Applying Open Science principles to make a better use of IOTC data

Julien Barde[*], Bastien Grasset[†], Emmanuel Chassot [‡], Sylvain Bonhommeau[§], Emmanuel Blondel[¶]

SUMMARY

In this paper, we further discuss known issues, achievements and challenges to improve IOTC resources management and we attempt to identify opportunities brought by Open Science principles to face them. For a regional organization like IOTC, we recommend to promote and implement these good practices internally at the Secretariat level as well as externally to expose and disseminate resources though multiple channels for the sake of users (CPCs or wider audience). Such practices appear to be instrumental as well when reusing external contributions or when sub-contracting or delegating tasks to CPCs. We also showcase how such good practices can be implemented "at no cost" in a simple and efficient way by relying on external infrastructures. As a demonstrator we showcase how IOTC size class datasets management can be improved to better feed external initiatives like the Global Tuna Atlas datasets update and related Shiny apps. We explain how this approach can be used beyond fisheries data management and visualization to foster the publication and reproducibility of additional research work like stock assessment models runs outputs.

KEYWORDS: Fisheries data, Open Science, FAIR data management principles, Reproducibility, Metadata, Interoperability, Data discovery, Data access, Fisheries, DOIs

[*]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, Av. Jean Monnet, 34200 Sète, France; julien.barde@ird.fr; Phone: +33 499 57 32 32 Fax: +33 499 57 32 15.

[‡]IOTC Secretariat (NFITD), C/O IOTC Secretariat, PO BOX 1011 – Blend Seychelles, Victoria, Mahé (Seychelles)

[§]Ifremer, avenue Jean Bertho, Le Port, Réunion, France

[¶]Food and Agriculture Organization (FAO) of the United Nations, Fisheries Division

1. Introduction

The Indian Ocean Tuna Commission (IOTC) collates and disseminates valuable materials and resources for marine research. Tuna and tuna-like fisheries play an important role in Ocean Observing Systems by providing Essential Ocean Variables (EOV) related to top predators, particularly in the open-sea ecosystems of the high seas. Generally, most of the available EOVs are physical and chemical variables. This is largely due to the underlying data collection and management infrastructures which are better aligned with open science and FAIR (Findable, Accessible, Interoperable, and Reusable) data management principlesWilkinson et al.. The lack of equivalent good practices in the biological domain is known, but it is not insurmountable. At IOTC, the main dissemination channel remains the website (https://iotc.org/). While it provides access to a valuable repository of fisheries data, the current structure limits data discovery and reuse. The website's navigation is not always intuitive, content appears underused, and usage statistics – such as page views or download counts – are not systematically tracked. A few years ago, IOTC began assigning Digital Object Identifiers (DOIs) to some resources via the Zenodo public data repository. However, beyond data, code, and document dissemination, Open Science fundamentally concerns the reproducibility of research which stresses the need to capture execution environments along with data and code. In the first section of this paper, without claiming to be exhaustive, we outline some of the main challenges that IOTC faces in improving data dissemination and supporting research reproducibility. In the second section, we suggest and discuss the potential interest of various technical solutions to address these challenges. In the third section, we provide examples to showcase how such a strategy could be implemented by leveraging existing public infrastructures and open-source tools that can (and should) be used by all stakeholders, including the IOTC Secretariat, Contracting Parties, and consultants.

2. Some challenges for IOTC data an other resources

From the perspective of the general public, IOTC resources are primarily made publicly available through its website which until recently was poorly indexed by search engines. Most of IOTC reports and publications, currently managed by the website's Content Management System, have no bibliographic references but need to be integrated into the FAO website in the forthcoming year. To date, only few documents have been assigned DOIs, such as those described by Nieblas et al. [2019] and more recently codelists and reporting forms via the (cf IOTC Zenodo community).

IOTC datasets are generally described by rich metadata, though these are not fully standardised, and can be downloaded directly from the website without restrictions. However, they are difficult to cite because they have not been assigned DOIs or standard bibliographic references. Time series datasets such as catch and effort data, are regularly updated – either by adding new statistical years or by correcting errors – but this is done by simply replacing previous files, without versioning to allow users to track or reproduce analyses. The metadata does not always describe provenance, including underlying workflows and code. Beyond datasets, IOTC also produces numerous documents, protocols, data collection forms, and scripts to handle these datasets, all of which are valuable resources that could benefit from sharing and versioning. Recently, the Secretariat has begun releasing scripts supporting projects linked to data curation, management, and dissemination through repositories hosted on GitHub (see IOTC organization). Working Parties and related datasets and papers also provide valuable resources that can be referenced and shared more efficiently. Some of these resources are already referenced by CPCs in external information systems.

In summary, most IOTC resources remain discoverable only via the website, making them underused. When accessed, users encounter non-standardised metadata, data structures, or formats, which hinders proper citation and reuse. Although IOTC relies on a skilled team and efficient information systems, the current configuration of its data dissemination system makes it difficult for data producers and users to ensure reproducibility of work conducted by the Secretariat, CPCs, and the general public.

Reproducibility – the ultimate goal of Open Science – depends on sharing and versioning open data, open-source code, and providing portable execution environments. These principles can serve as milestones for a way forward. Key assets to support this include:

- metadata :
  - Core metadata elements should be provided by IOTC, including standardised bibliographic references to enable proper citation by users and better track use, also to enable better migration from a system to another,
  - Provenance metadata can foster reproducibility by documenting the lineage of datasets, stock assessment model calibration, parametrization outputs and related workflows...
- data:
  - Prioritise standardised data structure promoted by the community
  - Prioritize open, interoperable data formats

- code:
  - Promote open-source code,
  - Describe and make environments restorable to reproduce some key work (e.g. stock assessment)
  - When possible, build and provide directly reusable computational environments to ensure reproducibility.
- Keep on exposing more resources in other infrastructures meant to implement Open Science and FAIR principles...

In the next section, we present possible technical solutions to address the issues outlined above.

3. What Open Science practices for IOTC resources ?

When looking at the main pillars of Open Science (see Figure 1), we can see that IOTC resources enumerated in section 2. are mainly dealing with following ones:

- Open Scientific knowledge
  - Scientific publications: e.g. working papers made public on IOTC website
  - Open research data: e.g. datasets made public on IOTC website
  - Open Source software and code: e.g. dynamic reports (R markdown or Jupyter notebooks), stock assessment models runs (e.g., Gitub repository)
- Open scientific infrastructures
  - Physical machines in Seychelles (e.g. database servers)
  - Virtual infrastructures (e.g. HPC services provided by CPCs, FAO, research projects or private infrastructures like Google, Amazon Web Services...)

In this context, Open Science good practices mainly consist in implementing some simple technical solutions expected to have a high impact:

- assigning generic DOIs by recording data on public data repositories (Zenodo, GBIF/OBIS...) and keeping track of the different versions disseminated over time by making use of versioned DOIs (Figure 2). This work has been initiated for code lists and reporting forms, using the Zenodo entry

Figure 1: Open Science pillars according to UNESCO

form interface, and should be extended and automated with workflow orchestrators such as the geoflow R package Blondel et al. [2020],

- implementing standards to foster resources discovery and interoperability:
  - standards for metadata : e.g. Datacite when assigning DOIs, EML for biodiversity and biological data to be displayed on GBIF /OBIS or OGC 19115 for spatial data to be displayed on e.g. FAO GeoNetwork, codemeta for code description on Software Heritage...
  - standards for data structure : e.g. CWP standards for fisheries, OGC standards for spatial data, GBIF / TDWG Darwin Core format Wieczorek et al. [2012] for biodiversity and biological data
  - standards for data formats : e.g. CSV instead of excel, cloud optimized formats like parquet...
- publishing and versioning code on a forge (e.g. GitHub), main releases being also assigned DOIs (e.g. with Zenodo & Software Heritage: cf example),
- ensuring reproducibility on different infrastructures by:
  - promoting open formats (e.g. CSV instead of xlsx)
  - using free and Open Source Software (e.g. Postgres instead of Oracle), open programming languages (e.g. R, Python...)
  - providing the code along with snapshots of virtual environments: e.g. R with renv package Ushey and Wickham [2025], Python with Conda,

- Containerizing applications by using e.g. Docker or Singularity (recommanded for HPC)
- promoting online environments for collaborative work (VRE including RStudio, Jupyter Lab..), or to host various applications by deploying containers.

It is worth noticing that for being FAIR, resources do not necessarily have to be open when recorded on data repositories (other status being restricted, embargo or closed).

On the same line, IOTC should not accept to fund, host or disseminate some work achieved by scientists (whether the work is an in-kind contribution or sub contracted) without imposing a certain level of maturity and reproducibility which fosters the deployment, maintenance and trust.

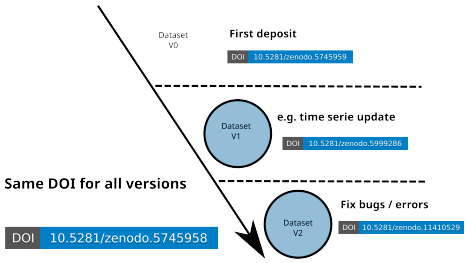

Figure 2: Assigning DOIs to different versions other same dataset

In the following section, we have chosen two examples showcasing how both IOTC resources and code provided either by CPCs or consultants might be better managed and what the expected benefits would be.

4. Expected benefits : some examples for IOTC ?

By applying Open Science principles, IOTC can quickly fix most of aforementioned issues and take advantage of this new configuration:

- bulk publishing: set up workflows Nieblas et al. [2019], Blondel et al., Blondel and Barde to upload an unlimited number of IOTC resources along with their metadata on data repositories (e.g. datasets, working papers...) providing an efficient backup and facilitating e.g. website migration when needed
- delegate usage statistics tracking to other infrastructures:
  - number of views and downloads can indeed directly be managed by e.g. Zenodo or GBIF if IOTC Web site simply reuses these external

links instead of using internal ones: e.g. from a DOI provided by Zenodo it is possible to create permanent links to e.g. download or view a specific file directly in Zenodo

  - publishing resources on such infrastructures foster the citation by individual researchers by providing standardized bibliographic references in addition to the DOIs also used as primary data to track the number of citations of a specific resource in the scientific literature at a global scale,

- reproducibility enabled anywhere : e.g. generic Shiny apps hosting services, HPC / Parallel computing

This set of simple practices would drastically improve the findability and access to IOTC resources, not being restricted anymore to the only website of the organisation, and being efficiently indexed on the Web.

For demonstration purposes, we applied several Open Science principles to demonstrate how sizeclass data could be more effectively managed.

4.1 Size class datasets

Within the context of Global Tuna Atlas, we aim to integrate size-class data alongside catch and efforts datasets. We also seek to provide reproducible workflows and visualisation tools (e.g. Shiny apps) that can be deployed in different computing environments. Currently, the structure of size-class data varies by species and does not follow standards promoted either by the Coordinating Working Party on Fishery Statistics (CWP) or Taxonomic Databases Working Group (TDWG)/ Global Biodiversity Information Facility (GBIF). In addition, datasets are disseminated in different data formats (e.g. some as CSV files, others as xlsx), a situation similar to that observed in other tuna RFMOs. The main steps of our workflow, implemented in the R programming language, consist in:

- Download the data from IOTC Web pages
- Transform and standardise the data structure by adopting the Darwin Core standard with its Measurement or Facts extension,
- Customise a generic Shiny app to display size- lass samples through interactive maps and plots,
- Ensure reproducibility of the workflow and the code deployment by providing a snapshot of the R environment (including versions of the language and associated packages) and by containerising the Shiny application in a Docker image,

- Publish and archive the code on a collaborative platform (e.g. GitHub), assign a DOI to a first release, and ensure long term preservation,
- Deploy a containerised Shiny application across multiple servers, hosted on different (physical or virtual) infrastructures, to validate both portability and reproducibility of the workflow (see screenshot 3).

In practice, for demonstration purposes, the code is hosted in a GitHub repository that is connected to Zenodo to assign DOIs to its main releases, with an additional record registered in Software Heritage. The archive includes an renv snapshot of the R execution environment, as well as a Docker image in which the Shiny application is containerised to ensure reproducibility and streamlined deployment. The container has been deployed on several infrastructures (e.g. D4Science, the IRD server, EDITO Datalab) demonstrating a satisfactory level of robustness and portability.

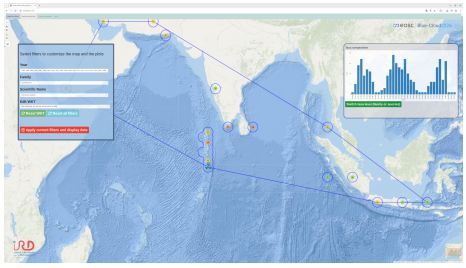

Figure 3: Assigning DOIs to different versions other same dataset

However, it is important to note that the IOTC datasets must be stored within our containers to ensure reproducibility, as these datasets have not yet been assigned DOIs directly by IOTC. Other Docker images of Shiny applications, based on DOIs of catch and effort datasets and thus ensuring full reproducibility of the images have already been deployed within the context of Global Tuna Atlas.

4.2 Stock assessment model

Open Science principles are also particularly relevant in the case of stock assessment when applied to model calibration, parameterization and execution as well standardising and visualising data outputs. In this case following practices would be key:

- Provenance metadata: to describe how model outputs have been generated and can be reproduced

- Standardisation of the data structure of the stock assessment model outputs would improve its reuse and exploration by e.g. easily building generic Shiny apps such the one in Figure to visualise and explore the content of models outputs (see Figure 4)
- Containerisation of code by using e.g. Docker images or Singularity containers (which becomes a de facto standard for parallelization High Performance Computing (HPC))
- Online execution of stock assessment models: either on virtual infrastructures by using Virtual Research Environments providing usual IDE like RStudio server Nieblas et al. [2017b], or directly on HPC

Figure 4: An R Shiny app to visualize the results of current or past model runs, allowing runs to be overlaid on interactive plots to compare different model parameterizations.

If funded, the past work described in Nieblas et al. [2017a] might soon be udpated and upgraded.

5. Conclusion and outlooks

Open Science and FAIR data management best practices are now widely regarded as standard recommendations for guiding scientific work Wilkinson et al.. Their adoption improves the quality of data and code publication methods, while enhancing the discoverability, accessibility, and reproducibility of research outputs. Moreover, recent studies have shown that applying these principles can significantly increase data citation rates Colavizza et al. [2025]. However, reproducibility should not be limited to the ability to repeat an analysis on a single personal computer. Rather, it implies that the work can be replicated and deployed across different infrastructures, including institutional servers and, ultimately, high-performance computing (HPC) systems. Currently, several methods allow researchers to capture and reproduce execution environments, such as containerisation (e.g. Docker) and environment management tools (e.g. renv), in addition

to sharing data and code. These approaches are essential for ensuring the reproducibility and reuse of scientific work across different computing systems, facilitating deployment on multiple servers, and supporting efficient data and code management. They are especially relevant for regional organizations like IOTC and should be imposed whenever possible.

Acknowledgements

This work has received support from the current HORIZON BlueCloud 2026 (grant number 101094227) and past European OpenAIRE-Connect H2020 research project (grant No. 731011).

References

E. Blondel and J. Barde. zen4r: R interface to zenodo REST API. URL https://doi.org/10.5281/zenodo.8365600.

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

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

G. Colavizza, L. Cadwallader, and I. Hrynaszkiewicz. An analysis of the effects of open science indicators on citations in the french open science monitor, 2025. URL https://arxiv.org/abs/2508.20747.

A. E. Nieblas, S. Bonhommeau, T. Imzilen, D. Fu, F. Fiorellato, and J. Barde. An online tool to easily run stock assessment models, using SS3 and YFT as an example. In 15th Working Party on Billfish (WPB15), volume 15 of IOTC Proceedings, page 16, San Sebastián, Spain, Sept. 2017a. IOTC. URL https://iotc.org/documents/online-tool-easily-run-stock-assessment-models-using-ss3-and-swo-example. https://iotc.org/meetings/15th-working-party-billfish-wpb15.

A.-E. Nieblas, B. 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. In 13th IOTC Working Party on Data Collection and Statistics, volume 13 of IOTC Proceedings, page 12, Victoria, Seychelles, 2017b. IOTC. URL http://www.iotc.org/documents/standardization-metadata-data-formats-access-protocols-and-statistical-visualization-ss3 https://iotc.org/meetings/13th-working-party-data-collection-and-statistics-wpdcs13.

A.-E. Nieblas, F. Fiorellato, E. Blondel, P. DeBruyn, and J. Barde. Assigning dois to publically-accessible iotc documents and their publication on the open-access data repository zenodo. page 27 multigr., 2019.

K. Ushey and H. Wickham. renv: Project Environments, 2025. URL https://github.com/rstudio/renv. R package version 1.1.5.9000.

J. Wieczorek, D. Bloom, R. Guralnick, S. Blum, M. Döring, R. Giovanni, T. Robertson, and D. Vieglais. Darwin core: An evolving community-developed biodiversity data standard. PLOS ONE, 7(1):1–8, 01 2012. doi:

10.1371/journal.pone.0029715. URL https://doi.org/10.1371/journal.pone.0029715.

M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons. The FAIR guiding principles for scientific data management and stewardship. 3(1):160018. ISSN 2052-4463. doi: 10.1038/sdata.2016.18. URL https://doi.org/10.1038/sdata.2016.18.