IOTC-2018-WPDCS14-34

# Describing and accessing biological and tagging data

Julien Barde*, Anne-Elise Nieblas†, Emmanuel Blondel‡, Nathalie Bodin§, Sylvain Bonhommeau¶, Emmanuel Chassot‖, Taha Imzilen**

November 26, 2018

## SUMMARY

*In 2017, a first attempt has been done to describe some IOTC datasets (dealing with stock assessment model outputs Nieblas et al. [2017] and fishing catch, effort, size class Barde et al. [2017]). Since, the method has been improved to make it more generic and reusable with other data sources. This paper gives an update of this work and focuses on ongoing efforts to describe other datasets of interests for IOTC. The description of datasets is achieved by using online collaborative environment which facilitate the contribution of the users. The descriptions of the users are then turned into proper metadata by implementing widely used standards to describe any kind of dataset or more specific kinds of data (eg spatial or biodiversity data). We present the main lines of the method, showcase some examples of outputs (metadata, datasets and related access protocols) which have been produced by focusing on two databases (RTTP tagging data, biological data, tracking data from pop-up). We finally discuss data management issues which go beyond metadata and data discovery in order to provide other services (standardization of data formats and access protocols).*

KEYWORDS: Indian Ocean, data standards, metadata, access protocols, data catalogue, interoperability, FAIR data management plans

*IRD, UMR MARBEC (IRD/Ifremer/Univ.Montpellier/CNRS), IOC, Rue de l'Institut, Ebène, Maurice; julien.barde@ird.fr

†IRD, UMR MARBEC (IRD/Ifremer/Univ.Montpellier/CNRS), IRD Réunion, 97744 St Denis, La Réunion, France

‡Independent Consultant

§Seychelles Fishing Authority (SFA), P.O Box 449, Fishing Port, Victoria, Mahe, Republic of Seychelles

¶IFREMER- DOI, rue Jean Bertho, 97822 LE PORT CEDEX, La Réunion, France

‖consultant, Seychelles Fishing Authority (SFA), P.O Box 449, Fishing Port, Victoria, Mahe, Republic of Seychelles

**IRD - UMR MARBEC 248, Av. Jean Monnet, 34200 Sète, France

## 1. Introduction

Some metadata have already been generated for fishing catch and efforts Barde et al. [2017] as well as stock assessment outputs Nieblas et al. [2017]. Since the method developed was generic enough to be reused with other data sources, we started to implement it with tagging data and biological data. By doing so, it is possible to generate additional standardized metadata for other IOTC datasets. Such metadata can then be loaded in metadata catalogs and served through standardized access protocols to foster data discovery in multiple data portals. In addition to metadata, if there is a will to do so, this method can make data themselves accessible through standardized data formats and access protocols.

In section 2. we briefly present the data sources which have been used for this study:

- data collected during the Regional Tuna Tagging Project,
- other biological data collected within different projects Chassot et al. [2017],
- tracking data (collected by Ifremer with pop up tags)

In section 3., we explain how we proceed to describe the different kinds of data sources by using either collaborative environment to foster the participation of users or worklows (R progamming language) for data managers to automate the work.

The section 4. showcase outputs of these method and workflows with examples (metadata sheets and related software) related to data sources presented in section 2..

We finally discuss the opportunity to make a deeper use of these standards (data formats and access protocols) and related benefits to handle both metadata and data. We present an example of rich application which can be built on top of these standards to visualize and access metadata and related data. We discuss as well the opportunity to use this method to publish data papers as done by Bodin et al..

## 2. Data sources

From a technical point of view, we consider two kinds of data sources:

- relational databases (SQL),
- NetCDF files to manage raster and in situ data.

In this work, we described tagging and biological datasets (managed in two relationnal databases) and a tracking data dataset (managed in NetCDF files).

### *2.1 Regional Tuna Tagging Project database*

The Regional Tuna Tagging Project (RTTP, 2004-2009, 8th and 9th European Development Fund) was driven by the Indian Ocean Commission and implemented by IOTC. The main datasets are public data managed and make accessible by IOTC. Data are stored in a database (ACCESS) and these public datasets are extracted from this database:

- Tagged tuna (during scientific cruises operated by RTTP project): accurate location, length,
- Recovered tuna (during professional fishing trips): estimated location, length, weight and, for some of them, biological samples (see Biological database in Chassot et al. [2017])
- Biological data (collected when tuna were damaged by the tagging activity in the RTTP project, or on dead tuna and tuna-like species caught by fishing vessels)

The figure 1 shows the spatial extent of the RTTP activities (tagging and recoveries from fishing vessels) and the figure 2 shows the structure of the database model used to store tagging data Nishan and IOTC [2010].

In 2010, the RTTP database has been first described by IRD by using ISO 19115 and 19110 ISO.

### *2.2 Biological database*

This biological database has been properly described in Chassot et al. [2017]. To summarize the database stores ecological data to describes fishing catches and environment. The biological data have mainly collected on tunas by various projects in Indian Ocean. Observations are of various kinds (morphometric measurements and ecological tracers observed from tissue samples...) and have been collected in collaboration with fishermen (onboard) and processing factories (landing).

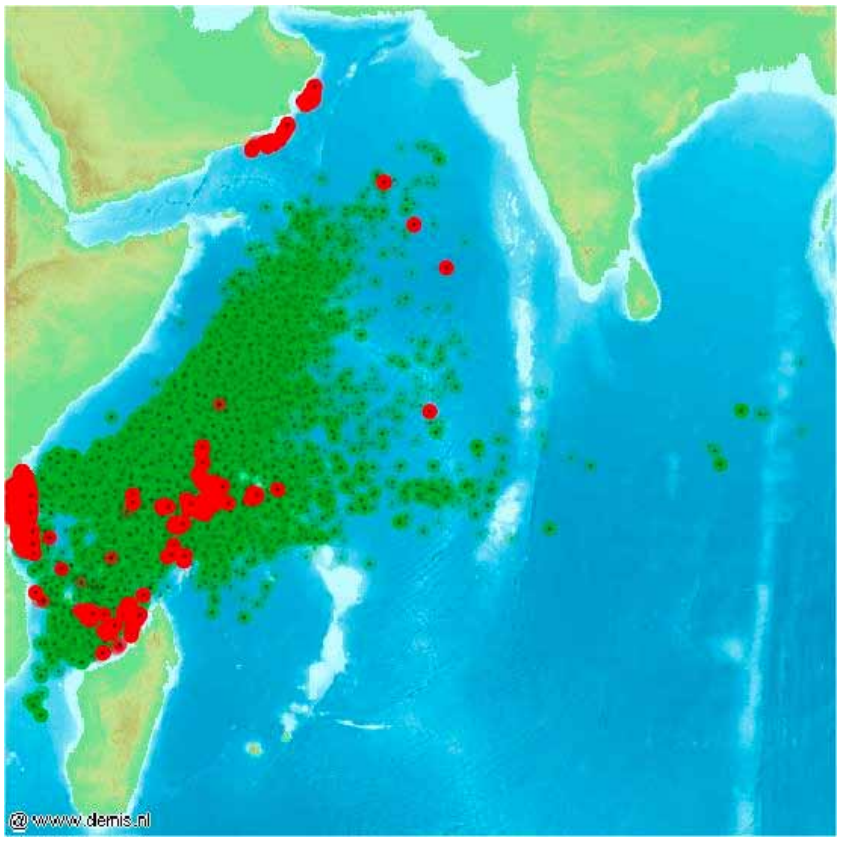


Figure 1: GPS location of tagging activities in red and estimated location of recoveries from involved fishing operations

The figure 3 shows the spatial extent of the observations stored in the database and the figure 4 shows the structure of the database model used to manage these biological data Chassot et al. [2017].

It is important to notice that this database embeds some metadata in a dedicated schema to describe the database:

- a data dictionary (table "ddd") which describes the tables and the columns they are made of by using the descriptors of ISO 19110:
  - entity: name of the table the column is part of
  - variable: name of the column
  - data_type: data type of the column
  - unit: unit of measure for values in the columns
  - description: definition of the column

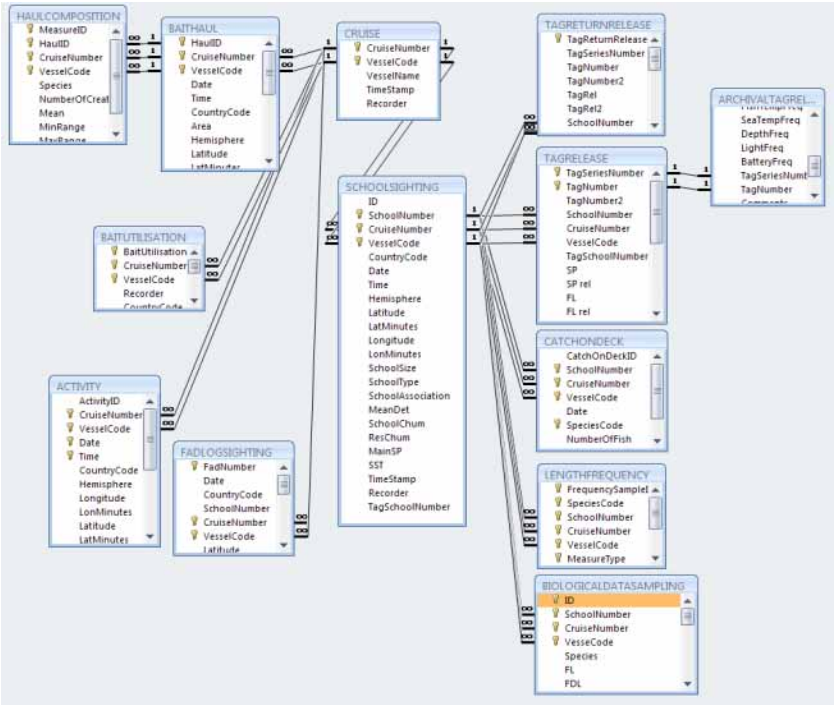


Figure 2: Physical model of the RTTP database

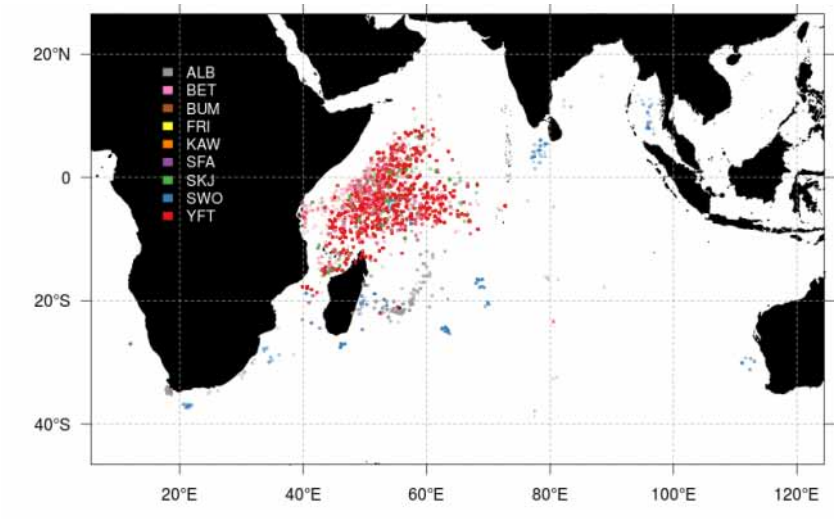


Figure 3: Spatial extent of observations stored in the biological database Chassot et al. [2017]

- description of measure names (eg "whole_fishweight")

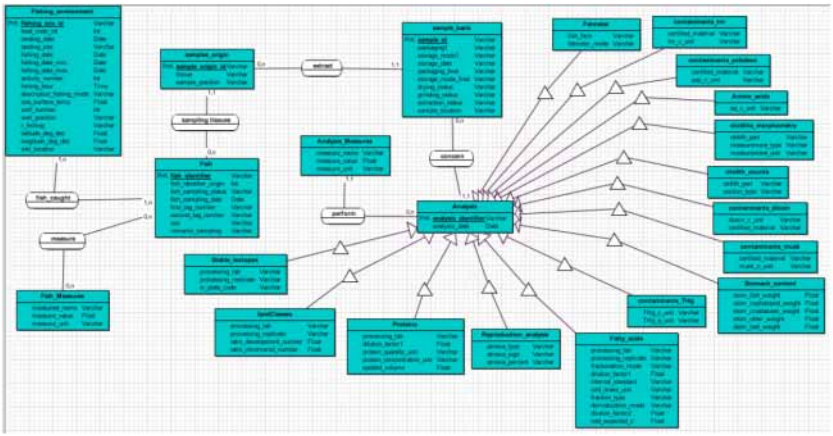

Figure 4: Physical model of the Emotion database to manage biological data Chassot et al. [2017]

- a table which describes the list of tracers ("analysis_tracers_details").

However, so far, this database didn't contain any table to describe the set of datasets which have been loaded in the database.

### *2.3 Tracking data*

These tracking data are fully described in Nieblas et al. [2018]. As explained, tracking data can be managed with standardized data format like NetCDF. Once these data are packaged in NetCDF files, it becomes then possible to directly harvest embedded metadata (contained in the header of the file) by using OPeNDAP access protocol. Headers of these files embed the main metadata elements as well as the description of the data structure (list of dimensions, variables and their metadata). It is also possible to read the data with the OPeNDAP protocol.

## 3. Method

### *3.1 Metadata Management*

Depending on the type of data source (see section 2.) the management of metadata is different.

- when dealing with NetCDF files, metadata can be embedded in the header of the file or managed within external (virtual) files using NCML language Nativi et al. [2005].
- for what regards relational databases, we suggest to add a new table which can be managed in a collaborative environment (eg google spreadsheet) and / or directly added within the physical data model.

### *3.2 Metadata standards*

We promote the use of following standards (ordered by priority):

- Dublin Core metadata elements (using a mapping with CF conventions for particular case of NetCDF files) Weibel et al. [1998],
- ISO/OGC standards for metadata: 19115, 19110 (data structure), 19119 (Web Services)
- Ecological Metadata Language is a TDWG standard for biodiversity data (and biological data)

Dublin Core is a simple metadata standards sufficient for data discovery whereas the others are more sophisticated but can be highly complicated to be implemented by newcomers. However, OGC and TDWG enable to describe the data structure and their content, and thus to build rich applications on top of them.

### *3.3 Inventory of reference datasets: catalog of queries*

We used a fairly simple method which consists in providing an inventory of reference datasets which can be extracted from each data source. This task is achieved by discussing with the users what are the most important datasets to be found and extracted in a given data source.

A database is a database management tool which factorizes and optimizes the quantity and quality of data. One can extract an infinite number of datasets by running different SQL queries. However some of them are more important than others and can be seen as reference datasets. Usually, we advice to distinguish the extraction of following datasets:

- original datasets which have been loaded in the data source. This enables to restore the input datasets (with a higher level of quality)
- new datasets which are a recombination of original datasets to study specific scientific questions (real outputs of the data source).
- datasets to be published as data papers (eg Bodin et al.).

The Figure 5 illustrates how the reference datasets are the outputs of SQL queries which can all be easily described by filling a simple table. This spreadsheet can be uploaded in collaborative environment to foster the participation of users (eg google spreadsheet) and / or by directly exploiting existing metadata when they are embedded in the data source (eg case metadata in the header of NetCDF files).

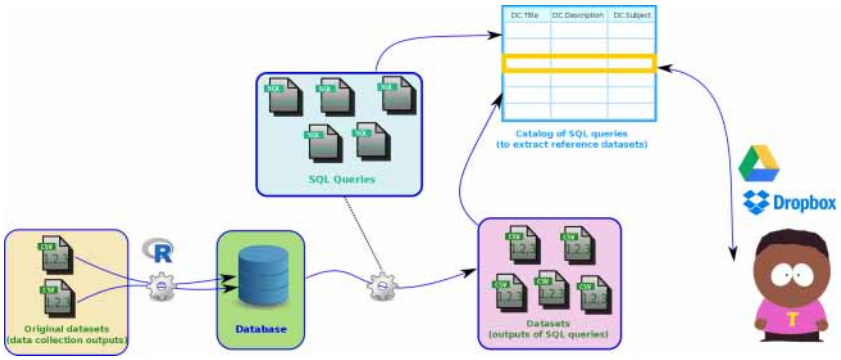


Figure 5: Reference datasets to be described are the outputs of SQL queries

The inventory of SQL queries in the table (spreadsheet) can then be refined to describe each dataset more accurately as explained in 3.4.

### *3.4 Description of each dataset*

Once the inventory of reference datasets is defined, each dataset can then be properly described with a set of main metadata elements. We use the Dublin Core metadata elements to set the names of columns. In addition we add a column to specify the related SQL query (the one which has to be executed to extract the dataset from the database).

The data structure is thus partly known since the SQL query gives the list of columns to be delivered. However some extra work is required to get the definition of columns and related data types.

Below, some examples of ongoing data descriptions done by using collaborative environments (Google spreasheet) to fill the main metadata elements:

- RTTP database and inventory of related datasets,
- Inventory of Biological data datasets

Once the list of reference datasets has been established with the users. We can run a workflow written in R programming language to generate standardized metadata, data and related access protocols.

## 4. Results

### *4.1 Metadata catalogs*

The figure 6 shows a metadata sheet describing RTTP database and the Figure 7 shows a metadata sheet describing stock assessment model output (from a NetCDF file). The main metadata elements come from the spreadsheets filled by the users and more specific metadata elements (eg spatial and temporal extent) are calculated on the fly by reading the data (executing the SQL query or OPeNDAP access).

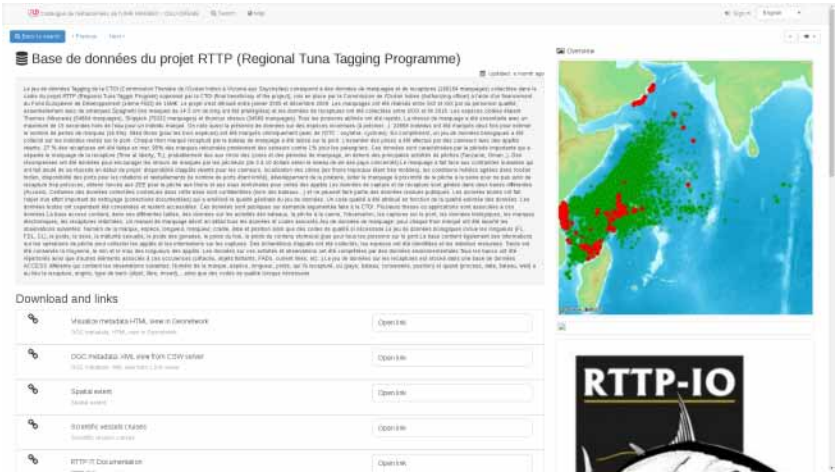


Figure 6: **Example of metadata outputs for RTTP databse**

### *4.2 Data access and visualization*

ISO standards can be used to set up rich applications and this has been already implemented in the marine domain Faucher and Lafaye [2007].

The South Pacific Commission (SPC) and Western & Central Pacific Fisheries Commission (WCPFC) have developed an application to manage biological data (cf SPC Specimen Tissue Bank, see snapshot in Figure 8), end-users can easily extract data by using following filters:

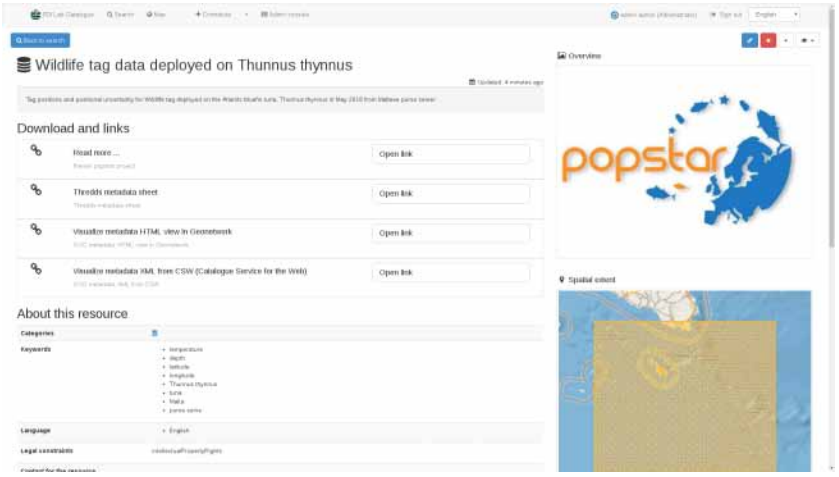


Figure 7: **Example of metadata outputs for tracking data of Popstar project**

- temporal extent of observations: by setting a start data and an end date,
- spatial extent of observations: by using EEZ of counries or Zoom in,
- species,
- types of tissue sample (blood, gonad, liver, muscle, otolith, stomach...),
- types of observations (port sampling, onboard observer...),

The visualization consists in a map which displays the location of samples and indicates the number of samples, the species and the type of tissue.

It is important to notice that we can generate similar applications by relying only on services provided by the method and related R workflow presented in section 3..

Indeed, a similar application has been generated by setting up a Javascript client which consumes OGC services to access metadata and data through standard Web Services (CSW and WMS/WFS). In the same way, the users can visualize and access data (see Figure 9).

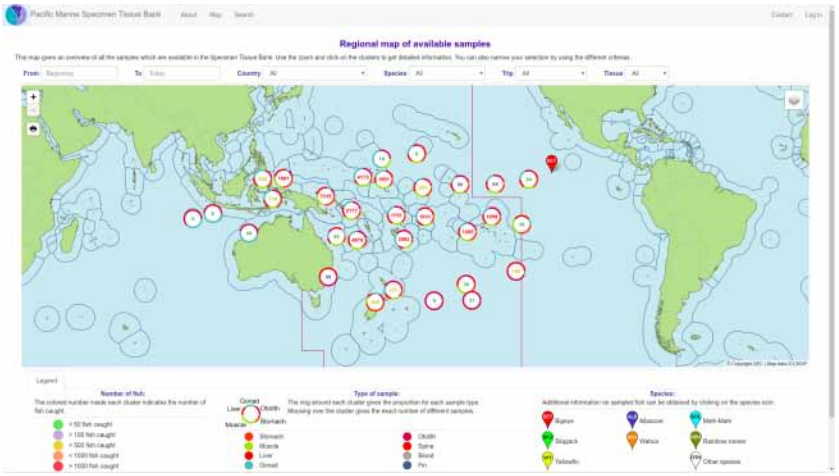


Figure 8: **Example of stock assessment metadata displayed by Geonetwork once transformed in OGC metadata**

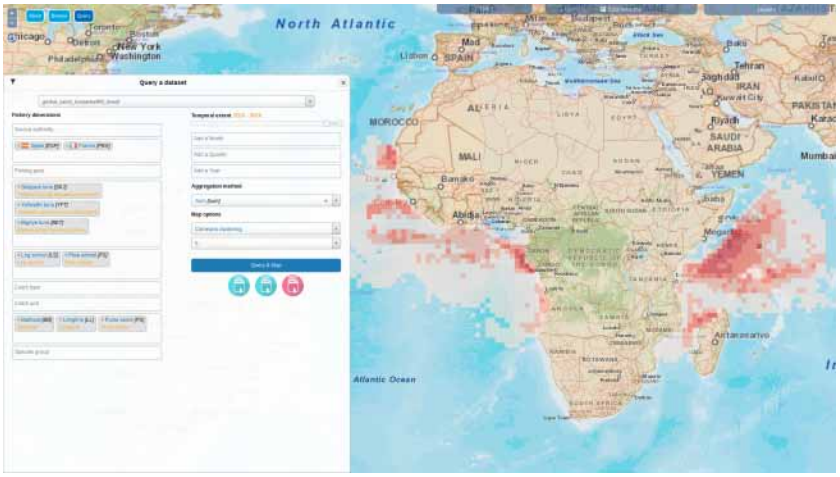


Figure 9: **Example of stock assessment metadata displayed by Geonetwork once transformed in OGC metadata**

## 5. Conclusion and outlooks

During the past two years, we have been able to describe some of IOTC datasets by reusing the same approach with different users in different projects. Step by step, the method became more generic and can now easily be reused to describe various types of data stored in widely used data management systems (relational databases and NetCDF/NCML files): catch and efforts observations, model outputs, tagging and tracking data. Indeed we provide workflows in R programming language whose outputs are standardized formats and access

protocols for both metadata and data. These standards are understood by multiple software (eg Geonetwork).

As presented in section 4.2, by making efforts to generate good metadata (including data structure description) it becomes feasible to plug generic clients to visualize and access the metadata and data provided by the workflow. We believe that such applications can become a motivation for users to describe their data in a better way.

Beyond technical aspects of data management, it is also important to consider the interest of this approach for data citation (when coupled with DOIs) and, ultimately publications like data papers which are basically another version of rich metadata (eg Bodin et al.) to be read by humans more than machines. However there is a mapping to be done between standardized metadata and data papers in order to fill some sections of the data papers with the values of standardized metadata elements. Indeed, most of the information that are already stored in XML metadata can be reused to fill the sections of metadata related to data papers (eg metadata of Bodin et al.).

**Acknowledgements**

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements No. 675680 (BlueBridge project) and No. 731011 (OpenAIRE Connect project).

## References


ISO 19110:2005(E). Geographic Information - Methodology for Feature Cataloguing.

J. Barde, E. Chassot, E. Blondel, T. Imzilen, A.-E. Nieblas, and P. Taconet. Collaboration between fisheries and computer scientists for improved data description : the case of IOTC data sets. page 11 p. multigr., 2017. URL http://www.documentation.ird.fr/hor/fdi:010071472.

N. Bodin, E. Chassot, F. Sardenne, I. Zudaire, M. Grande, Z. Dhurmeea, H. Murua, and J. Barde. Ecological data for western indian ocean tuna. Ecology, 99(5):1245–1245. doi: 10.1002/ecy.2218. URL https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/ecy.2218.

E. Chassot, J. Barde, L. Floch, L. Ibanez, R. Govinden, and N. Bodin. Open ecological data for tuna : the time has come ! page 10 p. multigr., 2017. URL http://www.documentation.ird.fr/hor/fdi:010071785.

C. Faucher and J.-Y. Lafaye. Model Driven Engineering for implementing the ISO 19100 series of international standards, 2007. URL https://hal.inria.fr/inria-00477571. http://www.coastgis07.com.

S. Nativi, J. Caron, E. Davis, and B. Domenico. Design and implementation of netcdf markup language (ncml) and its gml-based extension (ncml-gml). Comput. Geosci., 31(9):1104–1118, Nov. 2005. ISSN 0098-3004. doi: 10.1016/j.cageo.2004.12.006. URL http://dx.doi.org/10.1016/j.cageo.2004.12.006.

A.-E. Nieblas, Bonhommeau Sylvain, T. Imzilen, F. Fu, F. Dan, and J. Barde. Standardization of metadata, data formats, access protocols and statistical visualization of ss3 stock assessment outputs. 2017. URL http://www.iotc.org/documents/standardization-metadata-data-formats-access-protocols-and-statistical-visualization-ss

A.-E. Nieblas, Bonhommeau Sylvain, Imzilen Taha, K. Vincent, R. Tristan, and J. Barde. Enrichment of trajectories with environmental data, and standardisation of tagging data using netcdf. 2018.

S. Nishan and IOTC. REGIONAL TUNA TAGGING PROJECT - INDIAN OCEAN : IT CONSULTANT DOCUMENTATION. Technical report, IOC, Jan. 2010.

S. Weibel, J. Kunze, C. Lagoze, and M. Wolf. Dublin core metadata for resource discovery. Request for Comments 2413, 1998.

## Appendix

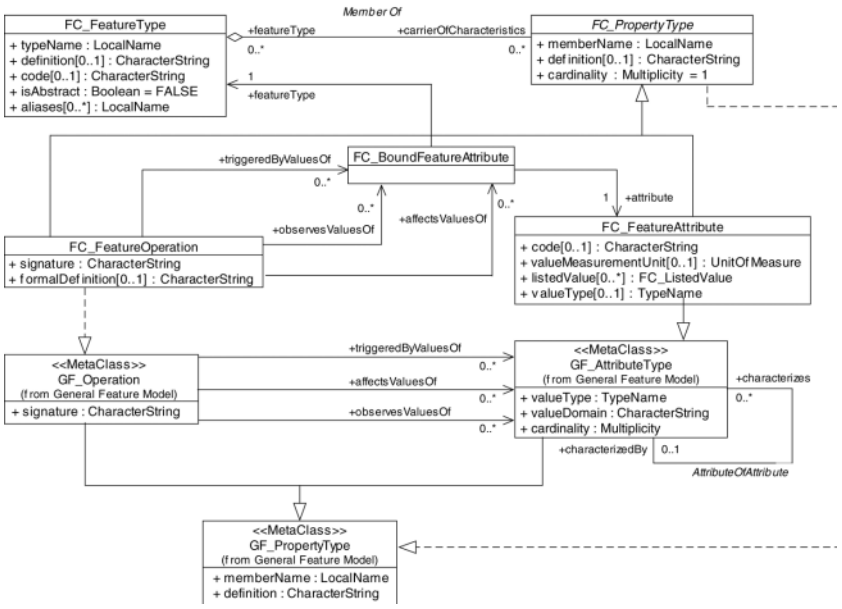


Figure 10: **Specifications used to set a data dictionnary: OGC Feature Catalog / ISO 19110**