, and SARDARA in particular - or on the public server for further sharing. Whether stored in the database or in the server, this dataset will be usable as input for the other data services (discovery, access and extraction, post-processing, communication and reporting).

Credentials to access the database:
host = db-tuna.d4science.org
database name = sardara_world
user= invsardara
password = fle087
port = 5432

Piece of R code to access the database and run queries:

```
1
# Load RPostgreSQL library
library(RPostgreSQL)

# Connect to Sardara DB
drv <- dbDriver("PostgreSQL")
con <- dbConnect(drv, dbname="sardara_world", user="invsardara", password="fle087",
    host="db-tuna.d4science.org")

# Run a query and get back the results as R data frame

# Example 1: get the list of all the raw datasets that have been imported and their
    metadata (description, release date, nature, etc.)

query_to_get_metadata_table<-"SELECT * from metadata.metadata WHERE table_type
    ='raw_dataset'"
metadata_table <- dbGetQuery(con, query_to_get_metadata_table)
# print first 10 lines
head(metadata_table)


# Example 2: get the catches by ocean, year, gear, species
query_to_get_catches_by_OceanYearGearSpecies<-"SELECT * from tunaatlas_indicators.
    tunaatlas_catches_by_ocean_year_gear_species"
catches_by_OceanYearGearSpecies <- dbGetQuery(con, query_to_get_catches_by_
    OceanYearGearSpecies)
# print first 10 lines
head(catches_by_OceanYearGearSpecies)
```

```
# Example 3: get the global time series of raw georeferenced catches catches (i.e. as
    distributed by the tuna RFMOs) limited to 10 rows
query_to_get_global_georeferenced_catches<-"SELECT *,st_astext(geom) as polygon_wkt
    from tunaatlas.catches_ird_raw_labels LIMIT 10"
catches_by_OceanYearGearSpecies <- dbGetQuery(con, query_to_get_global_georeferenced_
    catches)
# print first 10 lines
head(query_to_get_global_georeferenced_catches)
```

2.3 Data discovery: discover available data

Data discovery consists in providing access to the metadata and enable locate relevant datasets and processes for the users. Each data uploaded into the database or created out of already available data in the database comes with a set of metadata. These metadata are stored on a dedicated table - called *metadata.metadata* - of the database. The metadata provide information on:

- the dataset provider (person or institution),
- the dataset release or production date,
- a description of the dataset,
- the spatial and temporal coverage and resolution,
- the available dimensions in the datasets.
- in case of raw - or "primary" - dataset (such as the tuna RFMOs original datasets): a link to access to the dataset in its original format,
- in case of a processed dataset (such as the IRD version of global georeferenced catches): the code that has produced this dataset (e.g. R script, SQL query, etc.),
- in case of a mapping between code lists: date and information on the operator that has realized the mapping,

The raw metadata table is stored in the database, enabling anyone with SQL skills to explore it. In addition, work is ongoing to add to the toolbox some codes that extract metadata from the database to standard formats (e.g. ISO 19115). Metadata in such standard formats are very useful for dissemination of the information on the available data. In particular, work is ongoing to upload them to widely-used metadata catalogues such as the FAO Geonetwork. These web-based catalogues offer (i) ways to access metadata and corresponding data

through user interfaces using keywords or filters (such as spatio-temporal filters), and (ii) a gateway for users to discover the whole project and locate relevant datasets.

2.4 Data access and extraction: access and extract available data

The main challenges of data access and extraction services are to (i) propose ways (i.e. applications) to easily perform operations on the data - such as filters, aggregations, mappings, etc. - and (ii) enable the extraction of the data in several formats. Diversifying the ways to access and extract the data in order to adapt to the variety of potential users and uses of the data is also a major concern.

Some of the R scripts of the toolbox enable to extract the data from the database. These scripts take as input a given parametrization and return as output the data given the provided parametrization and in the selected format. The parametrization concern:

- the data to use (georeferenced catches, georeferenced efforts, total catches, catch-at-size)
- the level of processing of the data to use (raw or post-processed),
- the filters to apply (on gear, species, flag, area, time, catch unit, etc.);
- the aggregations to make (by gear, species, flag, area, time, etc.),
- the granularity to use for the aggregation dimensions (use gear or gear groups, species or species groups, time granularity - decade, year, semester, quarter, month),
- the code lists to use (either tuna RFMOs raw code lists or standard FAO code lists),
- the spatial intersection type to use if using a filter area.

The codes currently enable to extract data in CSV format. The following additional formats of extraction will be available in short term: MS Excel, ESRI Shapefile, NetCDF, JSON, GeoJSON, WFS, WMS. Work is ongoing to document and port these scripts open on line.

2.5 Data post-processing: create new data with own processing method

The concept of data post-processing is to identify a set of scientific corrections that can be applied to the raw data - e.g. units conversion (from number of fishes

to weight, or effort unit conversions), elevation methods, etc. - and write the corresponding scripts that perform these operations. Through these codes, it will be possible to parametrize the workflow with selected set of corrections and values to use for these corrections. These codes are currently under development. A data produced with these codes might then be used as input of all the other services (load into the database, create and publish metadata, manipulate data, computate indicators)

2.6 Data reporting and communication: compute and disseminate indicators

Reporting and communication aims at creating graphical indicators out of the data and disseminate them. A wide variety of indicators and visualisation formats does exist: static - images - or dynamic - web-based - charts or maps, automated reports (either static, i.e. in PDF or MS Word format, or web-based dynamic), etc.
Some R scripts have been developed to produce graphical indicators. These scripts produce:

- A time series plot, showing the temporal evolution of the fact (catch or effort),
- A pie map, showing the spatial distribution of the fact (catch or effort), aggregated by one dimension.

Work is ongoing to write other scripts to produce additional indicators. In particular, dynamic indicators and automated reports are interesting products to disseminate the data. As a first example, a dynamic map has been produced with the technical support of FAO though the BlueBRIDGE project (clic here to access). FAO is also working at creating parametrizable dynamic reports such at this (clic here to access).

3. Acknowledgements

We are grateful to all staff and personnel involved in collection and management of tuna fisheries data over the world over the last decades. We sincerely thank Nick Vogel and Carolina Minte-Vera from IATTC and Peter Williams and John Hampton from SPC for great help with the datasets from the Pacific. We are also grateful to our FAO colleagues, particularly Fabio Carocci and Anton Ellenbroek, for their work and support to the project. We finally thank Alain and Viveca Fonteneau for initiating the SARDARA project as early as the 1990s, Olivier Maury for continuing the work through the Pelagic Fisheries

Research Program of the JIMAR, the French ANR project REMIGE, and the IRD Tuna Observatory organization. PT acknowledges financial support from the the French National Research Institute for Sustainable Development (IRD), the French Ministry of Fisheries and Aquaculture and the International Seafood Sustainability Foundation.

References

P. Taconet, E. Chassot, J. Guitton, C. Palma, F. Fiorellato, E. Anello, and J. Barde. Global database and common toolbox for tuna fisheries.

4. Figures

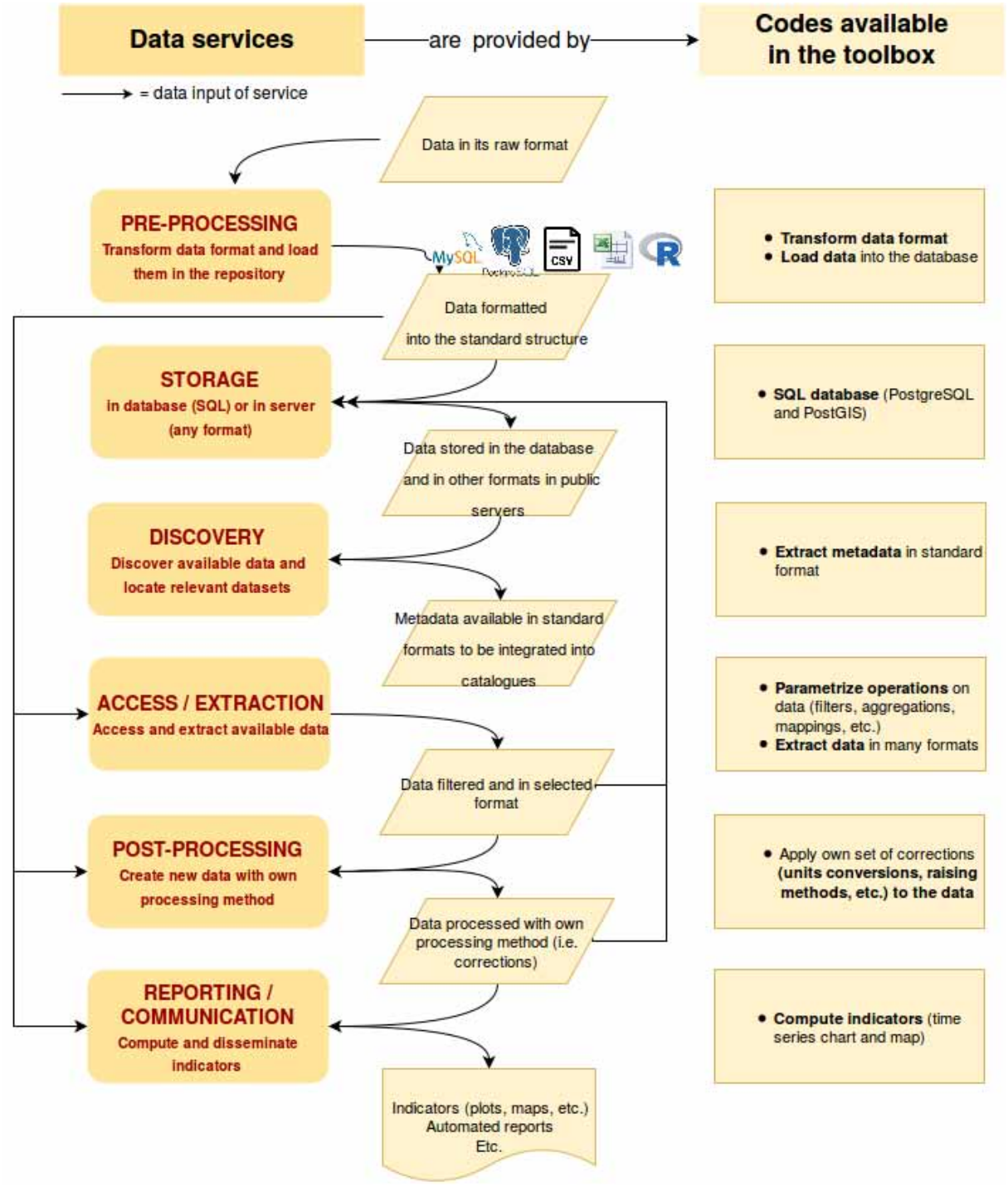

Figure 1: Description of the data services and corresponding codes available in the toolbox.

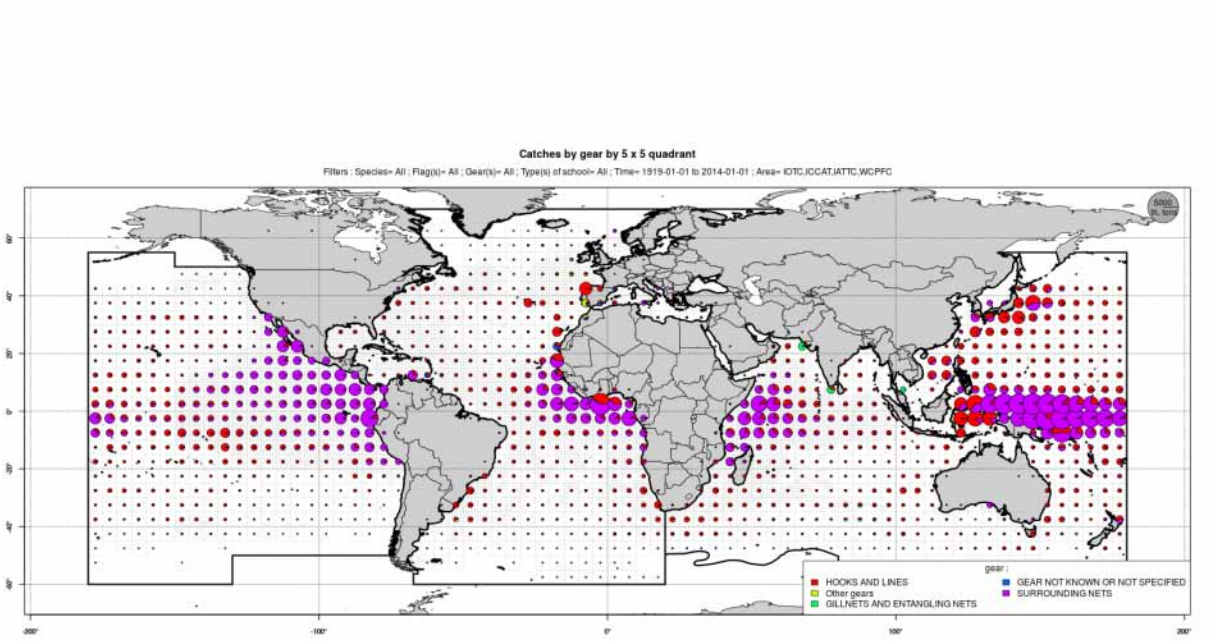

Figure 2: Map of catches by gear between 1919 and 2014. Extracted from SARDARA.

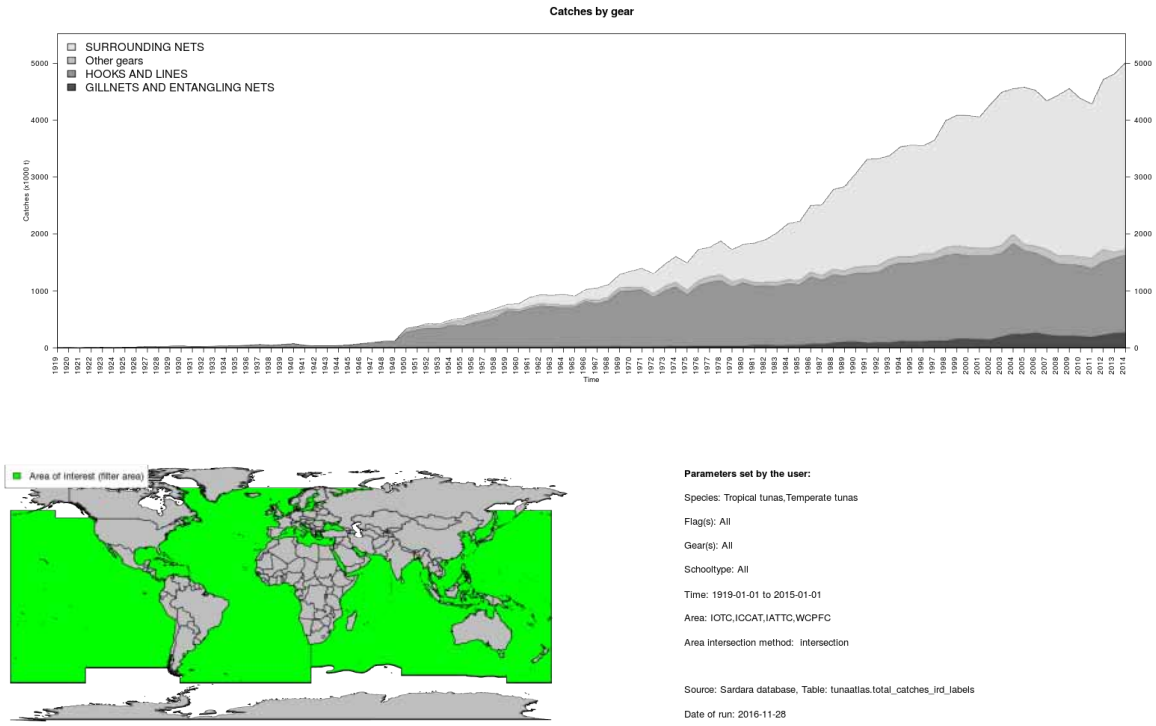

Figure 3: Time series of tropical and temperate tunas catches by gear between 1919 and 2014.Extracted from SARDARA.

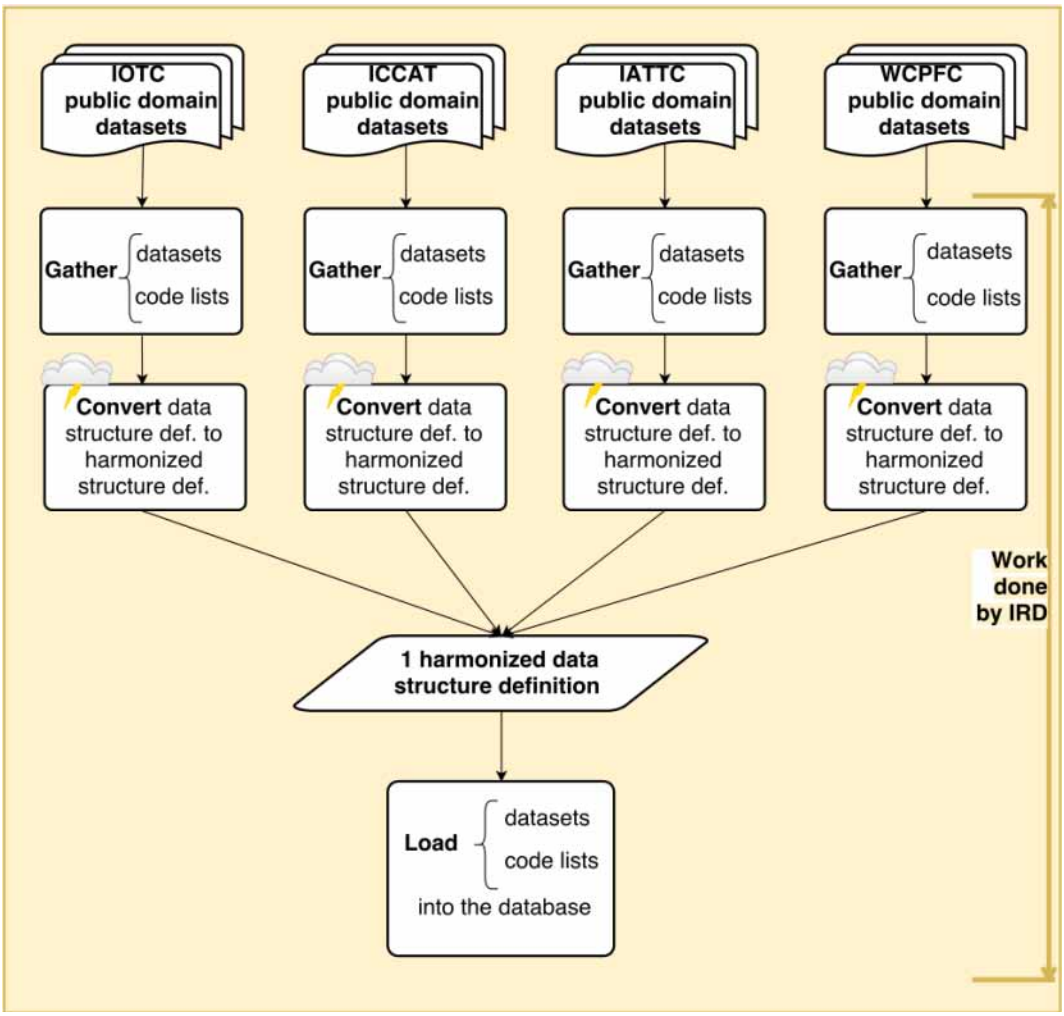

Figure 4: Methodological approach used to collate and merge public domain data from the tuna RFMOs.

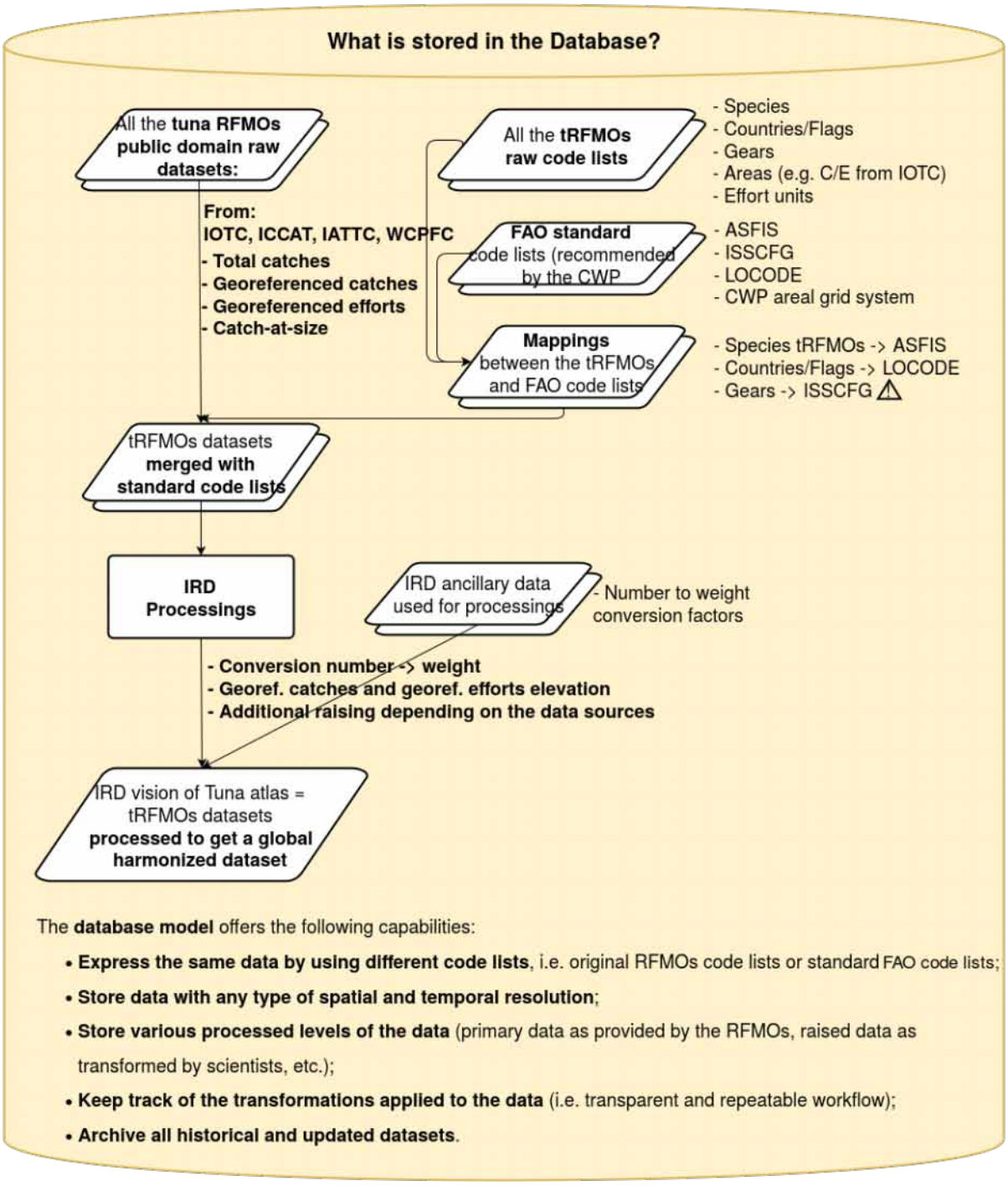

Figure 5: Data stored into SARDARA database

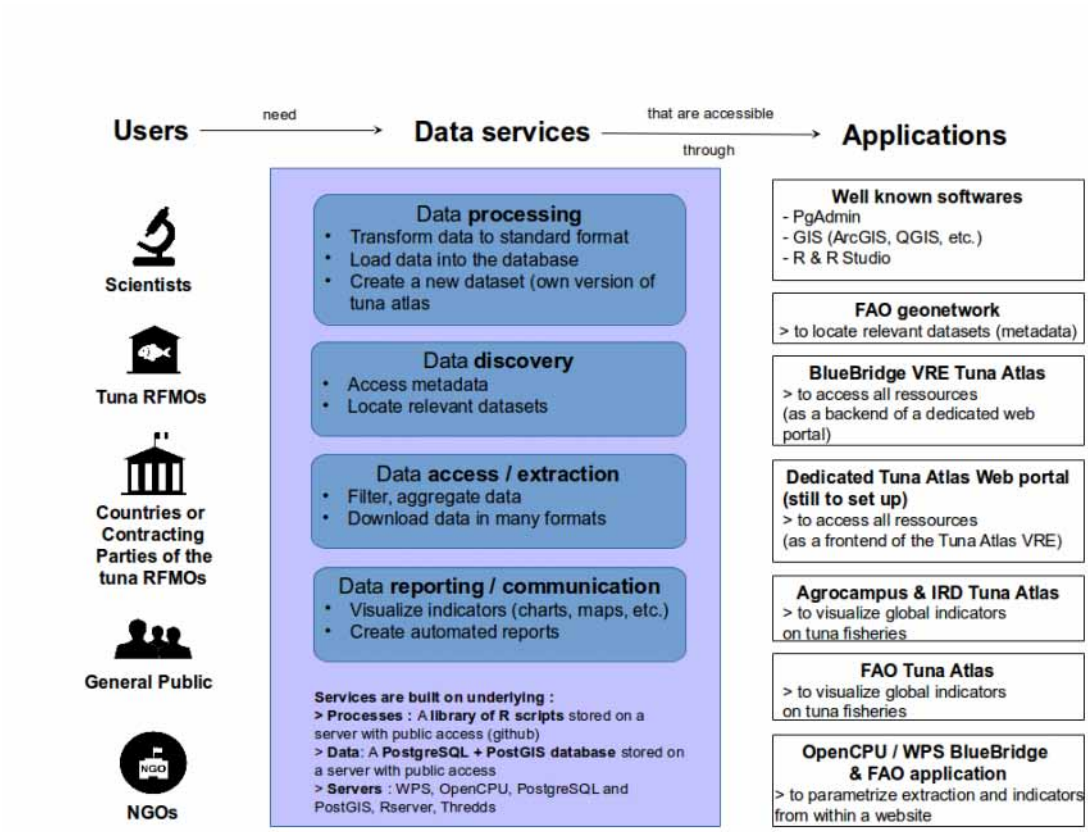

Figure 6: Diagram of potential users, data services and applications.

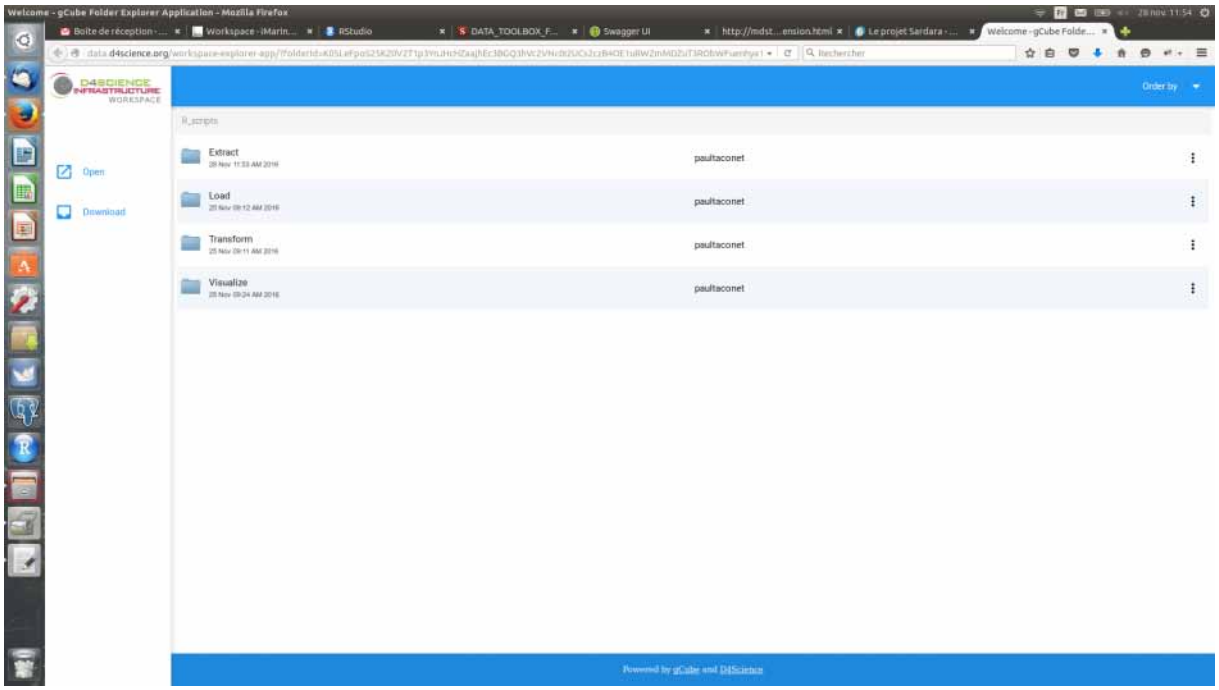

Figure 7: Codes available in the toolbox and stored in a public server. Clic here to access.

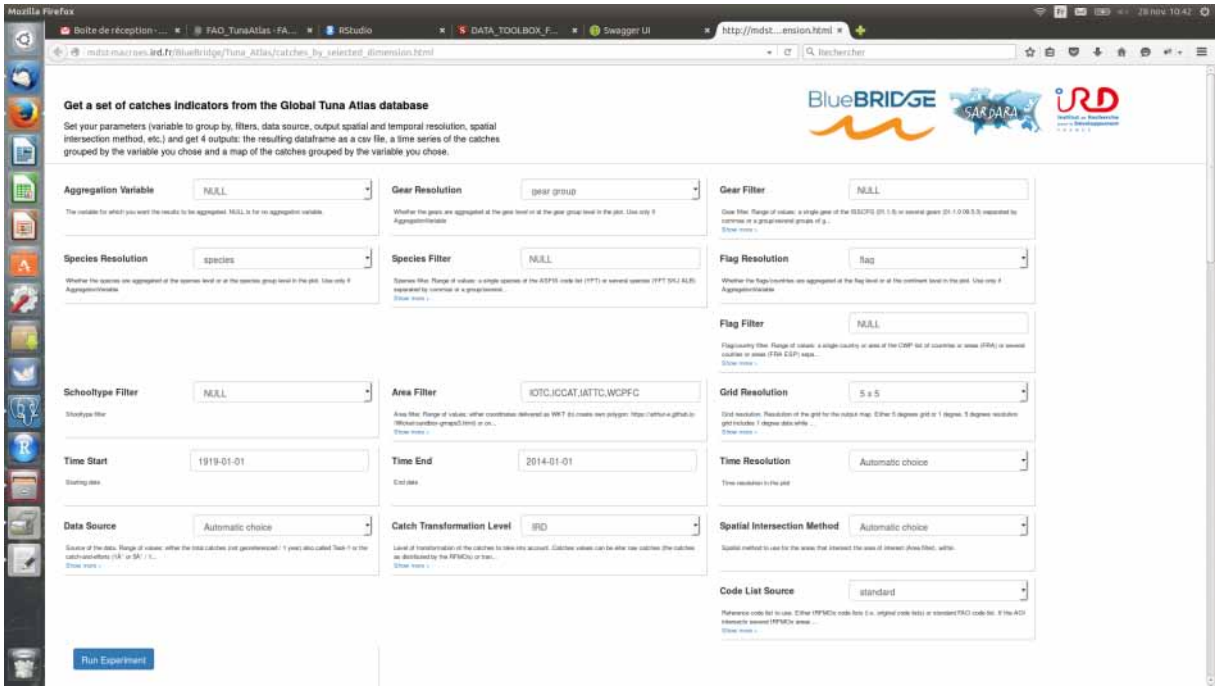

Figure 8: Web form to extract data and compute indicators. Developed by FAO within the BlueBRIDGE project. Clic here to access.

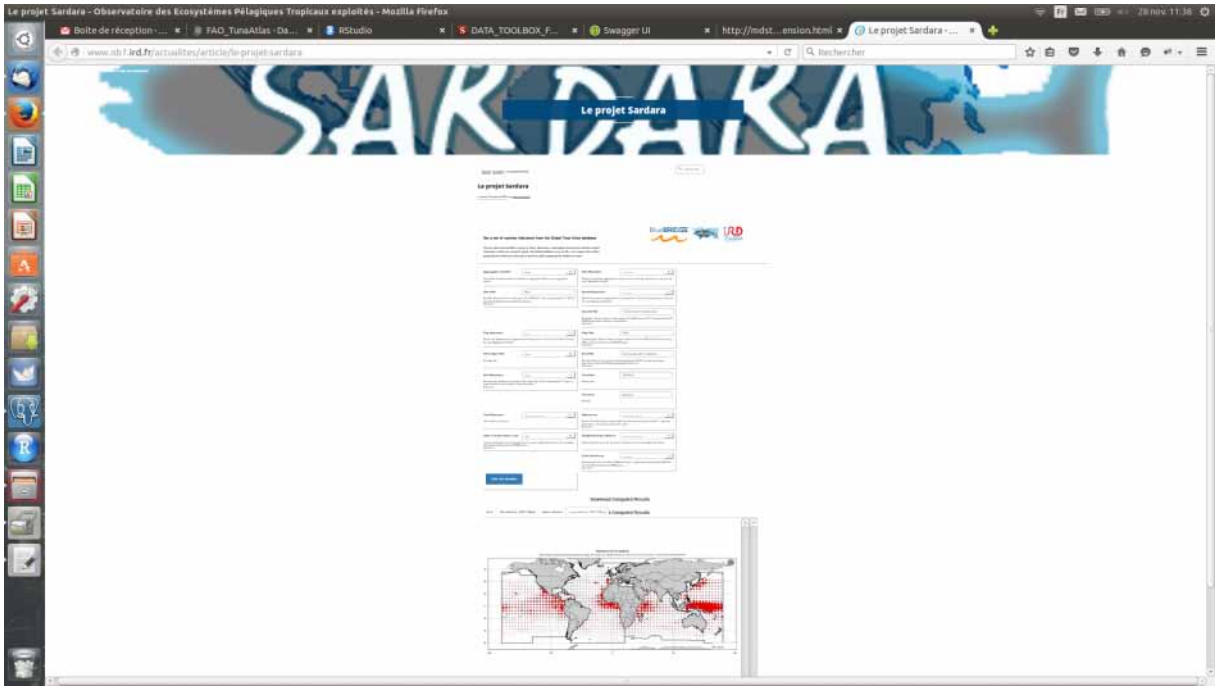

Figure 9: Web form to access data, wrapped into a web site. Developed by FAO within the BlueBRIDGE project.

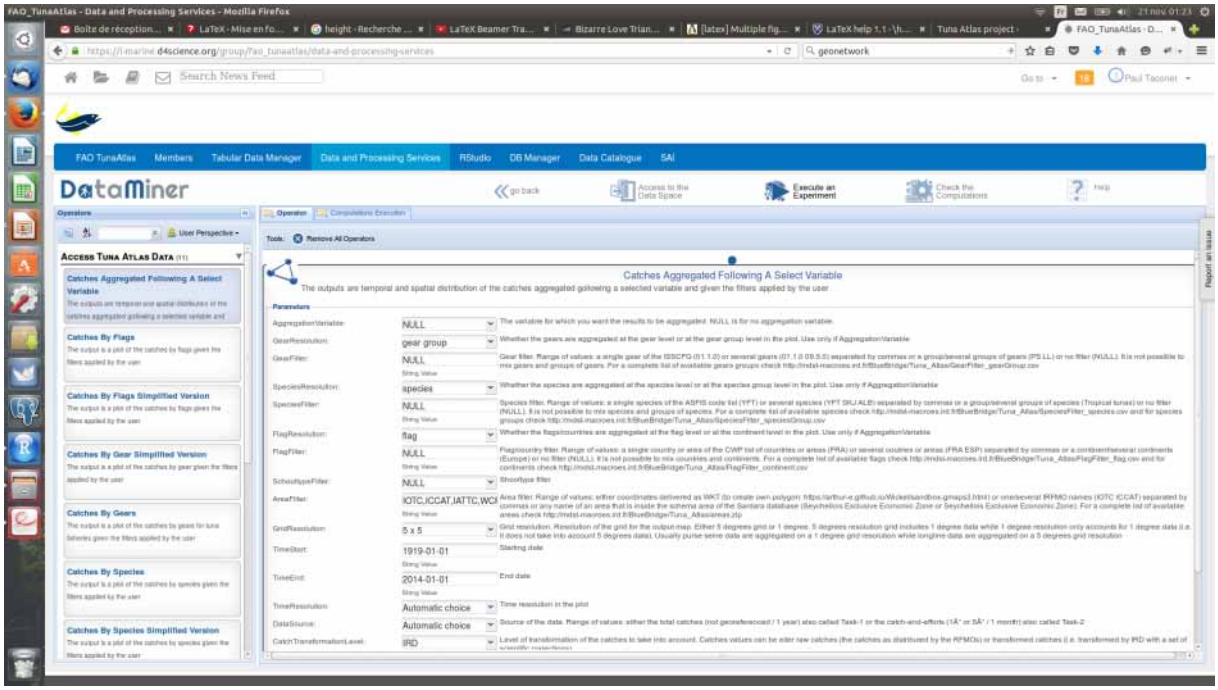

Figure 10: Web site (BlueBRIDGE Virtual Research Environment) to access all the data services: discovery, storage, extraction, processing, visualization.

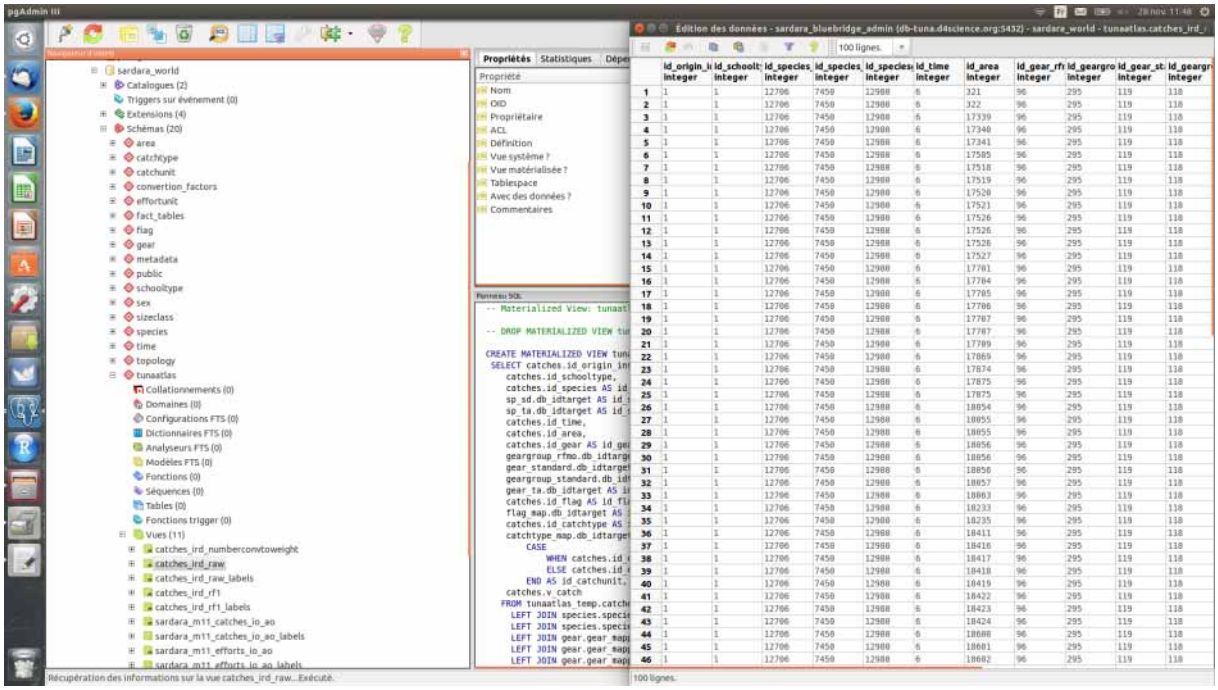

Figure 11: A desktop application - PgAdmin 3 - to access SARDARA database.