Metadata for fisheries: ongoing work, examples and outlooks

Julien Barde[*], Sylvain Poulain[†], Emmanuel Chassot[‡], Fabio Fiorellato[‡], Emmanuel Blondel[§], Bastien Grasset[*]

SUMMARY

In this paper we present a brief overview of the current work that has been undertaken for a few years to make fisheries datasets compliant with FAIR (Findability, Accessibility, Interoperability, and Reusability) data management principles in the context of different projects. We will first recall what are FAIR good practices and discuss existing options to set them up. We will present some technical aspects and relevant infrastructures which can be used to implement such standards and can provide efficient data discovery and access services. We will thereafter highlight some examples which showcase the collaboration between tuna Regional Fisheries Management Organisations (tRFMOs), FAO and IRD in the context of FIRMS and EU-funded research projects. In particular, the Global Tuna Atlas is the most advanced demonstrator, and is currently focusing on total and geo-referenced fisheries catches across the world oceans. We will also present preliminary results of the Blue-Cloud and G2OI INTERREG european projects which aim to complement catch datasets description with fishing effort, tagging, and size-frequency data. These metadata sheets describe datasets by using a variety of standards: from international domain agnostic to domain specific standards. Beyond data, similar standards and workflows can be used to describe reports or working papers which can also be assigned Digital Object Identifiers (DOIs) along with metadata to improve their discovery and access.

KEYWORDS: Metadata, Interoperability, Data discovery, Data access, Fisheries, CWP, FIRMS, catch, Global Tuna Atlas, fishing effort

[*]IRD - UMR MARBEC 248, Av. Jean Monnet, 34200 Sète, France; julien.barde@ird.fr; Phone: +33 499 57 32 32 Fax: +33 499 57 32 15.
[†]IRD - UMR MARBEC 248 - Station SEAS-OI, Saint-Pierre, La Réunion
[‡]IOTC Secretariat (NFITD), C/O IOTC Secretariat, PO BOX 1011 – Blend Seychelles, Victoria, Mahé (Seychelles)
[§]FAO (NFISI), Rome, Italy

1. Introduction

The need to comply with FAIR data management principles is widely accepted by scientific organizations and becomes a requirement for scientists under the increasing pressure from funding agencies, journals, ... In practice FAIR data management principles are basically a set of good practices which are just obvious goals for any data manager. This is all the more needed and relevant for organisations which have disseminated open data for years without proper means to track the impact of the datasets made publicly available (how they are reused or cited). Indeed FAIR services are meant to make data Findable (using discovery metadata), Accessible (using proper formats and protocols), Interoperable (using standards) and Reusable (using rich usage metadata and DOIs). For years, IOTC has provided fisheries datasets (i.e., catch, effort, size-frequency) in the public domain along with metadata that are directly accessible on the pages of the IOTC Web site and can be downloaded through http protocol. However, this work and the data impact could be made more efficient (FAIR) if simply repackaged with proper formats and access protocols.

Technical aspects to set up such services are now mature enough and well specified by standards. Multiple tools and infrastructures are also available to implement these standards easily. Currently the main blocking point for data managers is still an overload of work and a lack of time or human resources to deal with this additional task. This is the reason why many funding agencies make Data Management Plans (DMPs) available but also make costs for DMP implementation eligible. Current projects can help bringing the extra resources that are needed to deal with this additional layer of work, requiring a limited input from the IOTC Secretariat.

First, we will describe the generic method that can be reused in the long-term to expand the current work and cover the variety of IOTC datasets which need to be better described. In the following sections we describe examples of our ongoing work in the framework of different projects that can assist fisheries data managers in general and IOTC in particular in better complying with FAIR data management principles. In addition we also illustrate the need for domain specific (fisheries-oriented) metadata where standardization efforts are still limited out of initiatives led by the Coordinating Working Party (CWP) on fishery statistics which mainly targets data structure definition.

2. Materials and Methods for implementation

For some years, IRD, FAO and tuna RFMOs are collaborating to develop a Global Tuna Atlas Taconet et al. [2017a], Blondel et al. [2020b] and this

collaboration became official under the umbrella of FIRMS in 2019. The underlying method uses R scripts and packages Blondel et al. [2020a], Blondel [2020a,c,b, 2022] to set up workflows which create FAIR services for spatial data in general and spatial fisheries data in particular Barde et al. [2017].

R has been chosen since years Taconet et al. [2017b, 2016] because this programming language is widely used in the fisheries community. It thus enables a closer collaboration between fisheries scientists and data managers. The kind of workflows that we currently experiment are meant to better describe and share datasets which are managed as flat files (e.g., CSV) or in structured sources (e.g., relational database).

The code of the workflow is open while data are as open as possible and are assigned a Digital Object Identifier (DOI) when publicly accessible.

The process steps of the R workflow are making data FAIR by using various software which can be executed on different infrastructures:

- Data:
  - stored both on Google drive and Postgres SQL server provided by Blue-Cloud H2020 project,
  - made available with DOIs Zenodo or GBIF data repositories,
- Code:
  - can be executed online with RStudio server (provided by Blue-Cloud H2020 or G2OI ERDF projects),
  - open source and available on GitHub repositories
- Grid software and hardware to provide FAIR services:
  - data discovery: GeoNetwork (e.g., FAO Fishery GeoNetwork),
  - data access: GeoServer with OGC REST services (WMS, WFS, ...) which can be used with multiple prorgamming languages (R, Python, Java...)

The European infrastructure of the Blue-Cloud project and the G2OI INTERREG projects both provide a collaborative runtime environment where such workflows can be executed without having to configure any server or PC.

3. Datasets at regional and global level

In this section we present examples of fisheries datasets at different scales. Building global datasets is the ultimate effort in terms of interoperability and we showcase how it can be properly managed with the example of the Global

Tuna Atlas and how a similar workflow can be replicated for both regional and national levels.

3.1 Global tuna fisheries datasets

Since years, tuna RFMOs, FAO, and IRD have collaborated to manage a common Global Tuna Atlas (using the GTA acronym) officially endorsed within the FIRMS framework. This initiative showcases how fisheries data can be used to fit the growing need for Essential Ocean Variables (EOVs) dealing with biology at the global scale. This initiative is also meant to showcase how best practices for the FAIR management of fisheries datasets can be implemented with a generic workflow written in R. In particular, the datasets of the GTA have been well described by using different levels of metadata elements:

- discovery metadata
- usage metadata
- provenance metadata including information about the reproducibility
- domain specific metadata (see example in section 3.4)

FIRMS GTA datasets and related metadata are published by FIRMS on both FAO fishery GeoNetwork (see example of metadata sheet in Figure 1) and GeoServer for data access. DOIs have also been assigned (using Zenodo repositories, see example in Figure 2) for catch datasets only at this stage. However, upcoming work will replicate the same worklow for efforts, conversion factors, and size-frequency data.

A similar workflow has already been set up for regional organizations (tuna RFMOs, WECAFC, ...).

3.2 Datasets at regional levels

Ideally, just like global datasets, regional datasets (e.g., at ocean basin scale) should be managed with same (FAIR) principles. The ongoing European Regional Development Fund (ERDF) INTERREG G2OI project (discussed in 2018 and started in 2021) suggests a partnership with IOTC to describe and foster data discovery of different IOTC fisheries datasets:

- size class, fishing efforts, catch datasets by mapping the free text content of Web pages with metadata elements of widely used standards (e.g., Dublin Core or OGC standards) (e.g., **Figure** 3)
- tagging data collected though the Regional Tuna Tagging Program of the Indian Ocean (IO-RTTP) past project: discussed during IO-RTTP

Figure 1: A metadata sheet describing one of the datasets of the Global Tuna Atlas on FAO GeoNetwork (OGC standard)

Figure 2: A metadata sheet describing one of the datasets of the Global Tuna Atlas on Zenodo data repository (Datacite standard)

symposium in 2012 [Murua et al., 2015]. A recent discussion with EU has confirmed that there is no objection from EU for IOTC to open this dataset as suggested in previous IOTC working papers Barde et al.

[2018], Chassot et al. [2017] (e.g., by assigning a DOI on the Global Biodiversity Information Facility; GBIF). **Figure** 4 shows an example of OGC metadata (with WMS layer) that can be directly generated from the SQL database.

- DOIs can however be also assigned to IOTC reports and working papers.

Indeed, so far, DOIs have only been assigned to global datasets of the Tuna Atlas by using Zenodo data repository. Part of these global datasets are obviously made of IOTC data but, still, these datasets are not properly acknowledging the contribution of the IOTC. However, the same approach and method could be applied at regional, Indian ocean, scale to assign DOIs to IOTC datasets specifically. By doing so, the reuse of IOTC datasets within the GTA would be better described in the provenance section which describes the input data used to create the global datasets.

Regarding the spatial data infrastructure (SDI) to be used to expose IOTC metadata and data with FAIR services, we currently use the infrastructure of the G2OI project as a temporary solution. However, for an official publication, using FAO SDI would be an obvious choice since some of the IOTC datasets are already described in FAO GeoNetwork and made accessible with FAO GeoServer. In the medium term, IOTC might also want to set its own infrastructure but it would not change the workflow already in place.

Regarding data repositories, we would suggest to publish fisheries data on Zenodo and biodiversity data on GBIF (e.g., IO-RTTP).

3.3 Datasets at national level

In this section, we present an example of metadata describing a dataset collected and managed by a national institute (IRD). In this case, a dataset which has been provided by Ob7 tuna observatory with french purse seiners data. Usually national data are managed at a higher resolution and only part of data attributes can be shared. The scientific value can be high, in particular for biological data collected in high seas where fisheries data are, by far, the main source of scientific data.

Since this dataset is stored within a relational database which is regularly updated, we plan to directly generate and update the metadata with a workflow that directly run SQL queries to handle the current version of the dataset to be described. By doing so, we also want to automate the creation of FAIR services and to minimize the load of work for data managers.

In terms of interaction with data managers, such an approach basically requires to provide a set of reference SQL queries which generate the various datasets for which FAIR services are expected to be created. To facilitate the

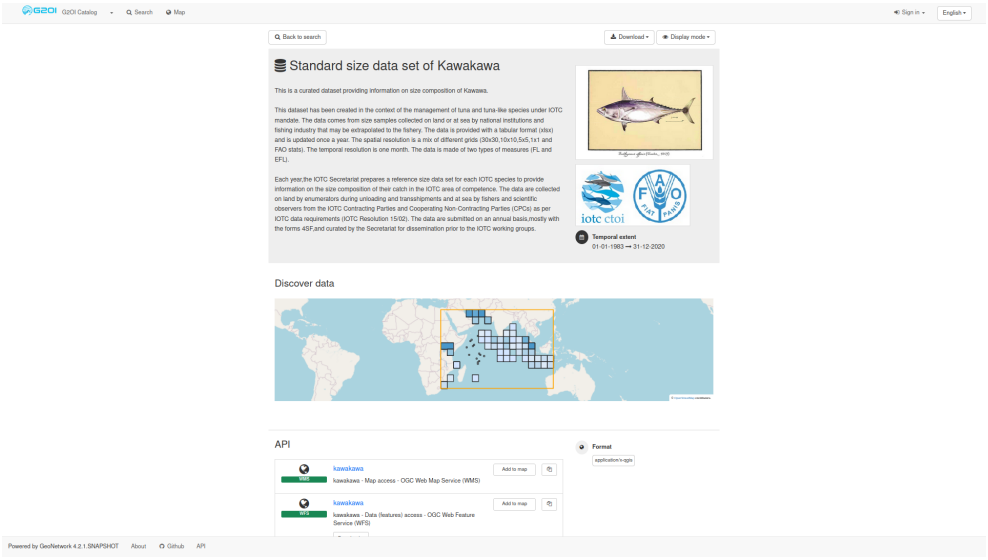

Figure 3: A metadata sheet describing the size-frequency data of kawakawa (*Euthynnus affinis*)

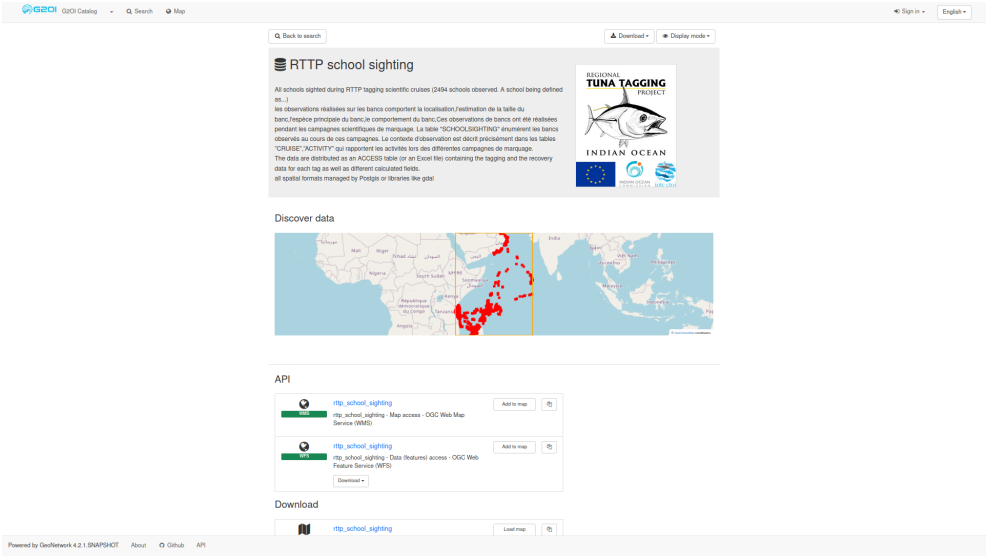

Figure 4: A metadata sheet describing the data collected through the Regional Tuna Tagging Programme of the Indian Ocean (RTTP-IO)

interaction, these queries can directly be provided through a GitHub repository. This enables the data managers to focus only on a set of reference queries which are transparently reused by the R workflow to feed spatial data infrastructures

with compliant (meta)data formats and protocols.

With recent incentives, scientists are more and more keen to share datasets and to assign DOIs either on GBIF (or Zenodo, Seanoe, ...) and, in the ideal case, by publishing data papers Taconet et al. [2017b], Guillou et al. [2022], Bodin et al. [2018]. The R workflow we use is helping to achieve this and the next section shows how it can be tailored to add domain-specific metadata.

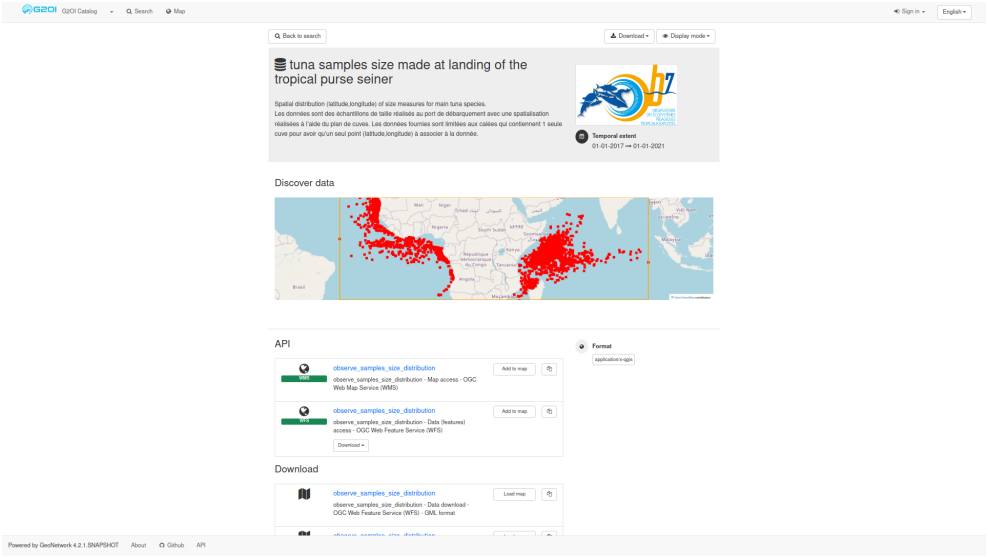

Figure 5: A metadata sheet on GeoNetwork describing French purse seiners data

3.4 Domain specific metadata

Some ongoing work, also funded by Blue-Cloud H2020 project, aims at generating more domain specific metadata that are relevant to describe and explore fisheries datasets through *ad hoc* metrics (maps, plots, tables...). To make it reusable, this work uses the current data structure that is promoted by the Coordinating Working Party on Fishery Statistics (CWP) as a standard to store gridded fisheries data which are multi-dimensional data cubes with following data structures:

- dimensions: time (period), area (lat / long coordinates of spatial objects), species, fishing gears, type of schools...
- variables: catch, fishing efforts, conversion factors, size-frequency data.

With an *a priori* knowledge of the data structure, dynamic reports that are domain specific can be set up to describe the main characteristics of fisheries datasets with metrics illustrated in Figure 6 such as:

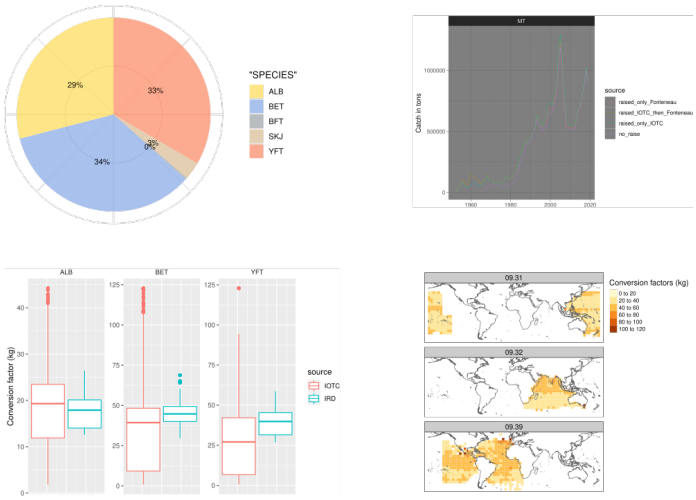

Figure 6: Example of metadata elements specific to the fisheries domain

- temporal coverage by using time series of catches, fishing efforts
- spatial coverage reported on CWP grids
- taxonomic coverage by using bar plots of pie charts

The main interest of this work is to provide multiple metrics (plots, tables. . .) to:

- quickly explore the content of fisheries datasets beyond usual, domain agnostic, metadata elements
- assess the quality and reliability of fisheries datasets
- point out some inconsistencies or errors.

Ultimately, such domain specific metadata are meant to be linked with more generic metadata described in previous sections (like Dublin Core, DataCite, and OGC standards) by relating the pdf report through http links or attaching it directly to the data (by adding it to the archive to which the DOI is assigned).

4. Results and Discussion

IOTC already makes some datasets public since years. However, some basic technical improvements can help to foster IOTC data discovery, tracking, citation and thus impact. These technical aspects mainly deal with an extra layer of interoperability to better connect IOTC datasets with other (spatial) data infrastructures meant to disseminate scientific data in general and sometimes

fisheries data in particular (e.g., FAO GeoNetwork and GeoServer). Such a technical work can be tackled by projects which bring the funds necessary to get the expected person-months. However, the IOTC secretariat will have to validate and drive this process to identify: (i) priorities for datasets to be described by rich metadata and (ii) the most relevant infrastructures to index these datasets. This is the reason why, in this paper, we recommend to make use of FAO SDI for fisheries and / or of widely used data repositories like Zenodo or GBIF to assign DOIs to datasets that are already public or others that are also supposed to be public. Regarding data, we recommend to repackage existing descriptions by mapping the content of IOTC web pages with OGC metadata standards that could then directly by published in FAO GeoNetwork. This approach would allow to get quickly published some metadata sheets describing catch, fishing effort, and size-frequency data. In addition, we recommend to publish other biological datasets like IO-RTTP data in GBIF data repository which better fits the need for biological / biodiversity data. Examples in this paper showcase what can be quickly achieved by re using a generic R collaborative workflow that can be directly managed by fisheries data managers in the short-term. Finally, domain specific metadata can also be automated (dynamic reports) when data structure datasets comply with CWP data structure.

Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Blue-Cloud project (Grant agreement No 862409). G2OI (Grand Observatoire de l'Océan Indien) ERDF project is co-funded by the European Union INTERREG V program, the French Republic and the Réunion Region.

References

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 multigr., 2017. URL https://www.documentation.ird.fr/hor/fdi:010071472.

J. Barde, A. Nieblas, E. Blondel, N. Bodin, S. Bonhommeau, E. Chassot, and T. Imzilen. Describing and accessing biological and tagging data. page 12 multigr., 2018. URL https://www.documentation.ird.fr/hor/fdi:010075959.

E. Blondel. ows4r: R interface to ogc web-services, May 2020a. URL https://doi.org/10.5281/zenodo.3860330.

E. Blondel. geonapi: R interface to geonetwork api, Aug. 2020b. URL https://doi.org/10.5281/zenodo.3975454.

E. Blondel. geosapi: Geoserver rest api r interface, Aug. 2020c. URL https://doi.org/10.5281/zenodo.3975455.

E. Blondel. OpenFairViewer: a FAIR, ISO and OGC (meta)data compliant GIS data viewer for browsing, accessing and sharing geo-referenced data, June 2022. URL https://doi.org/10.5281/zenodo.6652309. Funders/Sponsors: EC BlueBridge & Blue-Cloud projects; UN-FAO, IRD; INRAE.

E. Blondel, J. Barde, W. Heintz, and A. Bennici. geoflow: R engine to orchestrate and run geospatial (meta)data workflows, Nov. 2020a. URL https://doi.org/10.5281/zenodo.4275926. Beta release.

E. Blondel, J. Barde, A.-E. Nieblas, E. Chassot, F. Fiorellato, A. Ellenbroek, A. Gentile, and M. Taconet. The firms tuna atlas: a scalable open data portal for global tuna fisheries. page 15 multigr., 2020b. URL https://blue-cloud.org/sites/default/files/IOTC-2020-WPDCS16-22_-_FIRMS_Tuna_Atlas.pdf.

N. Bodin, E. Chassot, F. Sardenne, I. Zudaire, M. Grande, Z. Dhurmeea, H. Murua, and J. Barde. Ecological data for western Indian Ocean tuna [Data Paper]. Ecology, 99:1245–1245, 2018. ISSN 0012-9658. doi: 10.1002/ecy.2218. URL https://www.documentation.ird.fr/hor/fdi:010072870.

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

A. M. Guillou, N. Bodin, E. Chassot, A. Duparc, T. Fily, P. Sabarros, M. Depetris, M. J. Amande, J. Lucas, E. Augustin, N. C. Diaha, L. Floch, J. Barde, P. Pascual Alayon, J. C. Baez, P. Cauquil, K. Briand, and J. Lebranchu. Tunabio : biological traits of tropical tuna and bycatch species caught by purse seine fisheries in the Western Indian and Eastern Central Atlantic Oceans. Biodiversity Data Journal, 10:e85938 [18], 2022. ISSN 1314-2836. doi: 10.3897/{BDJ}.10.e85938. URL https://www.documentation.ird.fr/hor/fdi:010086075.

H. Murua, J. P. Eveson, and F. Marsac. The Indian Ocean Tuna Tagging Programme: Building better science for more sustainability. Fisheries Research, 163:1–6, Mar. 2015. ISSN 0165-7836. doi: 10.1016/j.fishres.2014.07.001. URL http://www.sciencedirect.com/science/article/pii/S0165783614002136.

P. Taconet, E. Chassot, J. Guitton, F. Fiorellato, E. Anello, and J. Barde. Data toolbox for fisheries : the case of tuna fisheries. page 23 multigr., 2016. URL https://www.documentation.ird.fr/hor/fdi:010067433.

P. Taconet, E. Chassot, E. Blondel, and J. Barde. Global datasets for tuna fisheries. page 10 multigr., 2017a. URL https://www.documentation.ird.fr/hor/fdi:010071471.

P. Taconet, E. Chassot, J. Guitton, C. Palma, F. Foirellato, E. Anello, and J. Barde. Global database and common toolbox for tuna fisheries. Technical report, Madrid, 2017b. URL https://www.documentation.ird.fr/hor/fdi:010071520.