Benefits of assigning DOIs to IOTC documents and data

Lucia Pierre[1] and Julien Barde[2]
[1] Indian Ocean Tuna Commission (IOTC), Victoria, Mahe, Seychelles
[2] MARBEC, Univ Montpellier, CNRS, Ifremer, IRD, France, julien.barde@ird.fr

Abstract: Only few organizations worldwide can ensure the long term storage and access for their numeric resources. Indeed this requires a replication of data in multiple sites within secured data centers. Consequently, even in scientific organizations, the rate of data loss is important and this situation is no longer accepted by funding agencies which are requiring research projects to implement data management plans to ensure that the funded data and resulting information and knowledge will not be lost. In addition, beyond economical considerations and for obvious ethical reasons, data have to be made accessible along with publications to enable the review and the reproducibility of the scientific work. To be found once published, resources also have to be described with standardized discovery metadata which foster interoperability on the Web. For years, multiple data repositories have offered free services for the long term storage of any kind of numerical resource. These repositories also assign unique identifiers (Digital Object Identifier; DOI) which foster their reuse and citation worldwide by complying with standardized metadata. In practice, scientists are already used to getting DOIs assigned to their articles by scientific journals but still few of them are assigning DOIs to other products like their data, code or reports. In 2019, a first work has been driven in collaboration with IOTC to start assigning DOIs to a selection of IOTC reports. In this paper we present a summary of the past work results and make some recommendations to update it by assigning new DOIs to other kinds of IOTC resources (e.g., datasets, working papers).

Keywords: DOIs, metadata, interoperability, workflow, FAIR data management

1. Introduction

IOTC is generating an important number of working papers, reports and datasets (annually around 400 meeting documents and reports, and 90 datasets) that are widely used by the scientific community. However, these resources are not, properly speaking, accessible in the long term, replicated in different locations, and are not assigned any Digital Object Identifier (DOI) to track their use and increase their dissemination worldwide. The impact of IOTC resources that are shared is thus lower than it could be with better indexing and citation strategies. Indeed they are already referenced, well described and publicly accessible but only by browsing the IOTC Web site. However, without replication elsewhere, the risk of data loss remains high. Moreover, the IOTC Web site is poorly referenced on the Web which makes its resources hardly findable when using search engines. These different issues can easily be tackled if IOTC makes the decision to assign DOIs to both past and future resources, in line with FAIR principles (Wilkinson et al..). In 2019, the OpenAire-Connect H2020 project provided some funding to better reference fisheries sciences resources in data repositories which are connected to global documentation systems managing DOIs, like the Datacite consortium which is perfectly indexed and harvested by Web search engines. This first work has enabled the publishing of some of IOTC documents in the Zenodo data repository. ZENODO, a free and open-access platform for preserving and sharing research output, is ideal for the IOTC to use as a repository for the datasets and documents. CERN, the European Council for Nuclear Research founded in 1952, hosts ZENODO, which is embedded in the IT Department, Collaboration Devices and applications Group, digital Repositories Section (ZENODO infrastructure), and funded by the European Commission for open access to science datasets and publications. Once a dataset or

document is in ZENODO, a unique DOI is assigned and remains the publication's universal identifier. ZENODO ensures that resources will be stored for at least 25 years, and once in the repository, they cannot be removed but can be updated whenever needed. ZENODO also provides other essential features like rich and standardized metadata to accurately describe the ressource, a preview facility, statistics (downloads, views) and citation. Zenodo is just one of the multiple free data repositories supported by funding agencies but it is one of the few which also provides an efficient API to automate the publication of multiple documents through programmatic access, without having to edit metadata manually. In the second section we present the result of the first initiative conducted in 2019. In the third section, we recall the main kinds of documents managed by IOTC and we discuss the opportunity and possible strategies to assign DOIs to all kinds of IOTC resources. In the fourth section we discuss the specific case of datasets and we finally discuss some outlooks before concluding.

2. Summary of past work

Figure 1 presents a screenshot of an IOTC document published on Zenodo in 2019. When uploading such a document, some standardized metadata elements need to be filled. This can be achieved in a few minutes and the resulting Zenodo Web page immediately displays a DOI along with the filled metadata and a preview of the document. Tracking statistics (view and downloads) are also made available as well as a standardized bibliographic citation.

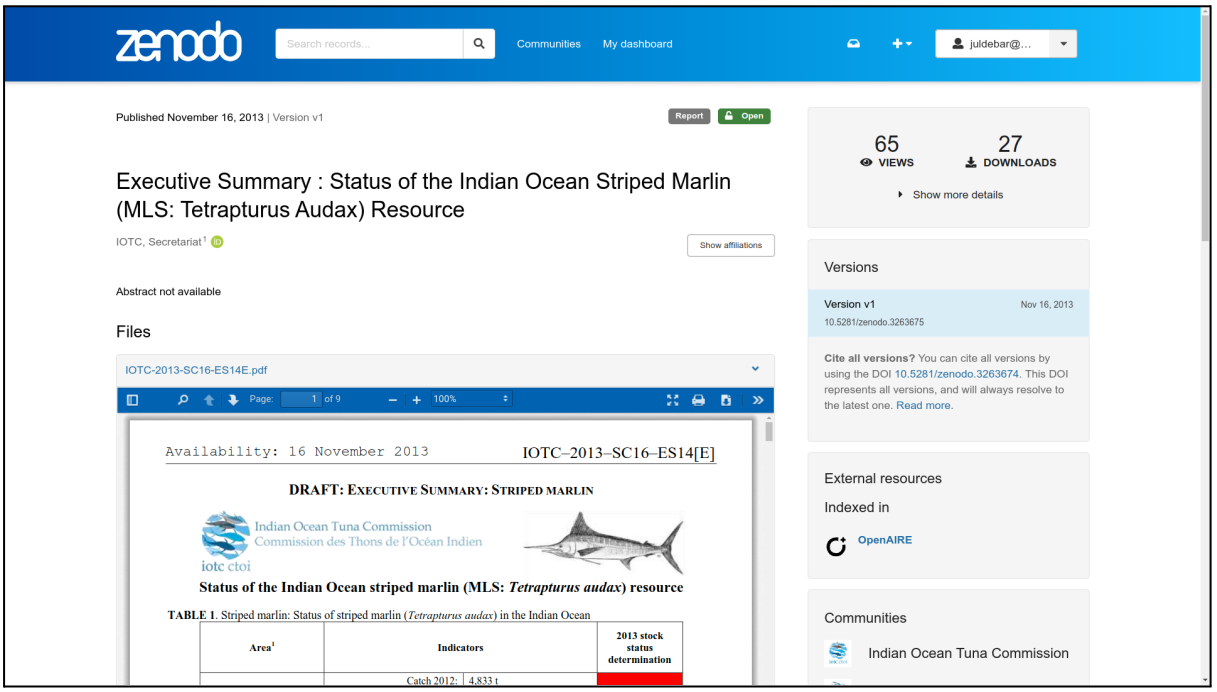

Figure 1. Screenshot of an IOTC report published on Zenodo

The OpenAire-Connect H2020 project funded this preliminary work which set up a first workflow to publish a selection of IOTC resources in Zenodo data repository. By doing so, a few hundreds of IOTC reports were assigned a DOI and these DOIs were displayed back on the IOTC Web pages describing the reports. Figure 2 shows the same resource as Figure 1 on IOTC Website enriched with a DOI. The same metadata was thus replicated on both IOTC Web site and Zenodo data repository. Such a process was too long to be managed manually and we used the Zenodo API instead.

Figure 2. Screenshot of an IOTC report Web page enriched with a DOI assigned by Zenodo

Indeed, the IOTC Web site references thousands of resources with a few metadata elements which are not standardized and are poorly describing the content of these resources. In such a case, the upload of resources, editing of metadata and assignment of DOIs in data repositories like Zenodo has to be automated by making use of an API. Zenodo API can be used with different programming language libraries. We used the zen4R R library (Blondel and Barde) to assign DOIs to a set of selected IOTC resource by using a basic workflow executing the sequence of following steps (illustrated in Figure 3):

- extraction of metadata for selected resources from IOTC Website from a SQL query
- SQL query results transformation to comply with geoflow R library input format
- assignment of DOIs in Zenodo with geoflow R library
- Adding DOIs link in the IOTC Web page of these resources

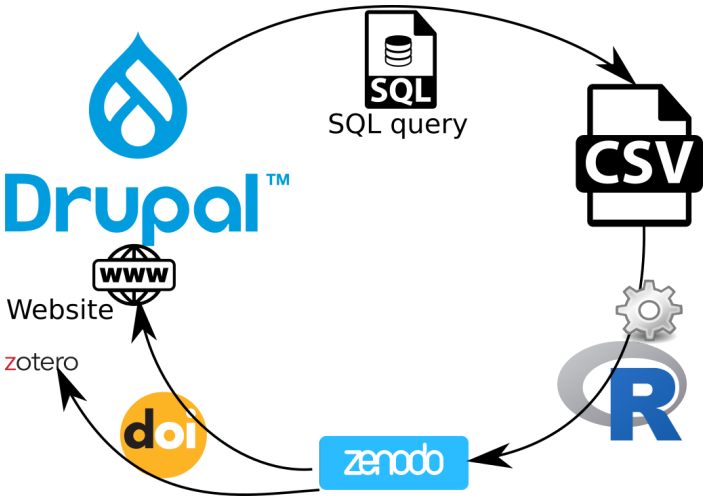

Figure 3. R workflow extracting IOTC resources from the Website and assigning them DOIs on Zenodo

All IOTC resources have been also gathered within a dedicated community as illustrated by Figure 4 which offers a focus on IOTC resources with dedicated filters to explore its content.

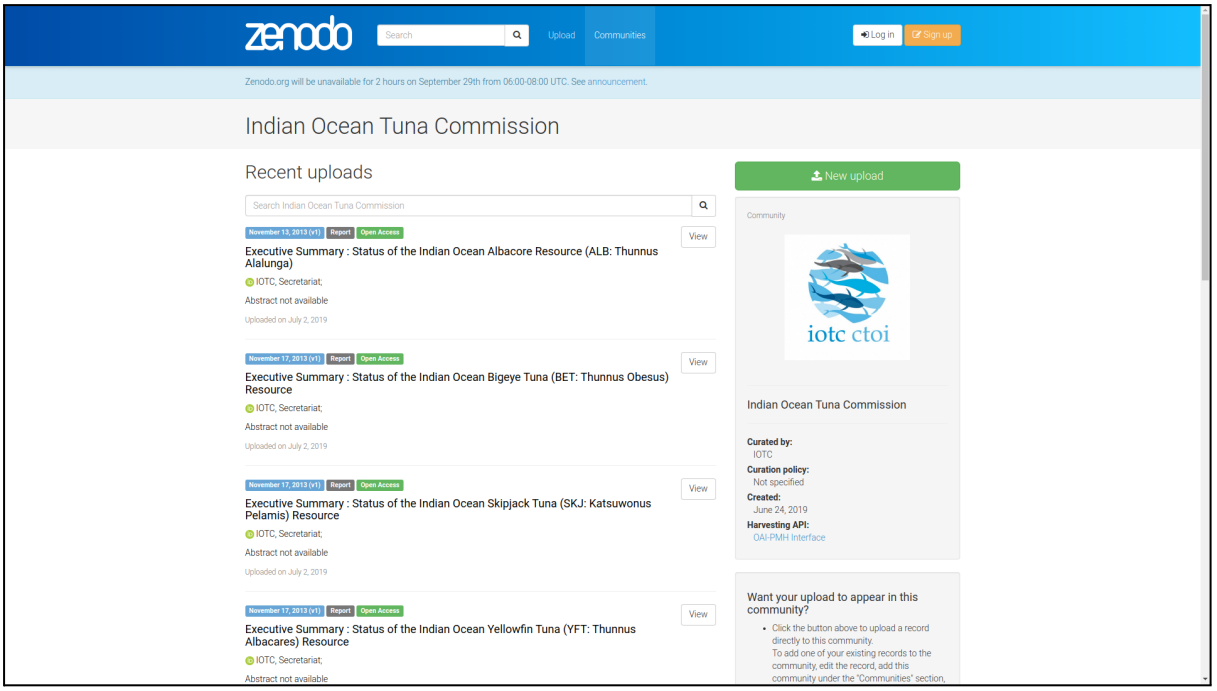

Figure 4. IOTC community created on Zenodo

This first attempt has demonstrated the feasibility and the interest of using data repositories to ensure long term access and dissemination of IOTC resources. This replication in data repositories provides added value for scientists to better share their work and ensure its proper citation when reused. Indeed IOTC documents with DOIs are now better indexed by search engines. Data loss is extremely unlikely since Zenodo ensures that all IOTC resources are duly managed in line with state of the art storage technologies. In addition, as Zenodo is the landing page of the DOI, users are also better redirected to the IOTC Web site which increases its visibility.

However, in this previous work, due to a lack of time to solve technical issues, key metadata elements (e.g. Abstract, Keywords..) were often missing. This issue has reduced the efficiency of data discovery services offered by Zenodo and this should be improved.

A working paper (Nieblas et al. 2019) details the whole process implemented in 2019 to achieve this work which could easily be updated. However, beyond technical aspects, this first attempt also raised multiple questions, in particular :

- IOTC needs to establish a strategy to figure out what already published resources should be assigned a DOI- and who could review and validate these workflows
- IOTC has to define a specific strategy to assign DOIs for future resources (a posteriori assignment of DOIs being more complicated to manage)
- even though discussed for some time, no DOIs were assigned to IOTC datasets. In the past, the case of IOTC data like RTTP has been discussed but no decision was made.

In the following sections, we discuss a set of use cases and some options to define new strategies to assign DOIs to other IOTC resources.

3. Overview of IOTC resources

The IOTC Secretariat provides resources that are either produced by the Secretariat or by meeting participants:

- the IOTC Secretariat produces datasets and documents annually for various meetings. Some documents are updates of activities, trends, or progress. The documents or datasets updated annually are overwritten on the IOTC Web site and past versions are not visible anymore for users to follow up on the trend or development in the IOTC fisheries. Only advanced users with a good knowledge of the layout of the IOTC web page would be able to browse and locate historical datasets and documents. The annual publications are in the form of (i) datasets, (ii) scientific papers, (iii) reports, and (iv) other information papers. For some scientific and information papers, related datasets are not in the public domain, although presented during the meetings. Indeed, part of IOTC resources are restricted or closed.
- the meeting participants present various scientific and informative documents which are, most of the time, accessible only through the IOTC meeting web pages (open access). However, participants are free to share these documents through other dissemination channels (e.g. HAL for IRD or Archimer for Ifremer..) and also free to assign a DOI on any data repository of their choice (Zenodo, Pangaea, GBIF, Seanoe..).

Besides publications for meetings, the Secretariat produces information documents which are made available for the public like reference documents, datasets on the level of tuna captured annually in the Indian Ocean or historical information:

- Reference documents:
  - Guidelines for reporting data;
  - Reference catalogue;
  - Guidelines on implementing the conservation and management measures;
  - Glossary for scientific terms;
  - Glossary of terms for implementation of conservation and management measures
- Datasets:
  - Project data – Tagging data
  - Yearly captured data for tuna and other species
  - Spatial fishing effort catch data
  - Spatial size frequency data for major IOTC species and bycatch species
  - Fishing craft statistics and fishing capacity report
  - Species conversion factors
- Historical information:
  - Collection volume of meeting proceedings
  - Project data sets – sampling data collection
  - Manuals – data collection
  - Expert consultation reports

It is thus customary for all the resources related to the IOTC activities to only be available on the IOTC website, although not always very visible or easily accessible to readers or scientists. The IOTC publications, if not published on another platform, do not hold a digital identifier of any form (either DOI, ISBN or others from relevant publishers), which makes them hardly findable by the general public and hardly citable when reused. Furthermore, in upcoming years, the resources of IOTC

Secretariat will probably have to use another software if a modification or a complete restructuring of the website happens and this could generate some data or documents loss.

As a publisher IOTC could either, like scientific journals, unilaterally decide to assign a DOI to IOTC resources (including working papers presented in the working parties) or IOTC let scientists decide whether or not they want a DOI. By using free data repositories like Zenodo, IOTC doesn't have to be one of the organizations registered in the DataCite consortium.

3.1 Proposed IOTC resources to be assigned a DOI

Resources stored in data repositories are not necessarily in the public domain (open access). Access can be either open or restricted (with or without embargo), and even closed. However, to be more findable, all resources (even with a restricted access) should be properly identified through a digital identifier and described with standardized metadata. Different platforms provide digital identifiers to various types of publications, but not all are freely available or with open access like Zenodo (e.g. Dryad, Figshare..). Using such repositories for IOTC resources would benefit and facilitate the discovery and access to several of the IOTC datasets and documents without browsing the IOTC web pages.

IOTC publications which could be published in the ZENODO repositories are:

- **Datasets for IOTC scientific working parties**: updated annually under the same name. With ZENODO, each new version of the same dataset can be stored under the same general DOI and be assigned a specific DOI. All versions are then listed on the same page. This would give scientists and other data users better access to historical updates and enable specific metadata for each version.
- **Regular datasets**: yearly capture data, spatial fishing effort, size frequency and fishing capacity. On ZENODO, the datasets could also be stored under one page per data type listing all versions updated annually. The main description in the metadata would remain the same, with modifications in case of any changes. With DOI, users could access the historical revised datasets, following the trends and updates through different versions.
- **IOTC revised documents for IOTC scientific working parties**: annual updates with the same name. In this case, a single general ZENODO DOI could list the different versions for the annual updates within a single page, each update being assigned a version DOI.
- **Reference documents**: produced in the form of booklets or reference guides (glossaries, guidelines, catalogues.) These reference documents are valuable and not updated as frequently as the datasets.
- **Historical information**: IPTP collection volumes and sampling programmes. The historical information is unknown to scientists unless cited in papers. This valuable resources are thus difficult to find and thus access and reuse. They will be more visible to scientists and other data users by assigning them a DOI. In circumstances where raw data of the programmes are available, with ZENODO, they could be published also.
- **Tagging data**: a valuable dataset for scientific research which currently has limited visibility on the IOTC website. Access is still restricted but could now be made open by considering a long embargo period. Moreover, some standardized discovery and usage metadata have been created by past projects to better find and reuse this dataset. As ZENODO allows the publication of resources with different access levels

(including open, restricted or closed access), a hybrid approach could be chosen: for example, display of the research outcome could be made public, and users interested in the detailed data could request access to the raw data (being assigned a specific DOI with restricted access).

- **meeting participant papers:** working papers are important scientific contributions that could also be more visible. In this case, the participants should be given the opportunity to publish their paper in an open-access platform, which could provide a DOI, preserve, curation and sharing of their research. For example, either participants could register their documents for digital identifier ahead of submitting to the IOTC meetings or IOTC could obligate the use of ZENODO for publishing meeting documents by participants, with sets of required metadata.

3.2 Methods and tools for assigning DOI to IOTC resources

According to the kind of resources presented in the previous sub-section, the strategies and validation process might thus differ. Indeed, some resources are directly produced by the secretariat (e.g. data, scientific reports) when others are produced by scientists (e.g. working papers). According to the use cases, the method and the review process can be adapted. DOIs could for example either be assigned programmatically by IOTC or manually by scientists, or by using a mixed approach.

The most basic method consists in making use of Zenodo GUIs to manually upload the resource, describe it with required and optional metadata elements. This can be done by anybody as long as the user has the right to assign the DOI on behalf of the authors. However, this approach is time consuming and can not be used to publish thousands of historical resources that are already online. Indeed, resources managed manually have to be processed one by one and metadata elements already existing in the Web site have to be copied and pasted in the Zenodo repositories.

This is the reason why we recommend setting up reproducible workflows for each kind of IOTC resource which can be assigned a DOI. A first workflow has been successfully implemented in 2019 (see Figure 3) to publish IOTC reports in Zenodo and newer versions have been updated and improved since. In particular, in the context of G2OI ERDF INTERREG projects, we have successfully tested and validated workflows created with geoflow and zen4R R libraries which facilitate the automated publication of numerous resources from a CSV file. However, in the case of IOTC, this CSV file has to be generated first by extracting existing metadata from IOTC webpages. Some additional metadata elements could also be enriched thereafter by reusing the content of the resources (e.g. Abstract, Keywords..). The Figure 5 illustrates examples of workflows to create FAIR services on top of IOTC resources management systems.

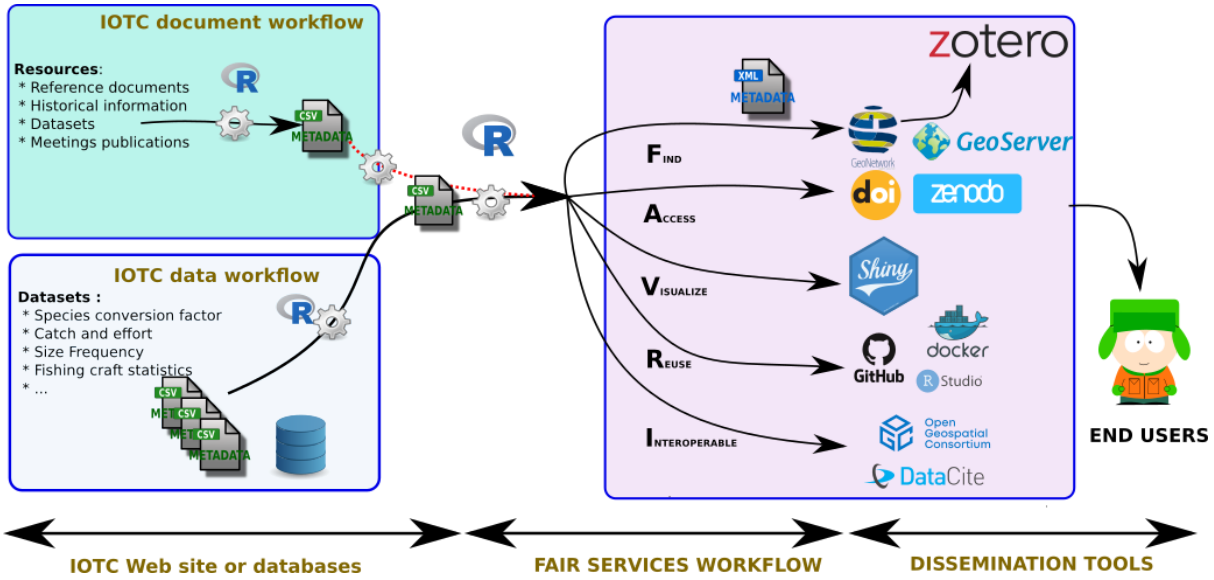

Figure 5. Resources publication workflow

In the specific case of datasets, additional options exist to improve the metadata and data publication workflow which are described in the next section.

4. The case of IOTC datasets

IOTC datasets are less numerous than documents and are usually very well described and accessible in different ways:

- Some datasets have been publicly available for years without any restrictions: CE...
- Some other datasets are available with restricted access

However, even if the quality of the metadata content is good, the format used to package metadata is just free text (html tags) which doesn't enable real interoperability with other systems. The Figure 6 shows a screenshot of the IOTC Web page describing available datasets.

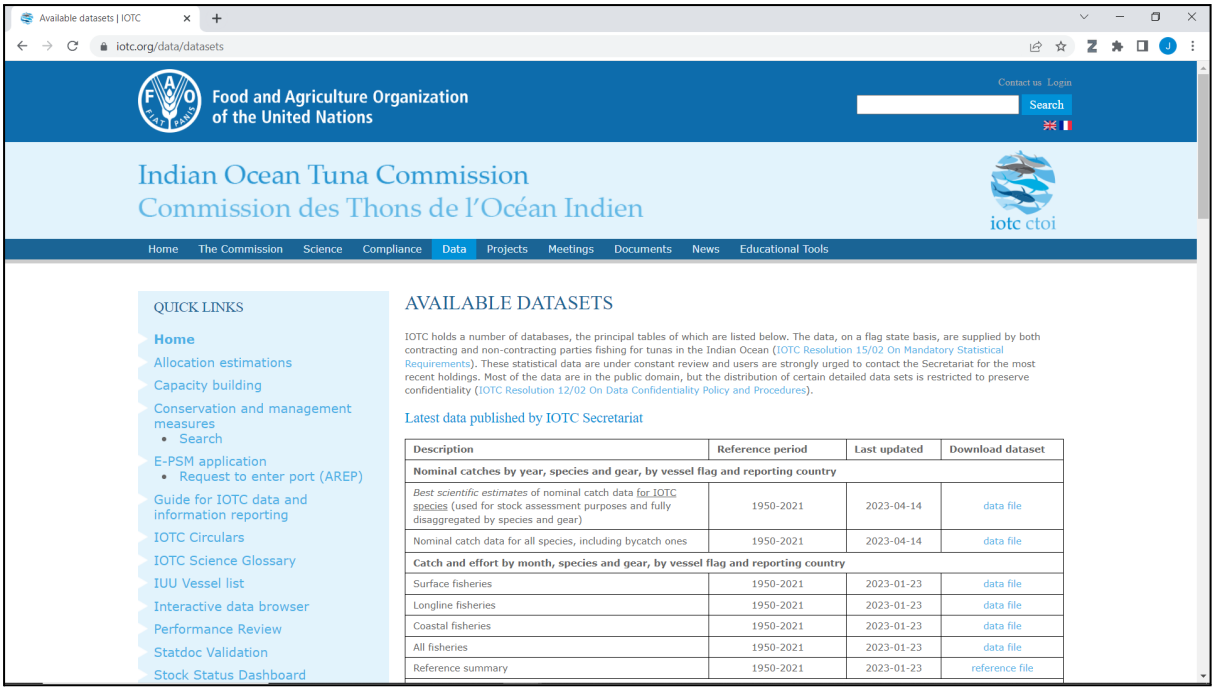

Figure 6. IOTC Web page describing available datasets

The workflows presented in Figure 3 and Figure 5 show how IOTC metadata can easily be extracted from the IOTC Website or databases and transformed into standardized metadata and data format. Once standardized, metadata can then be published in any standardized infrastructure and data access can be open, restricted or closed. In addition, by browsing data, important metadata elements can also be inferred and automated (e.g. spatial extent or temporal extent of the dataset). Figures 7 and 8 shows example of metadata extracted either from IOTC Website or directly generated by browsing the data, transformed into standardized metadata (e.g. OGC for spatial data, EML for GBIF or Datacite for DOIs) and published in data infrastructures which improve data discovery and access services (e.g. GeoNetwork and Zenodo). Indeed, data publication can be synchronized in multiple data infrastructures beyond data repositories. Assigning a DOI is probably the best way to disseminate resources but other metadata and data standards exist which are made for more specific domains targeting specific communities of users. For example, standards for species occurrences are widely used in the biodiversity domain (e.g., TDWG standards used by GBIF) or OGC standards used to improve the management of spatial data. Complying with these standards does not represent an additional work since the kind of workflows we implement enable to synchronize the publication of metadata and data not only in repositories like Zenodo but also in other information systems like spatial data infrastructures (e.g., FAO GeoNetwork or Geoserver) which are widely used, in particular by ocean observatories.
In the framework of past and current projects (e.g., ERDF INTERREG G2OI or Horizon BlueCloud 2026 european research projects) we also illustrated how IOTC data could be published in other channels.
Figure 8 shows an example of IOTC dataset whose metadata has been extracted from IOTC Web site, converted in ISO 19115 / OGC standard for geographic information and loaded in GeoNetwork by using a standard API.

Such metadata catalogs are widely used worldwide (e.g., FAO, Ifremer, CSIRO, IRD..) and can be interconnected.

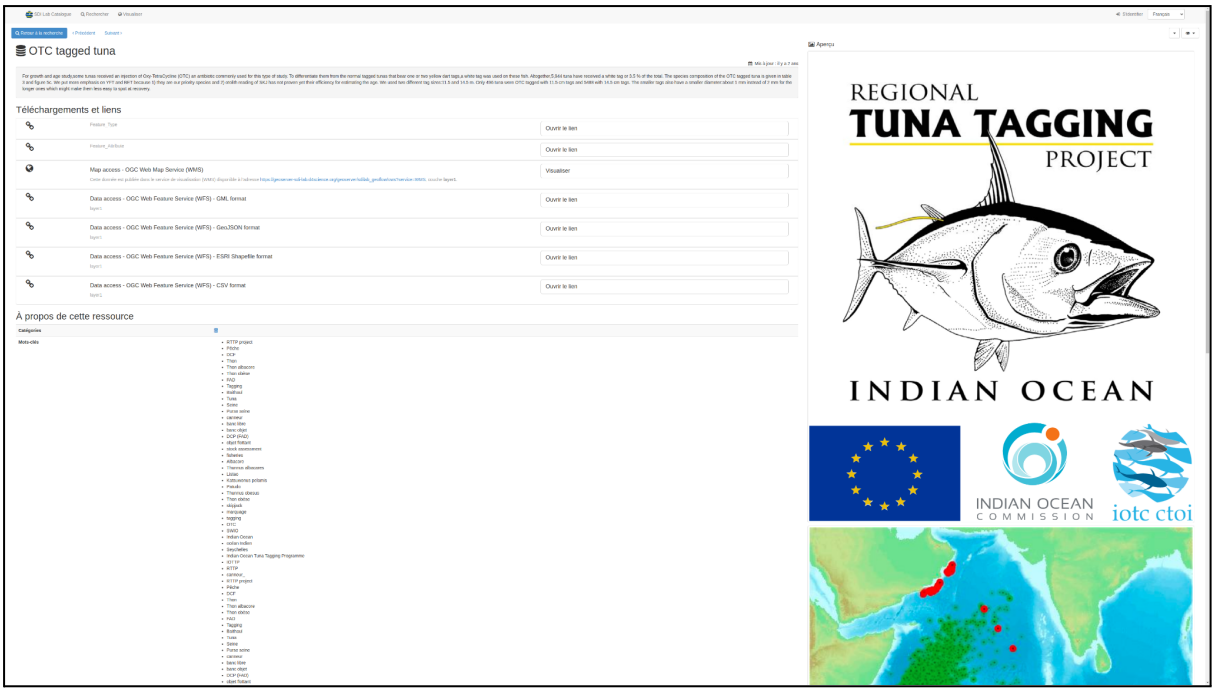

Figure 7. RTTP data described with OGC metadata published on GeoNetwork metadata catalog

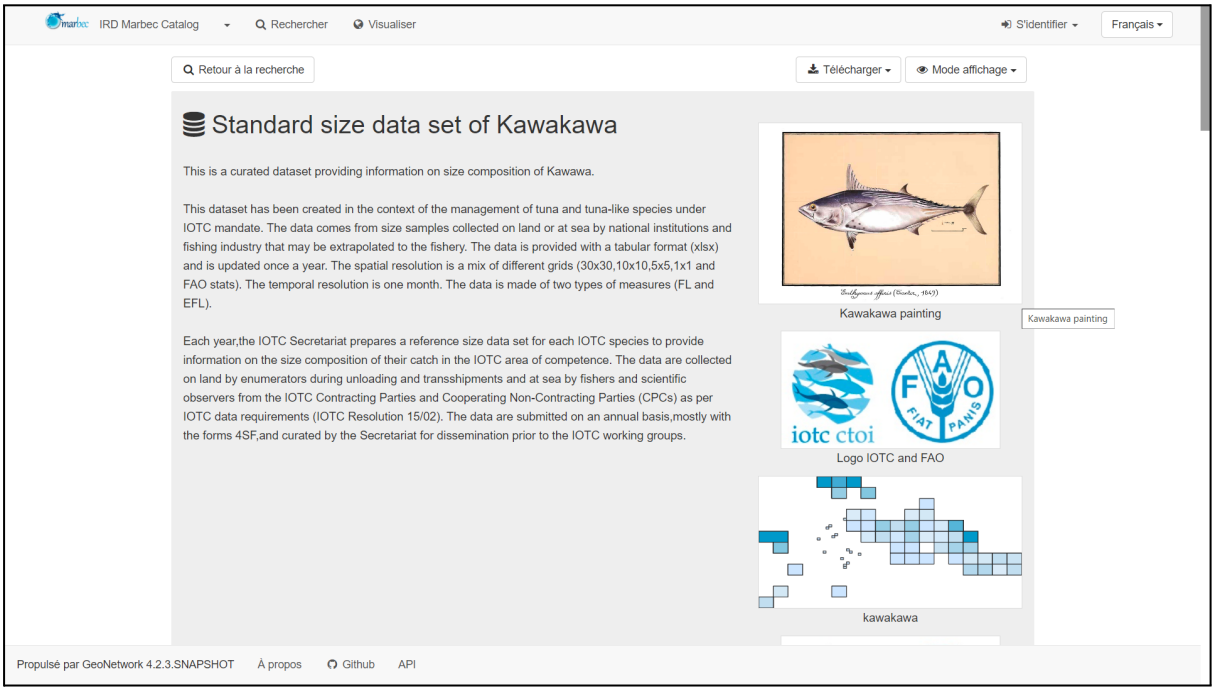

Figure 8. IOTC size class dataset published in a GeoNetwork metadata catalog

To summarize, as explained in section 3, datasets should be assigned a DOI but also disseminated through multiple channels to foster data discovery and reuse. Datasets that are regularly updated should also keep the same DOIs over the years by making use of version DOIs to identify the different versions of the same resource over its life cycle, e.g. :

- time series (catch, fishing effort, size class..)

- nominal catch and effort time series
- spatial catch and effort time series
- size class of all species.

Figure 9 shows an example of one of the Global Tuna Atlas dataset which is meant to be updated every year but which keeps the same global DOI and gets a specific version DOI for each yearly update. This dataset is also published in FAO fisheries spatial data infrastructure (GeoNetwork, GeoServer..).

Figure 9. Example of DOI assigned to one of the Global Tuna Atlas dataset

5. Outlooks and benefits for the IOTC information system

As explained, all IOTC resources might be assigned a DOI and then also be advertised in other data infrastructures. By doing so, IOTC resources would be better indexed and findable on the Web. This would improve the impact of IOTC resources but also the visibility of the IOTC Website since all DOIs are pointing back to the IOTC Web site.

Another benefit is that DOIs are a standard for many tools (e.g., Zotero bibliographic management system), in particular the main Web CMS (e.g., Wordpress..) which all have plugins to import DOIs. This means that Zenodo might be used first to reference IOTC resources with DOIs and then, IOTC Web page for the resource might be easily created by importing its DOIs instead of the other way around. Such an approach is thus particularly interesting when the underlying software of a Website has to be changed for another one. In such a case, DOIs can be used as a mediation to export and import the content from the old to the new Website without losing a stable URL.

In terms of the ergonomics of the GUIs, it is worth noticing that Zenodo recently upgraded its GUIs and now offers an enriched editing environment as well as a new search engine within a given community. These services, open to all scientists and accessible through APIs, are also a good alternative to complete the current IOTC Website forms used to edit metadata (Figure 10) and search engine, used to explore IOTC resources (Figure 11).

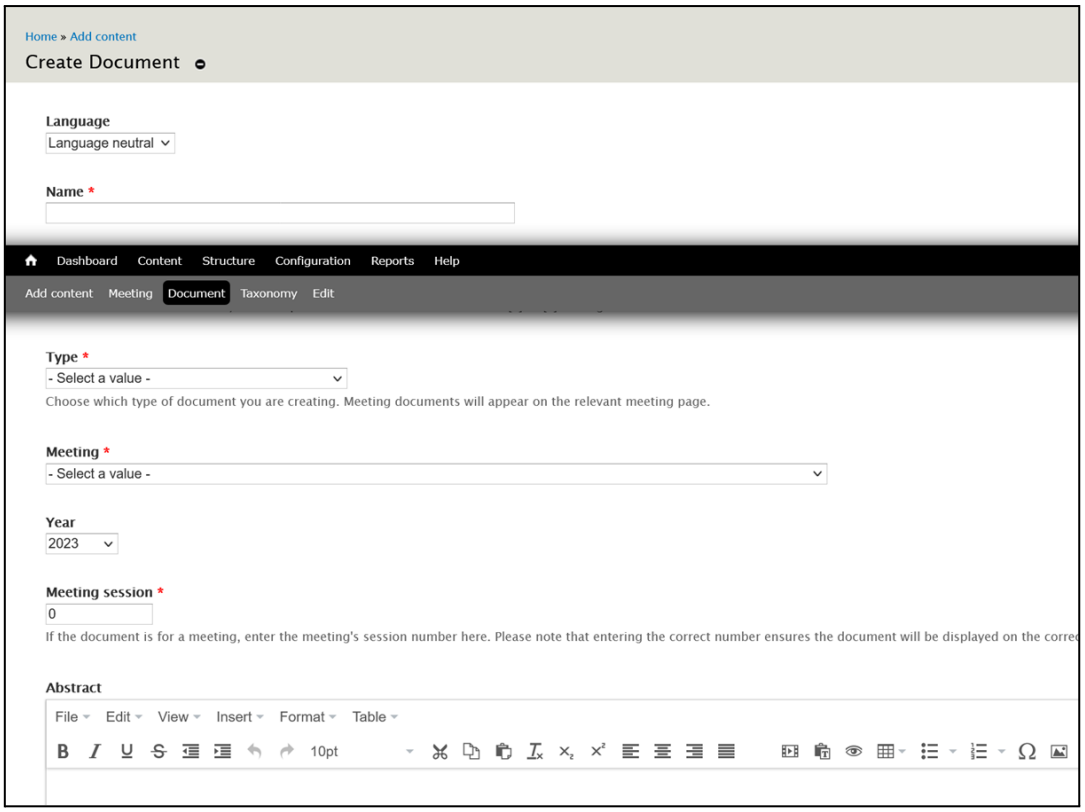

Figure 10. Current IOTC Secretariat form for uploading and modifying metadata of resources

Figure 11. Search engine for documents on IOTC Web page

6. Conclusion and outlooks

In the scientific context, whether resources are publicly available or not, there is no reason not to assign them a DOI. Indeed, Web sites are not meant to ensure long term storage and access and are

not as efficient as data repositories to advertise scientific work. Moreover, the replication of numerical resources in multiple infrastructures is a requirement not to lose the data whether or not data access is closed, restricted (with or without embargo) or open. IOTC is now familiar with the use of data repositories since 2019 but, at this stage IOTC, like other RFMOs, is mainly using Web sites to disseminate its resources and is not taking advantage of assigning them DOIs (easy replication fostering long term storage, discovery and access, interoperability..). Given the limitations of the IOTC website, we thus strongly recommend, at least for some of the resources, to assign DOIs in data repositories (like Zenodo or GBIF..). For example, most of IOTC datasets are only made publicly available on the Web site which makes their discovery, the tracking of their reuse or their citation currently difficult. DOIs can drastically increase the visibility, reuse and citation of IOTC data, papers and reports. When assigned by repositories like Zenodo, DOIs and resources will be stored and accessible for at least the next 25 years in a highly secured data center. As the main limitation for this work is the time availability of IOTC staff, part of it can be subcontracted and funded by projects. However, there is first a validation process and a set of technical specifications to be defined, in particular for selecting the past resources already published on the Web site which should also be uploaded on data repositories. Recognising the limitations of the IOTC Secretariat, such as human resources to upload the documents and datasets on an open access repository, the Secretariat could:

- Consider outsourcing for uploading IOTC Secretariat resource information, including past resources;
- Encouraging meeting participants to provide their papers with a valid DOI ahead of the working parties, Zenodo being one possible solution like other platforms (e.g. Pangaea, GBIF, Seanoe..)
- Reformatting the layout of the IOTC resources web pages to display richer metadata which would be directly harvested from its DOI (ideally by making use of existing plugins for DOIs) or other metadata formats (e.g. GeoNetwork).

The first step however remains a governance issue with the adoption of an updated strategy to improve IOTC resources management.

Acknowledgments

This work has received support from the G2OI ERDF INTERREG project, co-financed by the European Union, the Reunion region, and the French Republic, as well as from 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. Zenodo. https://doi.org/10.5281/zenodo.8365600

E. Blondel, J. Barde, W. Heintz, and A. Bennici. (2020). geoflow: R engine to orchestrate and run geospatial (meta)data workflows (0.0.20201116). Zenodo. https://doi.org/10.5281/zenodo.4275926

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. URL. Zenodo. https://iotc.org/documents/WPDCS/15/24

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18