DOIs & entrepôts de données

accessibilité et reproductibilité des données, codes, rapports..

https://orcid.org/0wo000-0002-3519-6141

DOI 10.5281/zenodo.14795560

https://ror.org/049dk3691

Introduction

Pause café ?

- Contexte de la crise de la reproductibilité => données et codes associés aux articles (+ intéressants que l'article ?)
- Mieux comprendre les DOIs et les entrepôts de données pour les gérer
- Tester la mise en place de DOIs, de communautés (test avec Zenodo)
- Versionner les objets de la recherche
- Gérer les accès
- Reproductibilité, pérennité…
- Exemples : drone / photogrammétrie, pêche, biodiversité, IA, acoustique..

Crise de la reproductibilité

=> pression légitime de la société civile et des éditeurs scientifiques privés

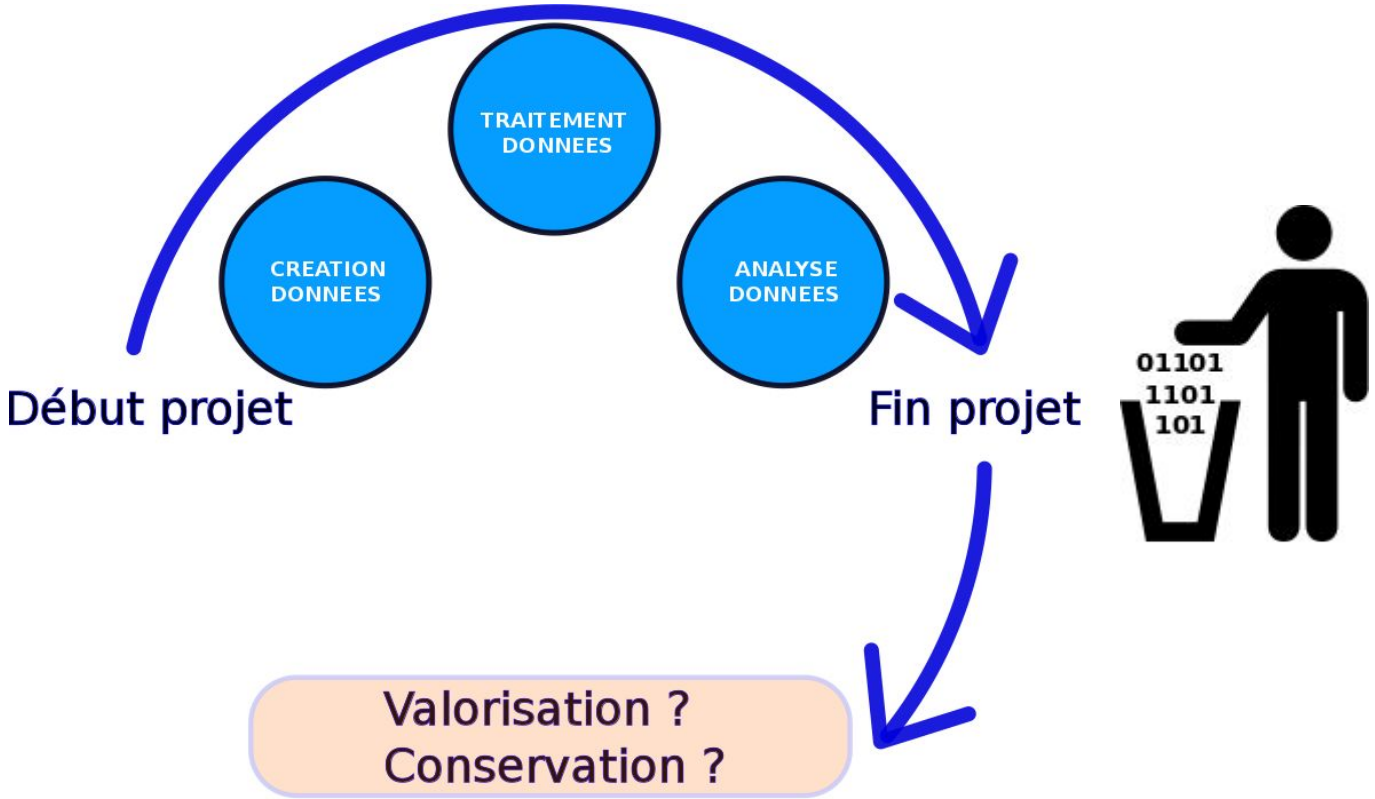
Discipline	% d'échec d'expériences de collègues	% d'échec de sa propre expérience
Chimie	90	60
Biologie	80	60
Physique et ingénierie	70	50
Médecine	70	60
Géologie et environnement	60	40

Source : wikipédia

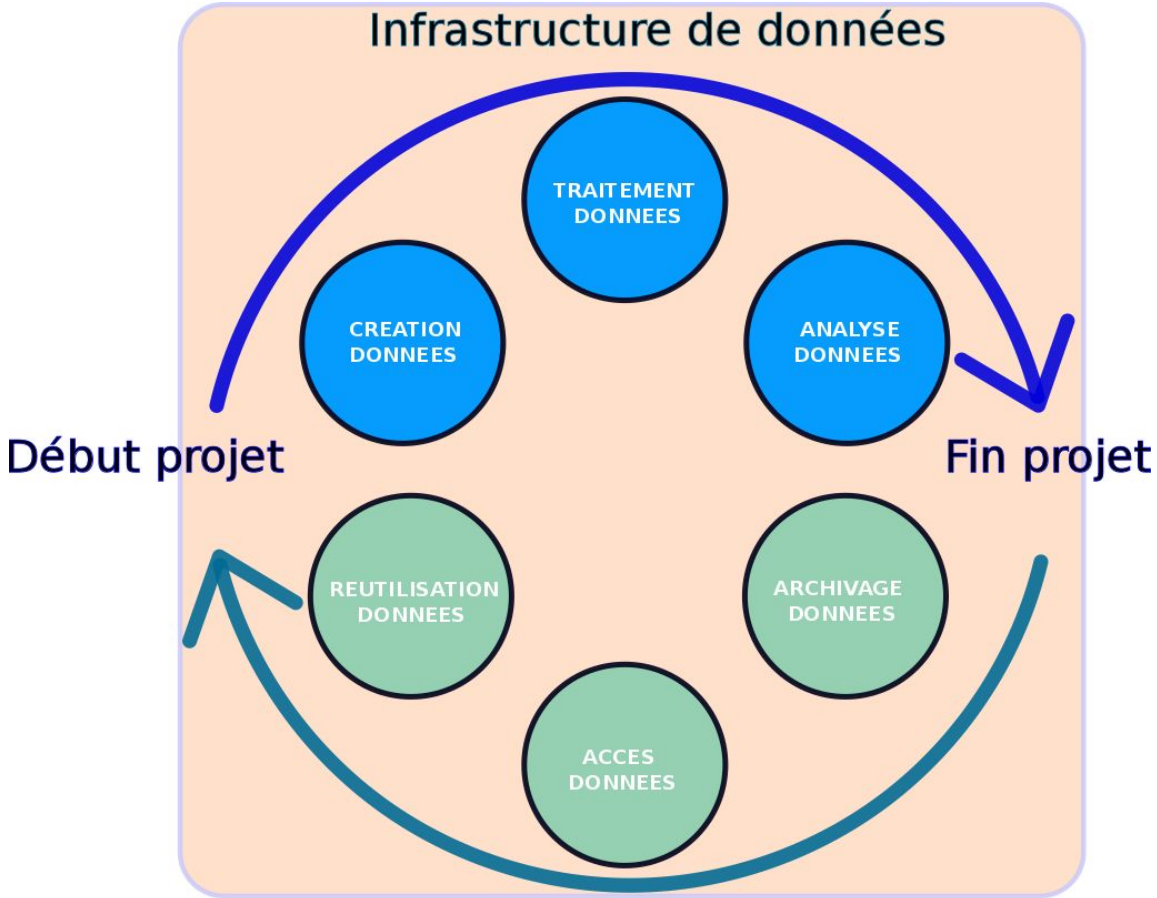
Des DOIs pour quels objets ?

- Documents :
  - Poster, rapports, conference proceedings…
  - Articles scientifiques
  - Data paper
  - ..
- Données
- Code
  - Script
  - Software
  - Package R..
- Protocoles
- Projets : e.g. CORDIS

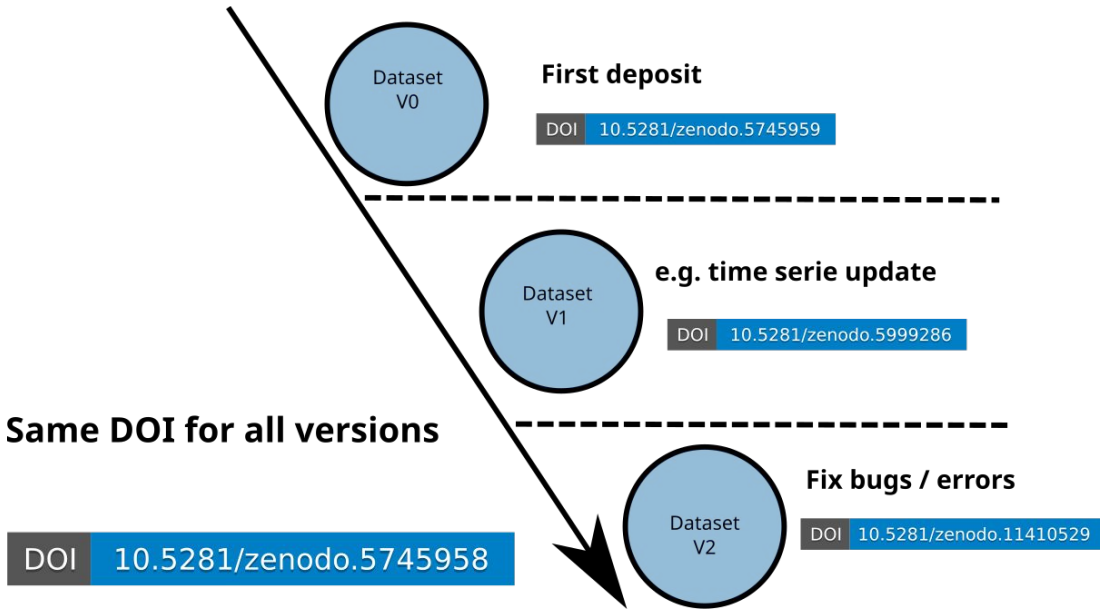
Cycle de vie des données: DMP & DOIs



Cycle de vie des données: DMP & DOIs

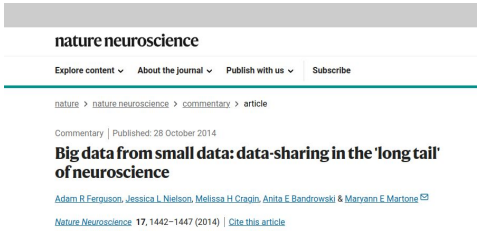


Cycle de vie des objets numériques

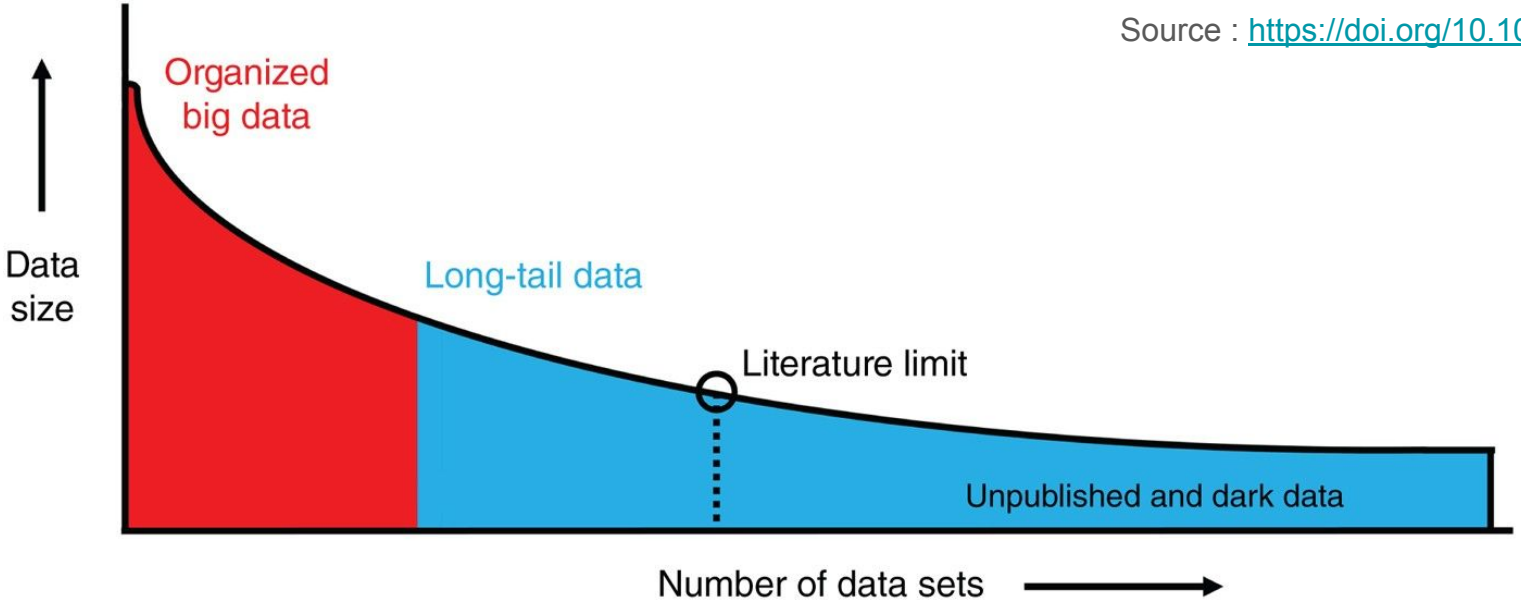


Différents DOIs de version pour un même objet qui évolue : à chaque version son DOI

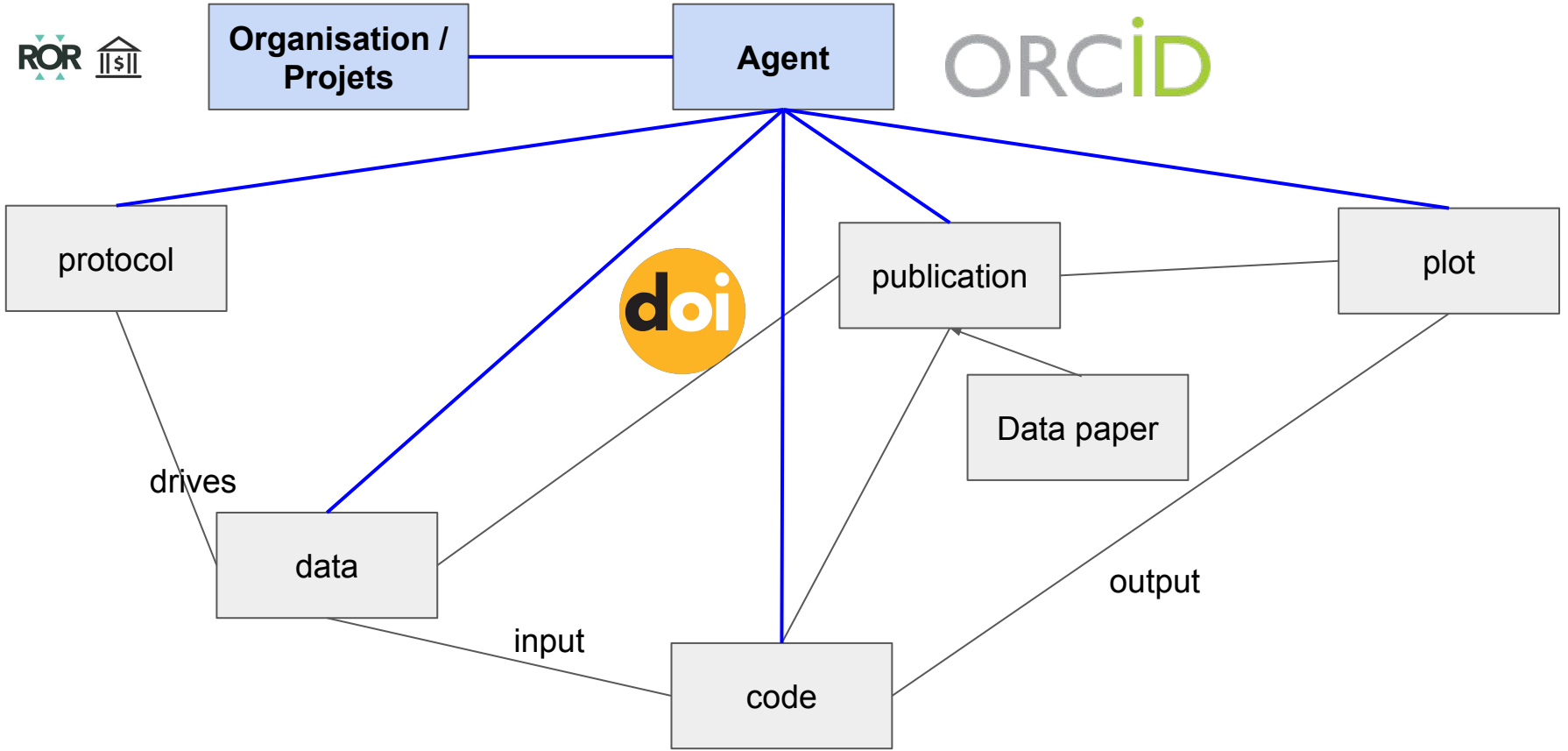
Long tail of research data



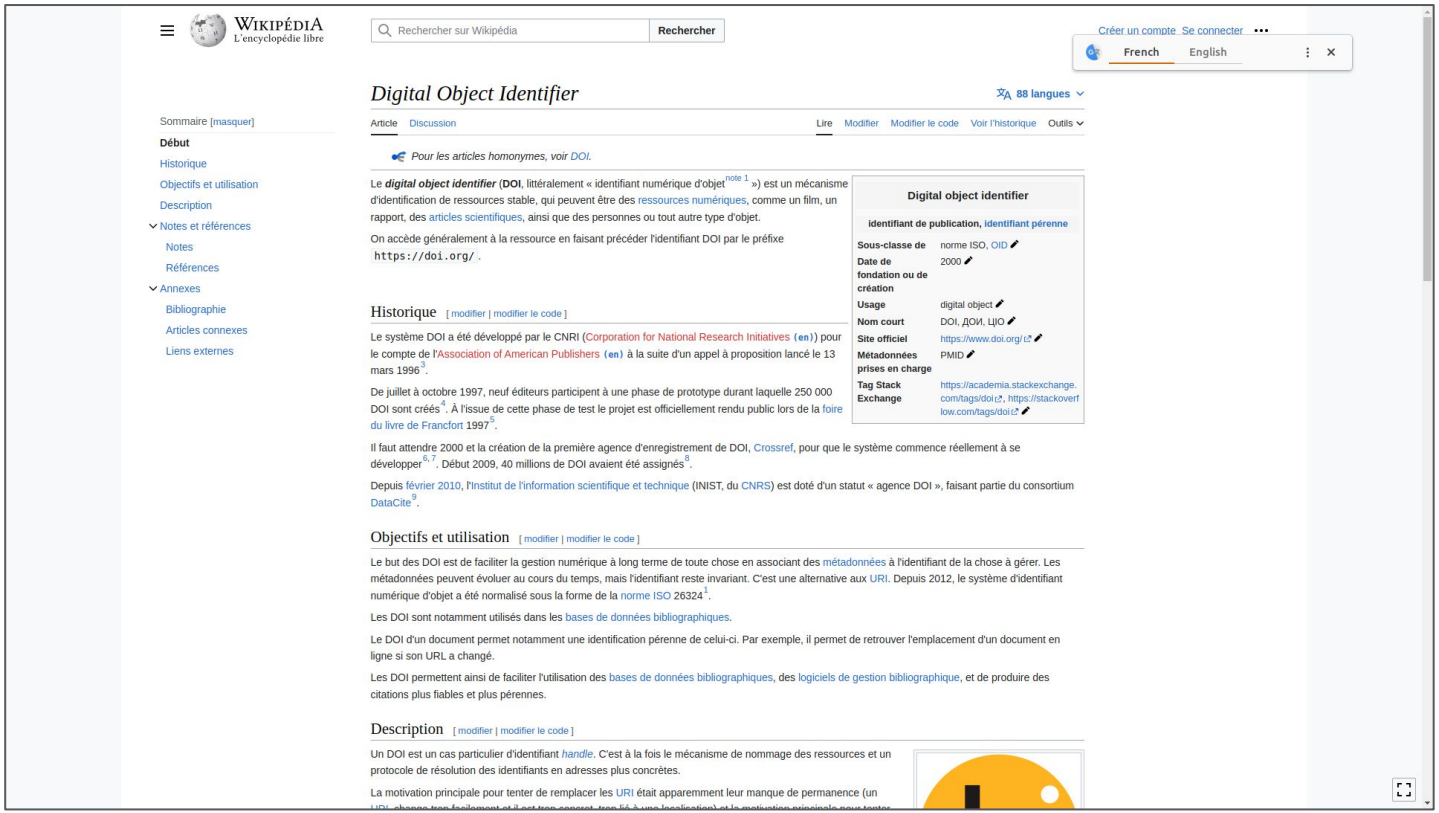
Source : https://doi.org/10.1038/nn.3838



Contexte : Open & Reproducible Science



DOIs : définition (wikipédia)



WIKIPÉDIA
L'encyclopédie libre

Rechercher sur Wikipédia — Rechercher

Créer un compte Se connecter

French English

Digital Object Identifier

88 langues

Article Discussion — Lire Modifier Modifier le code Voir l'historique Outils

Pour les articles homonymes, voir DOI.

Sommaire [masquer]

- Début
- Historique
- Objectifs et utilisation
- Description
- Notes et références
  - Notes
  - Références
- Annexes
  - Bibliographie
  - Articles connexes
  - Liens externes

Le ***digital object identifier*** (**DOI**, littéralement « identifiant numérique d'objet[note 1] ») est un mécanisme d'identification de ressources stable, qui peuvent être des ressources numériques, comme un film, un rapport, des articles scientifiques, ainsi que des personnes ou tout autre type d'objet.

On accède généralement à la ressource en faisant précéder l'identifiant DOI par le préfixe `https://doi.org/` .

Digital object identifier

identifiant de publication, identifiant pérenne

Sous-classe de	norme ISO, OID
Date de fondation ou de création	2000
Usage	digital object
Nom court	DOI, ДОИ, ЦЮ
Site officiel	https://www.doi.org/
Métadonnées prises en charge	PMID
Tag Stack Exchange	https://academia.stackexchange.com/tags/doi, https://stackoverflow.com/tags/doi

Historique [modifier | modifier le code]

Le système DOI a été développé par le CNRI (Corporation for National Research Initiatives (en)) pour le compte de l'Association of American Publishers (en) à la suite d'un appel à proposition lancé le 13 mars 1996[3].

De juillet à octobre 1997, neuf éditeurs participent à une phase de prototype durant laquelle 250 000 DOI sont créés[4]. À l'issue de cette phase de test le projet est officiellement rendu public lors de la foire du livre de Francfort 1997[5].

Il faut attendre 2000 et la création de la première agence d'enregistrement de DOI, Crossref, pour que le système commence réellement à se développer[6,7]. Début 2009, 40 millions de DOI avaient été assignés[8].

Depuis février 2010, l'Institut de l'information scientifique et technique (INIST, du CNRS) est doté d'un statut « agence DOI », faisant partie du consortium DataCite[9].

Objectifs et utilisation [modifier | modifier le code]

Le but des DOI est de faciliter la gestion numérique à long terme de toute chose en associant des métadonnées à l'identifiant de la chose à gérer. Les métadonnées peuvent évoluer au cours du temps, mais l'identifiant reste invariant. C'est une alternative aux URI. Depuis 2012, le système d'identifiant numérique d'objet a été normalisé sous la forme de la norme ISO 26324[1].

Les DOI sont notamment utilisés dans les bases de données bibliographiques.

Le DOI d'un document permet notamment une identification pérenne de celui-ci. Par exemple, il permet de retrouver l'emplacement d'un document en ligne si son URL a changé.

Les DOI permettent ainsi de faciliter l'utilisation des bases de données bibliographiques, des logiciels de gestion bibliographique, et de produire des citations plus fiables et plus pérennes.

Description [modifier | modifier le code]

Un DOI est un cas particulier d'identifiant *handle*. C'est à la fois le mécanisme de nommage des ressources et un protocole de résolution des identifiants en adresses plus concrètes.

La motivation principale pour tenter de remplacer les URI était apparemment leur manque de permanence (un URL change trop facilement et il est trop concret, trop lié à une localisation) et la motivation principale pour tenter

Rappels sur les DOIs

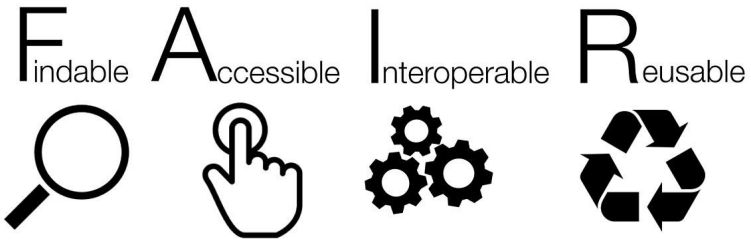
- Identifiant unique = URI (Web sémantique..)
- Suffisant pour mettre en œuvre un Plan de Gestion de Données !
- Différents systèmes de DOIs : Datacite, crossref…
  - https://commons.datacite.org/?query=tuna++atlas+firms
  - google dataset search
- Trois étapes :
  - téléverser / uploader l'objet
  - décrire les métadonnées associées :
    - Titre, mots-clés, licence,
    - Auteurs, financements …
  - Définir le niveau d'accès
    - Ouvert | embargo | restreint | fermé

DataCite
FIND, ACCESS, AND REUSE DATA

DOI => données FAIR ?

Avec un DOI :

- Findable => Oui
- Accessible => Oui mais pas forcément ouvert
- Interopérable => pas forcément (sauf entrepôt métier type GBIF)
- Réutilisable => pas forcément



Question fil rouge

Quelles sont les raisons de ne pas mettre un DOI ?

=> intérêt de ne pas mettre un DOI si les ressources sont déjà en ligne et accessibles publiquement ? Exemple

Les bonnes raisons de mettre des DOIs :

- Individuelle
  - Assurance de ne pas perdre son travail
  - Permettre la reproductibilité
  - Valoriser son travail
- Déontologie / société
  - Légale
  - Urgence écologique

Entrepôts de données & DOIs

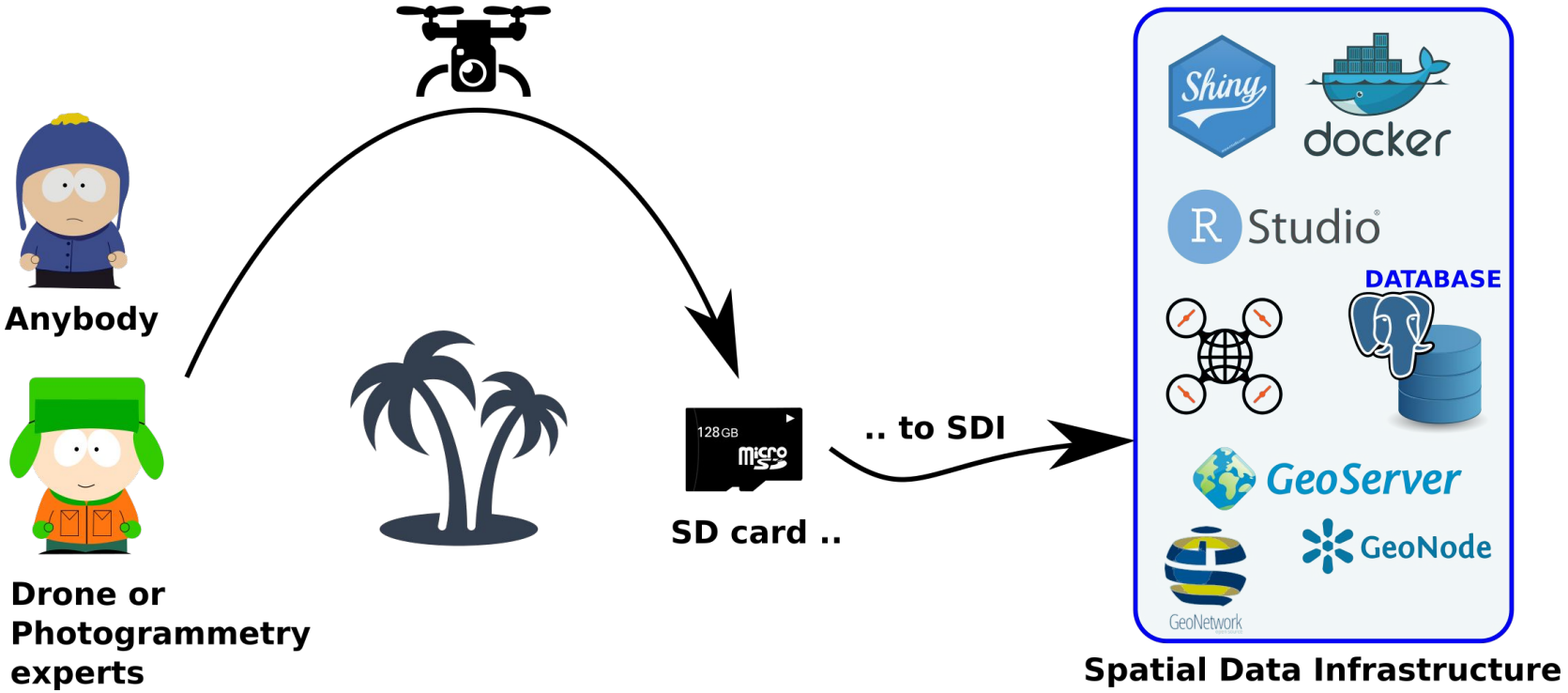
Internationaux

- Zenodo (CERN) : “*domain agnostic*” et communautés libres
- Pangaea (U. Bremen) : communautés thématiques
- GBIF (OCDE)
- libres et privés : figshare (Elsevier..)
- autres payants et privés : Dryad..

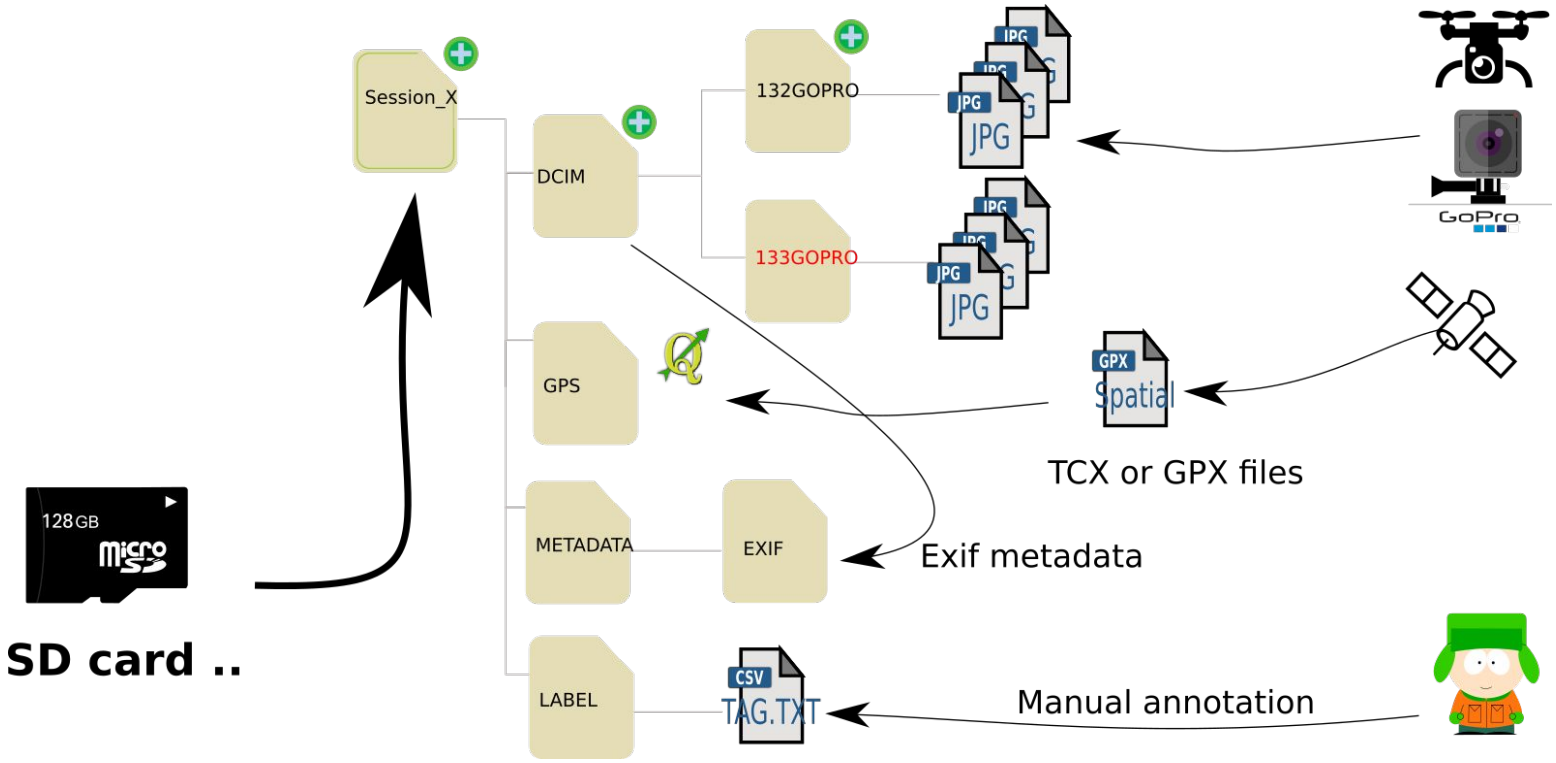
Entrepôts de données & DOIs : France

- Gouvernementaux / multi-tutelles
  - recherche.data.gouv.fr
  - Data Terra / ODATIS / EaSy Data
- Institutionnels :
  - INRAE : Data INRAE & Geodata INRAE
  - IRD : DataSuds & DataSuds-geo
  - Ifremer : “Seanoe” + Sextant
  - CNRS InDoRES : Data.InDoRES & cat.InDoRES
  - https://entrepot.recherche.data.gouv.fr/dataverse/umontpellier etc..

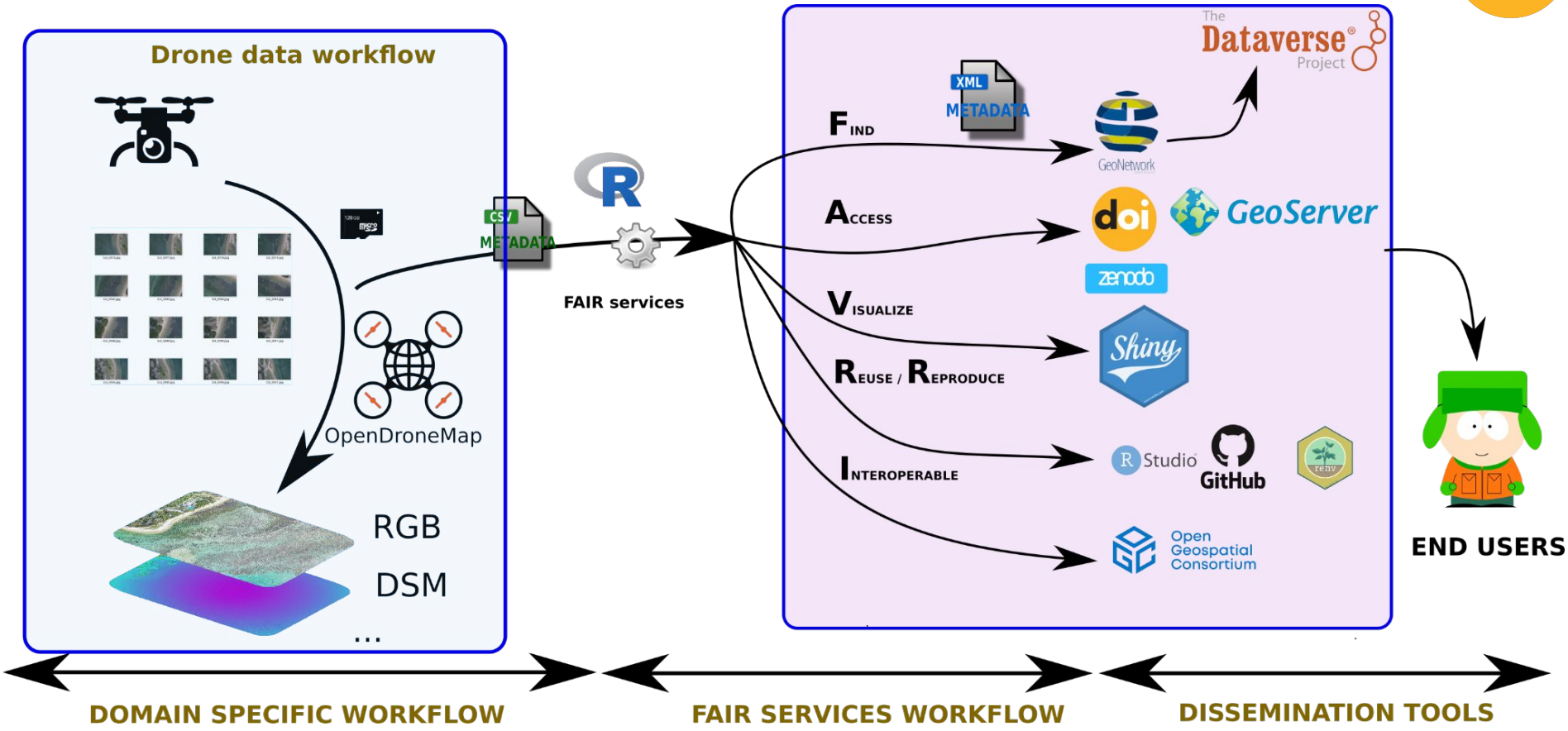
Exemple des données de drones



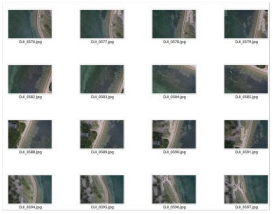
Conventions pour nommer / structurer les données

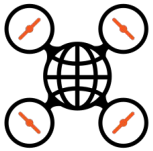


Exemple des données de drones : workflow

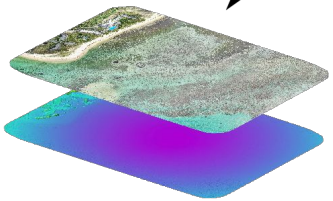


Exemple des données de drones : workflow





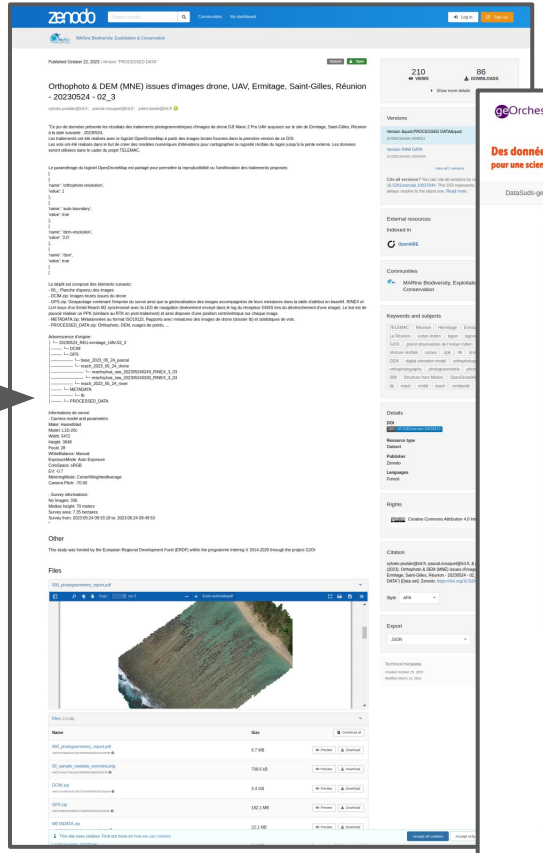
OpenDroneMap

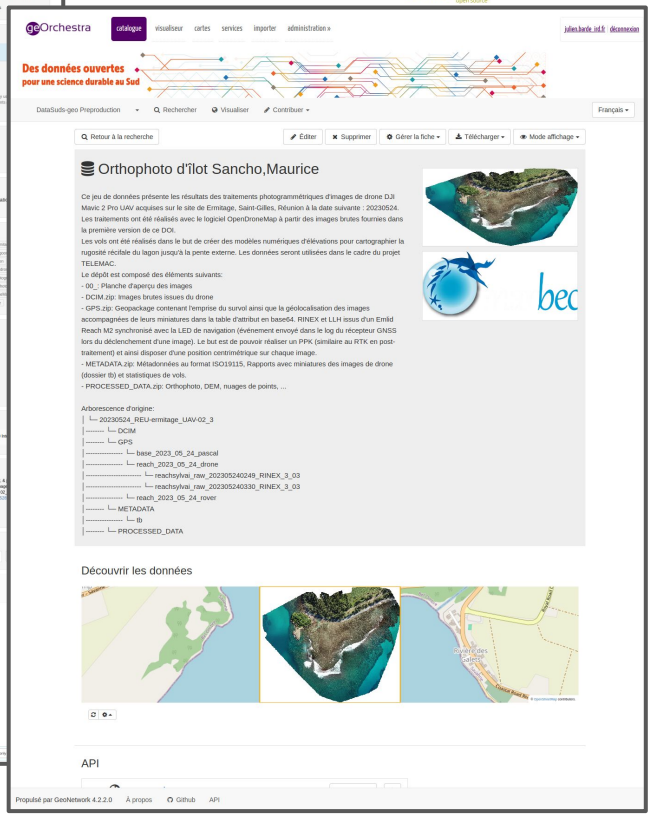


RGB

DSM

...





Principales caractéristiques

- GBIF (OCDE) : + 110k datasets
  - Authentication pour accès / téléchargement
  - Scope: species occurrences and specimen data & extensions Darwin Core
  - Orienté métier (biodiversité, barcoding) : Interopérabilité
  - DOI sur les requêtes query
  - Média associées, Système utilisé par iNaturalist, Pl@ntnet..
  - API en lecture (Package R..)
- ZENODO (CERN): +4.4M objects
  - En ligne au moins 25 ans
  - Pas d'authentification
  - Up to 50 GB, 100 files maximum
  - Généraliste : domain agnostic
  - API en lecture / écriture
  - Sandbox : https://sandbox.zenodo.org/
  - Login : ORCID / ROR, GitHub, Grants

Principales caractéristiques

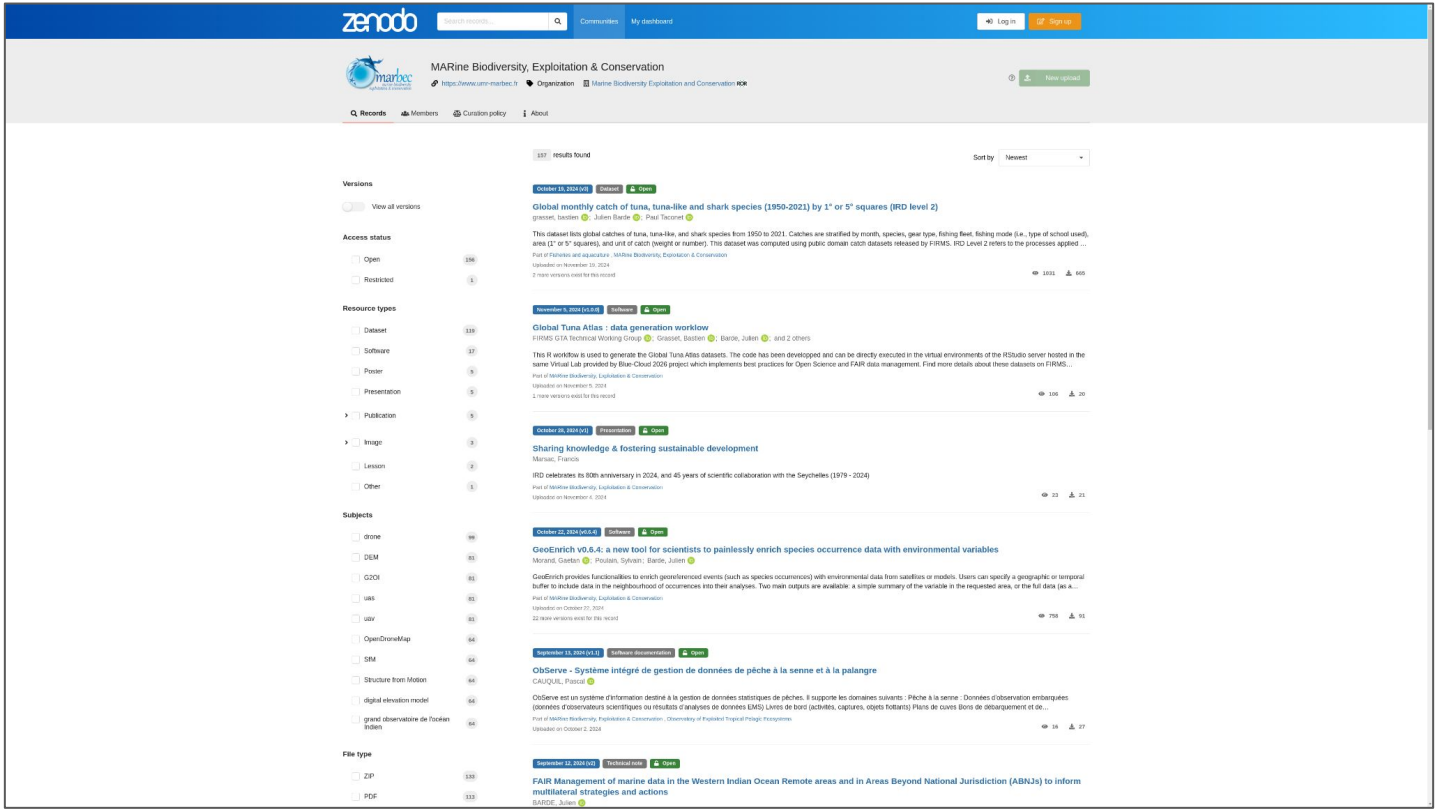
- Pangaea (Université de Bremen) : +432k datasets
  - Communautés métiers : “disciplines from earth and life sciences.”
  - Accès sans authentification
- Seanoe (+1,6k datasets) : 100GB, embargo 2 ans maximum, relecture
  - Exemples MARBEC
- DataSuds & DataSuds-geo (+1,3k datasets)
  - Exemples MARBEC
- Data INRAE : + 3,7k datasets & Geodata INRAE
- etc…

Entrepôt européen : Zenodo

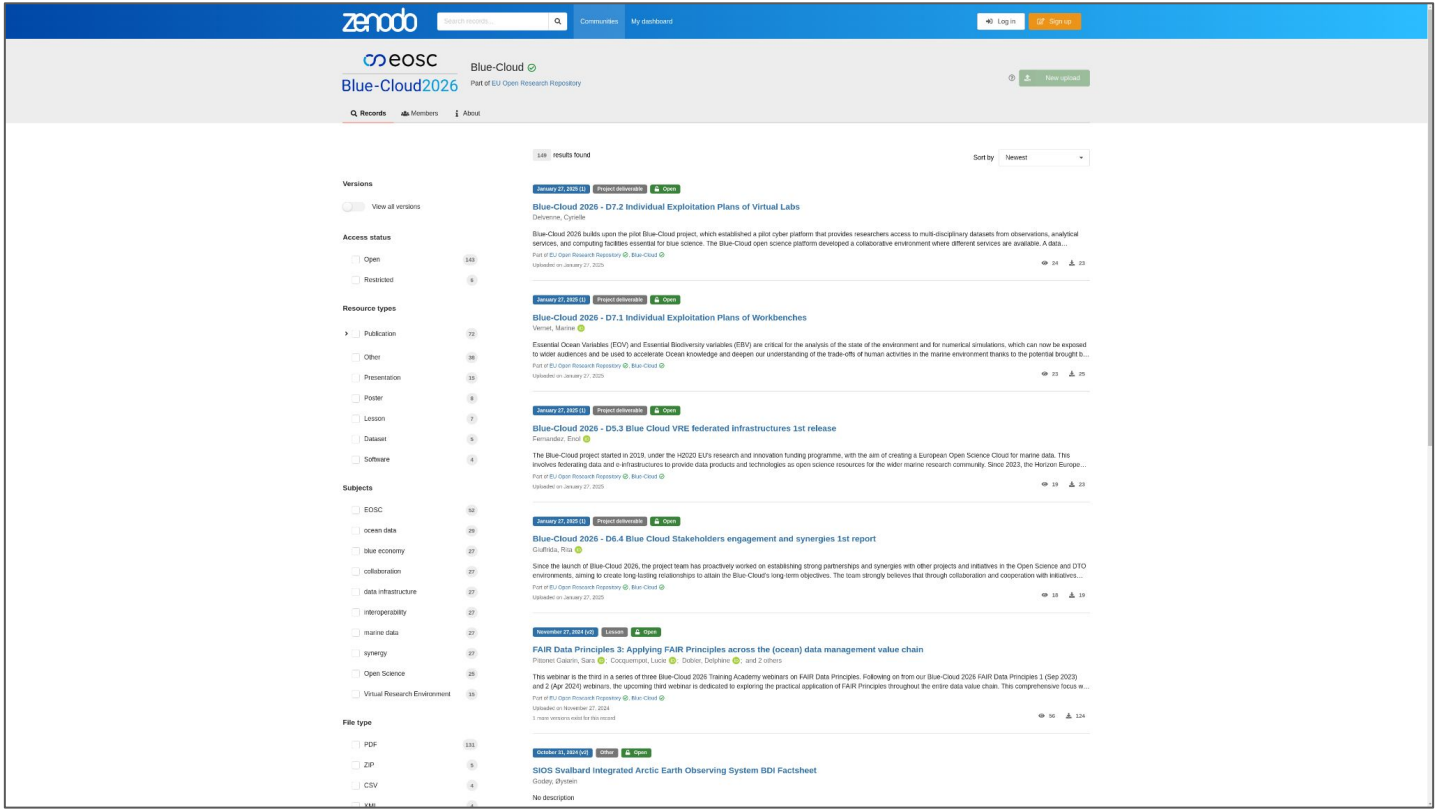
Plus de 4,4M DOIs

- \+ :
  - Tout objet de la recherche & domain agnostic
  - Création de communautés avec ou sans modération
  - Collaborative “shared” edition
  - DOIs accessibles sans authentifications
  - Sandbox : https://sandbox.zenodo.org/
  - API en lecture & écriture
  - ORCID / ROR, GitHub, OpenAire…
- \- :
- Communautés : exemple MARBEC

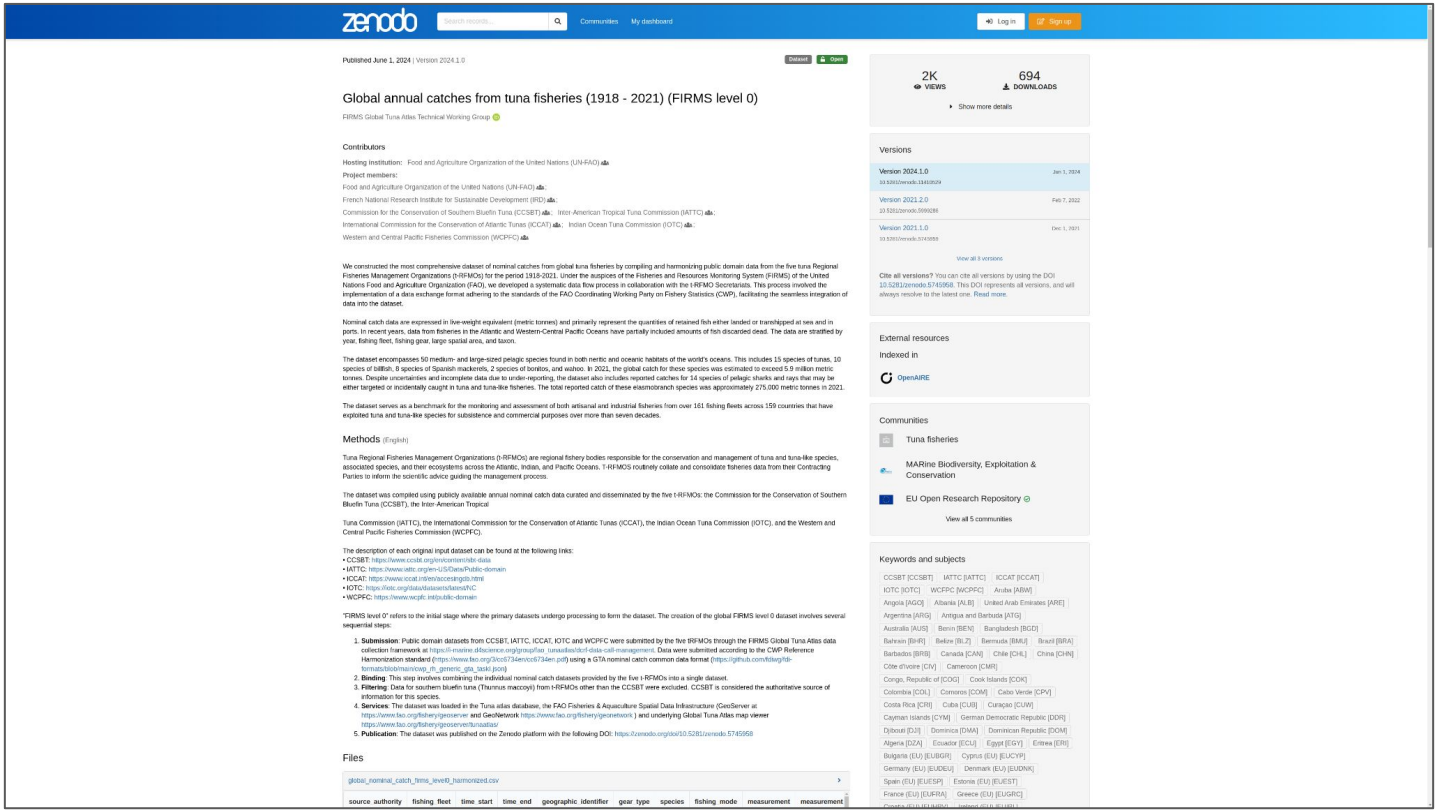
Communautés Zenodo : MARBEC



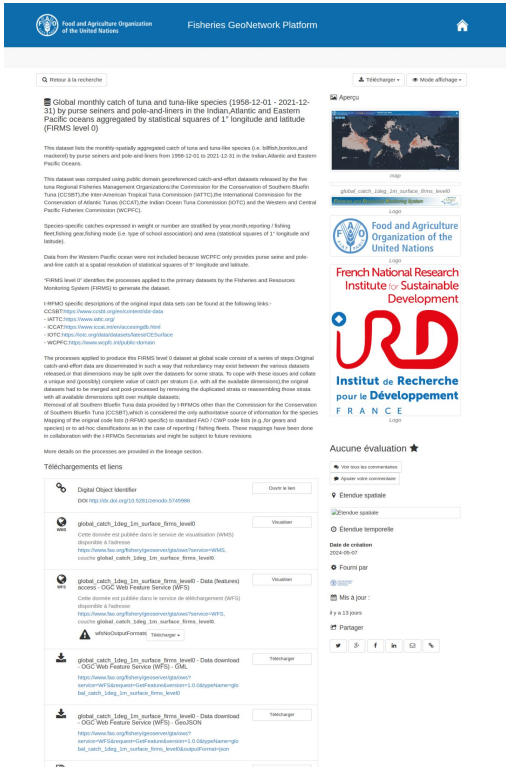
Communautés Zenodo : projet HORIZON



Dépôt Zenodo : data pêche



Dépôt Zenodo : data pêche



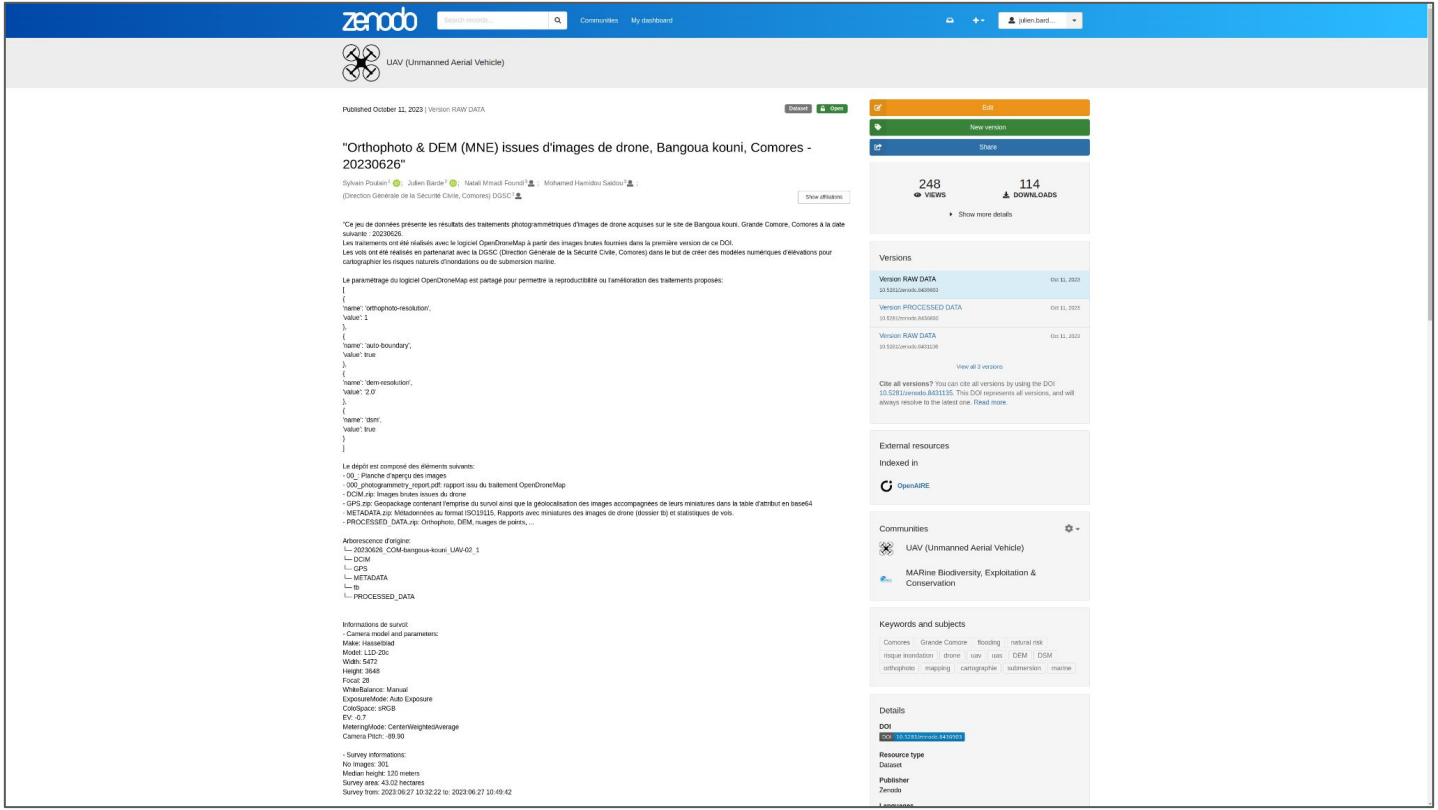
GeoNetwork

PostGIS

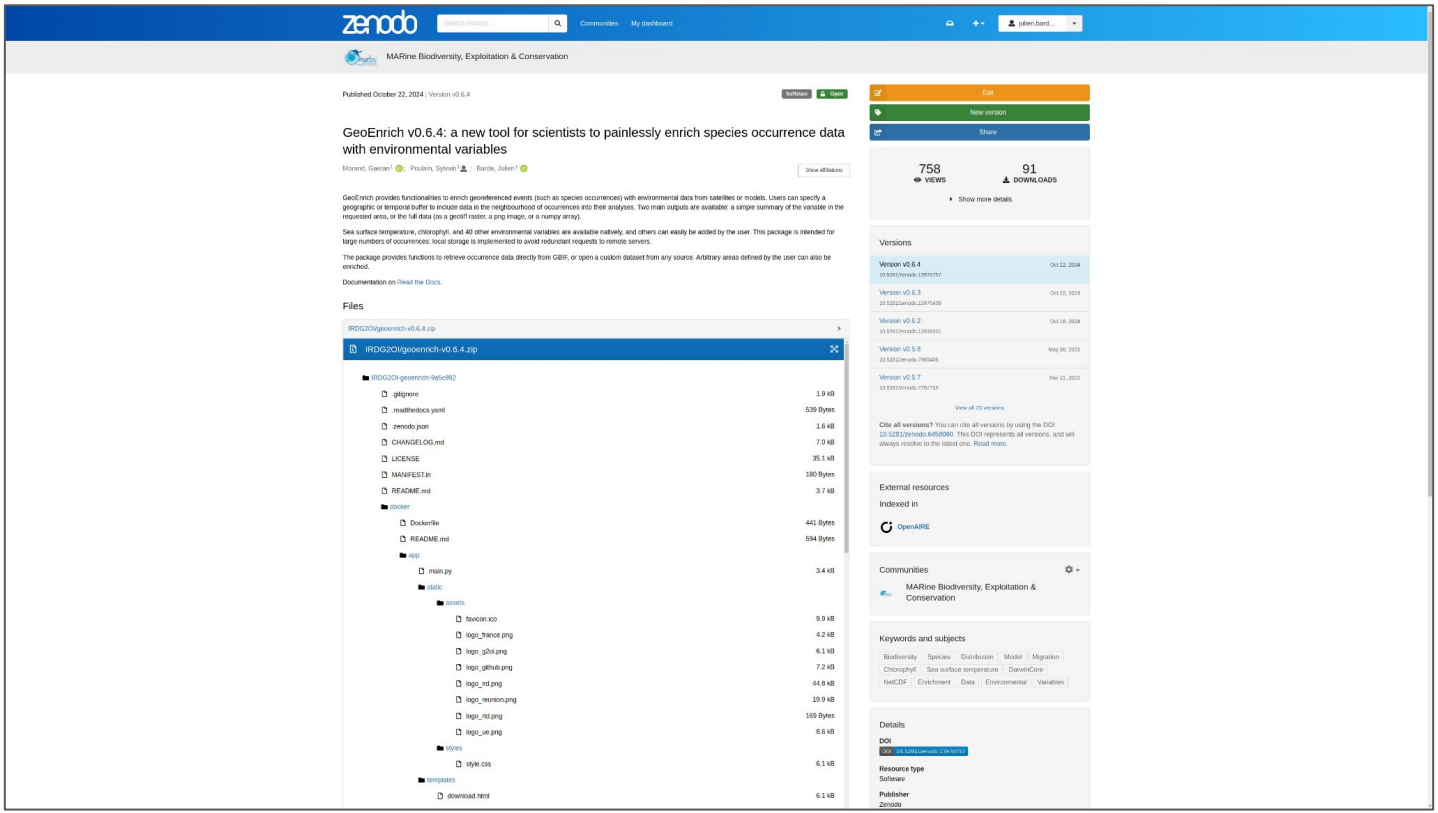
GeoServer

OpenLayers

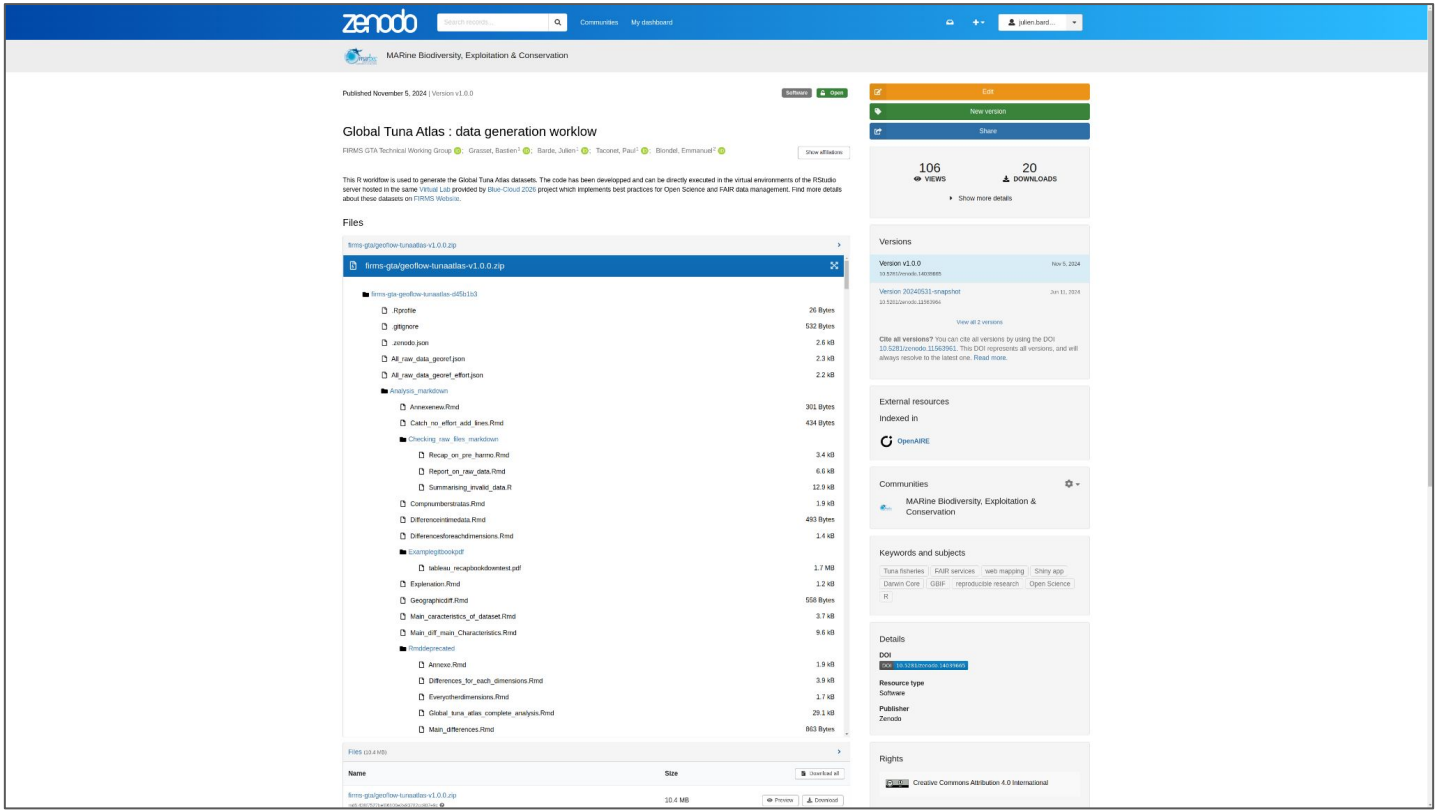
Dépôt Zenodo : data drone



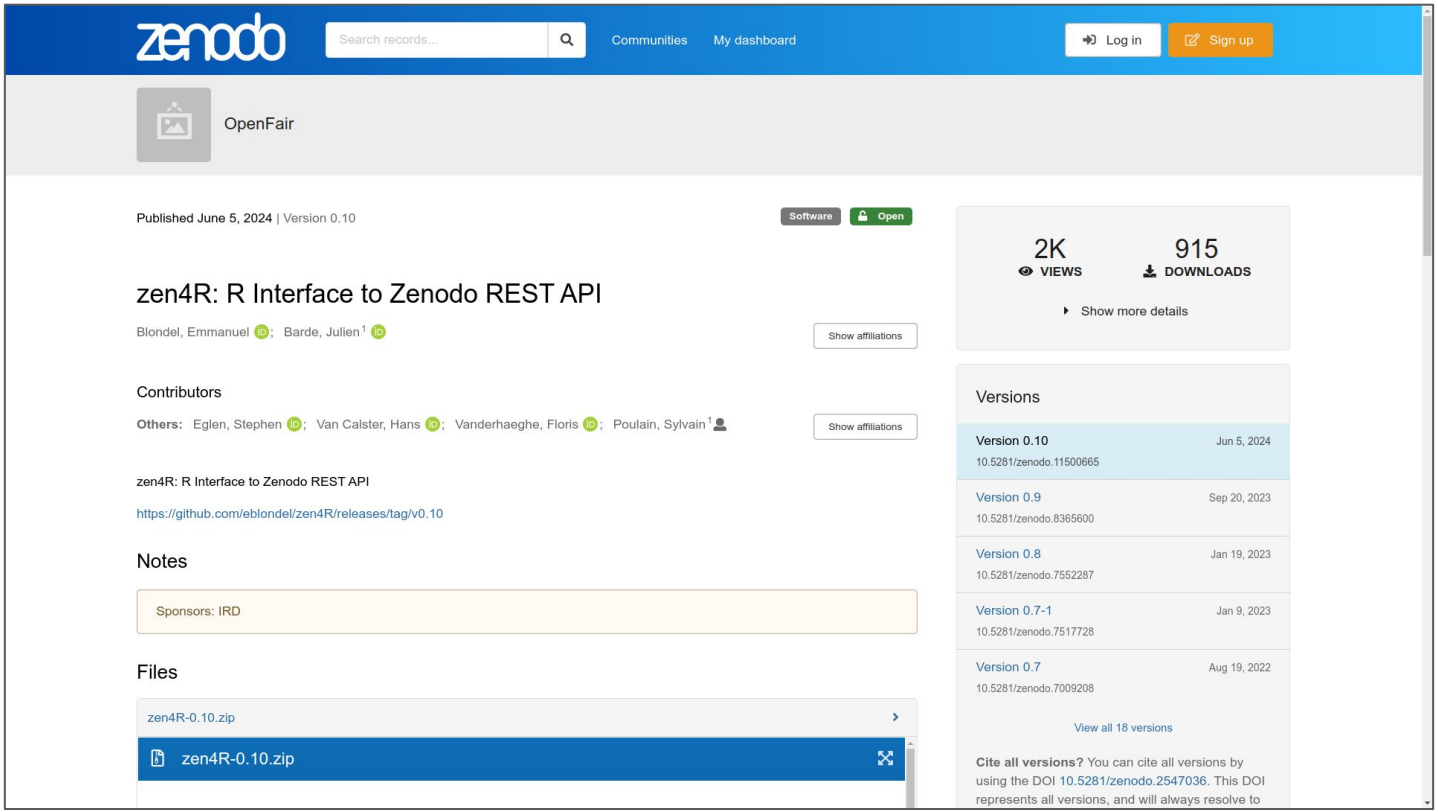
Dépôt Zenodo : code Python



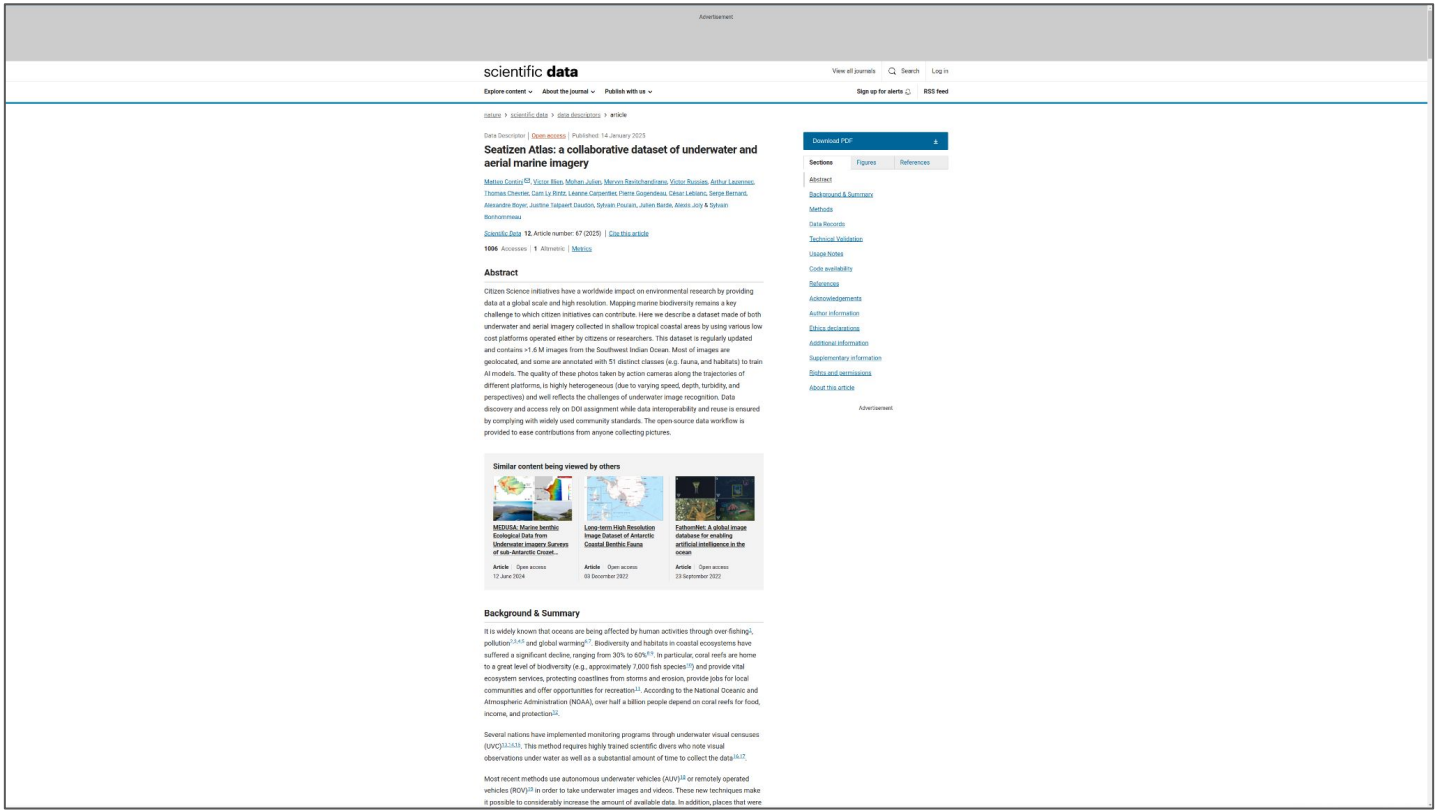
Dépôt Zenodo : code R



Dépôt Zenodo : releases package R (CRAN)



Data paper (thèse en cours, Matteo Contini)



scientific data

Seatizen Atlas: a collaborative dataset of underwater and aerial marine imagery

Abstract

Citizen Science initiatives have a worldwide impact on environmental research by providing data at a global scale and high resolution. Mapping marine biodiversity remains a key challenge to which citizen initiatives can contribute. Here we describe a dataset made of both underwater and aerial imagery collected in shallow tropical coastal areas by using various low cost platforms operated either by citizens or researchers. This dataset is regularly updated and contains >1.6 M images from the Southwest Indian Ocean. Most of images are geolocated, and some are annotated with 51 distinct classes (e.g., fauna, and habitats) to train AI models. The quality of these photos taken by action cameras along the trajectories of different platforms, is highly heterogeneous (due to varying speed, depth, turbidity, and perspectives) and well reflects the challenges of underwater image recognition. Data discovery and access rely on DOI assignment while data interoperability and reuse is ensured by complying with widely used community standards. The open-source data workflow is provided to ease contributions from anyone collecting pictures.

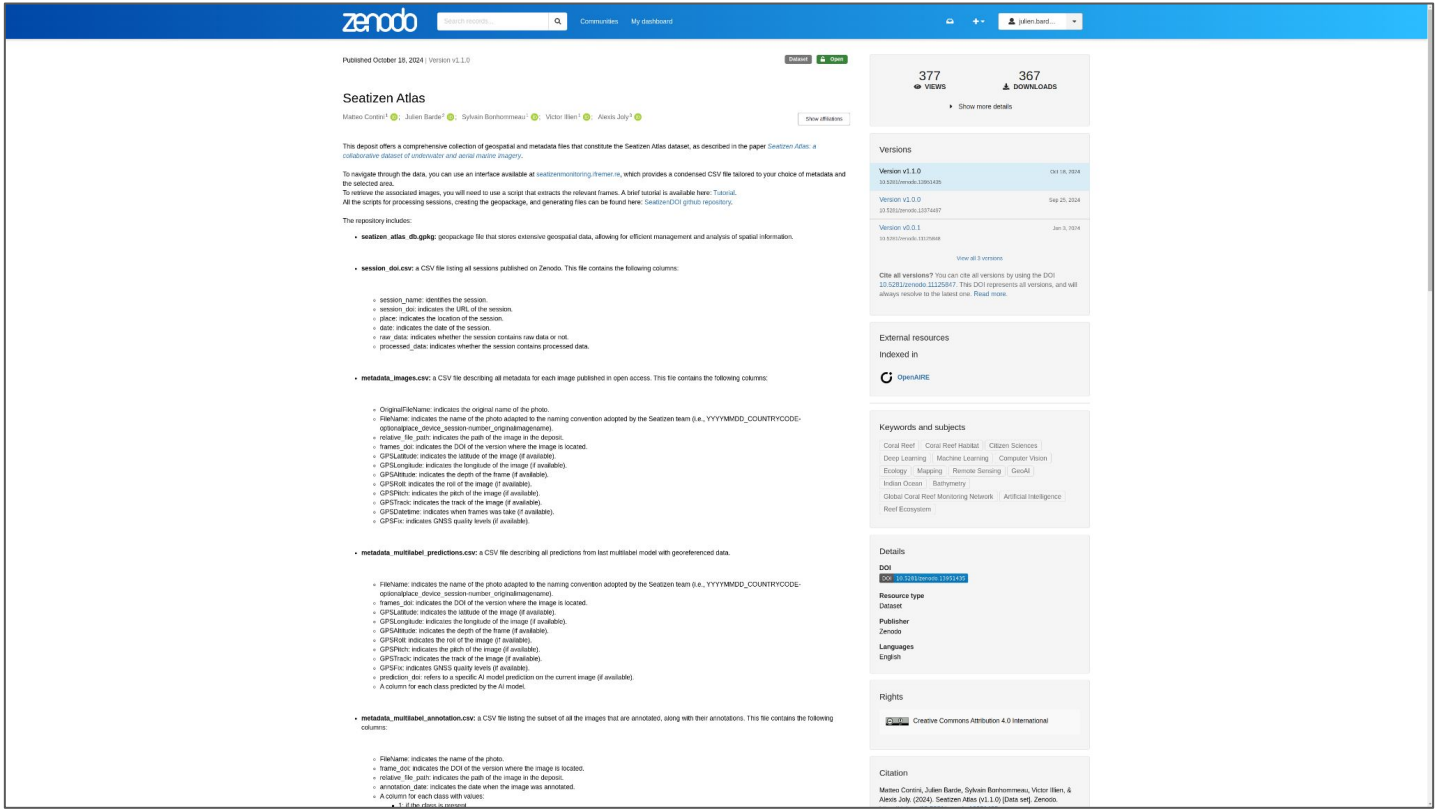
Background & Summary

It is widely known that oceans are being affected by human activities through over fishing[1], pollution[2,3,4,5] and global warming[6,7]. Biodiversity and habitats in coastal ecosystems have suffered a significant decline, ranging from 30% to 60%[8,9]. In particular, coral reefs are home to a great level of biodiversity (e.g., approximately 7,000 fish species[10]) and provide vital ecosystem services, protecting coastlines from storms and erosion, provide jobs for local communities and offer opportunities for recreation[11]. According to the National Oceanic and Atmospheric Administration (NOAA), over half a billion people depend on coral reefs for food, income, and protection[12].

Several nations have implemented monitoring programs through underwater visual censuses (UVC)[13,14,15]. This method requires highly trained scientific divers who note visual observations under water as well as a substantial amount of time to collect the data[16,17].

Most recent methods use autonomous underwater vehicles (AUV)[18] or remotely operated vehicles (ROV)[19] in order to take underwater images and videos. These new techniques make it possible to considerably increase the amount of available data. In addition, places that were

Data paper : données associées sur Zenodo



Article PCI (Gaétan Morand)

Peer Community Journal

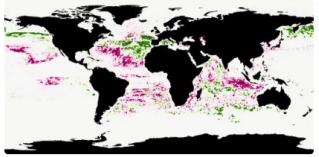
Search articles, authors.. Search Browse by volume Browse by section Browse by topic Browse by conference

« Browse issues Volume 4 (2024) article no. e93 »

Section: Ecology

Topic: Biophysics and computational biology, Ecology

Predicting species distributions in the open ocean with convolutional neural networks



Morand, Gaétan[1] ; Joly, Alexis[2] ; Rouyer, Tristan[1] ; Lorieul, Titouan[2] ; Barde, Julien[1]

Corresponding author(s): Morand, Gaétan (gaetan.morand@umontpellier.fr)

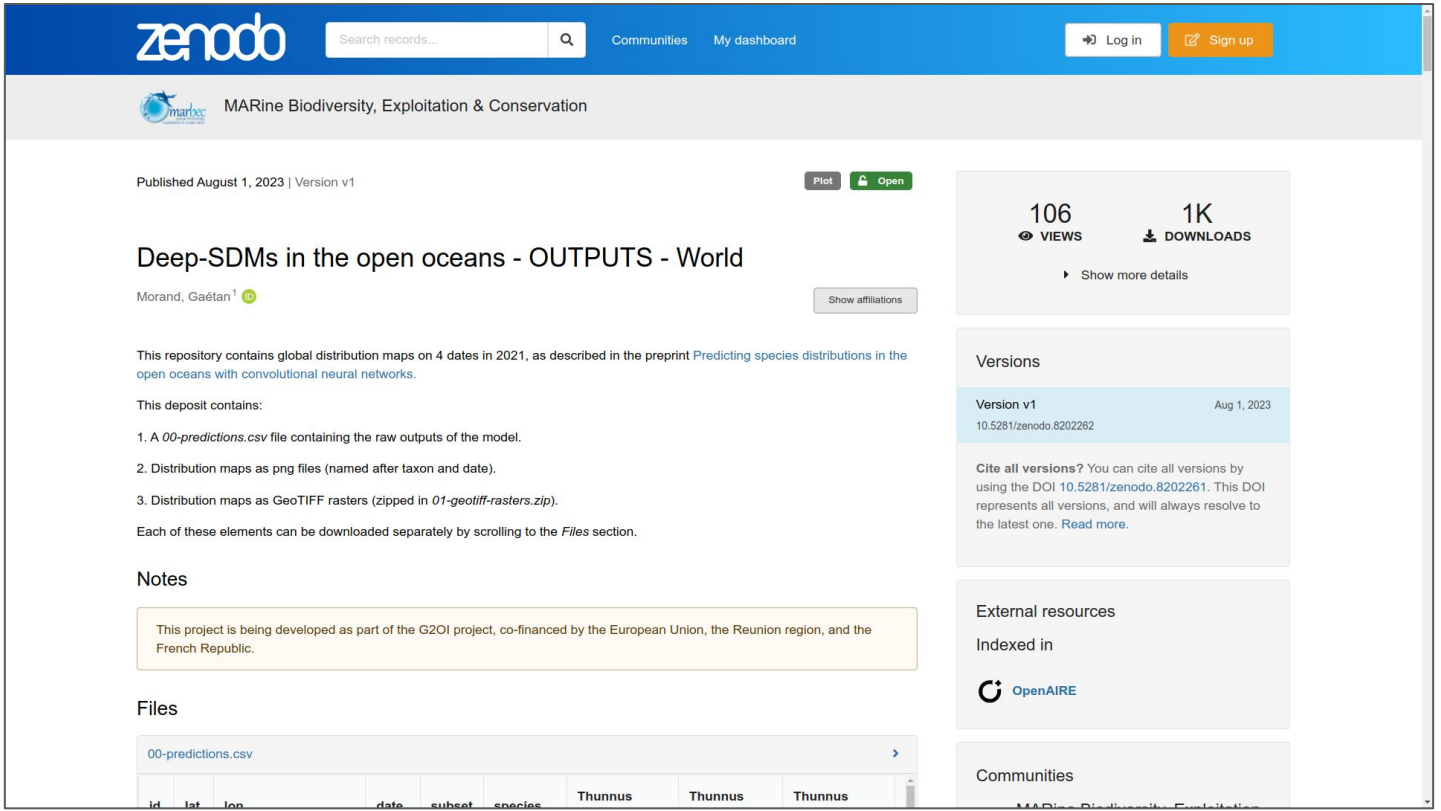
10.24072/pcjournal.471 - Peer Community Journal, Volume 4 (2024), article no. e93.

Get full text PDF Peer reviewed and recommended by PCI

Abstract

As biodiversity plummets due to anthropogenic disturbances, the conservation of oceanic species is made harder by limited knowledge of their distributions and migrations. Indeed, tracking species distributions in the open ocean is particularly challenging due to the scarcity of observations and the complex and variable nature of the ocean system. In this study, we propose a new method that leverages deep learning, specifically convolutional neural networks (CNNs), to capture spatial features of environmental variables. This novelty eliminates the need to predefine these features before modelling and creates opportunities to discover unexpected correlations. Our aim is to present the results of the first trial of this method in the open ocean, discuss limitations and provide feedback for future improvements or adjustments. In this case study, we considered 38 taxa comprising pelagic fishes, elasmobranchs, marine mammals, marine turtles and birds. We trained a model to predict probabilities from the environmental conditions at any specific point in space and time, using species occurrence data from the Global Biodiversity Information Facility (GBIF) and environmental data from various sources. These variables included sea surface temperature, chlorophyll concentration, salinity and fifteen others. During the testing phase, the model was applied to environmental data at locations where species occurrences were recorded. The classifier accurately predicted the observed taxon as the most likely taxon in 69% of cases and included the observed taxon among the top three most likely predictions in 89% of cases. These findings show the

Article PCI (Gaétan Morand) : données & code

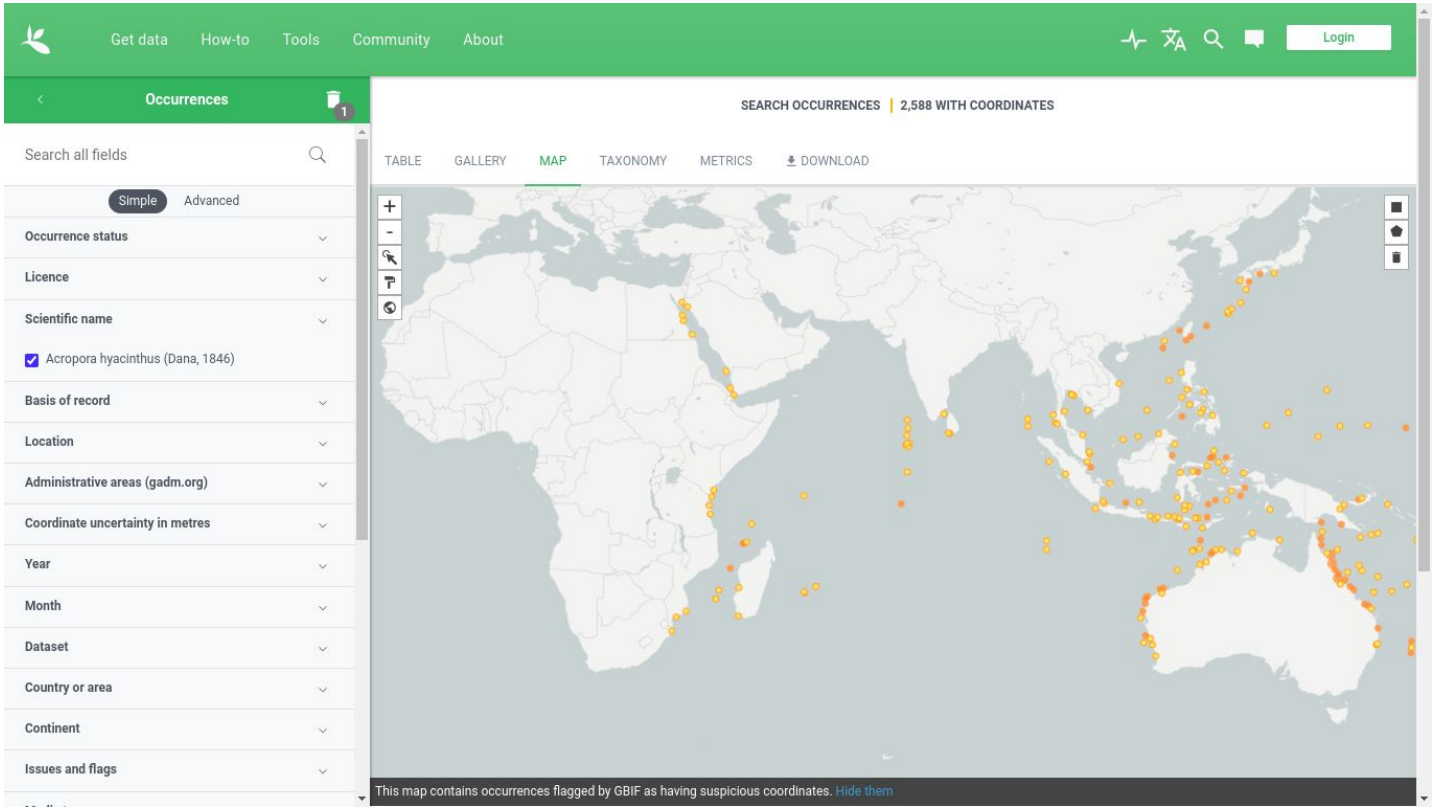


Entrepôt international : GBIF

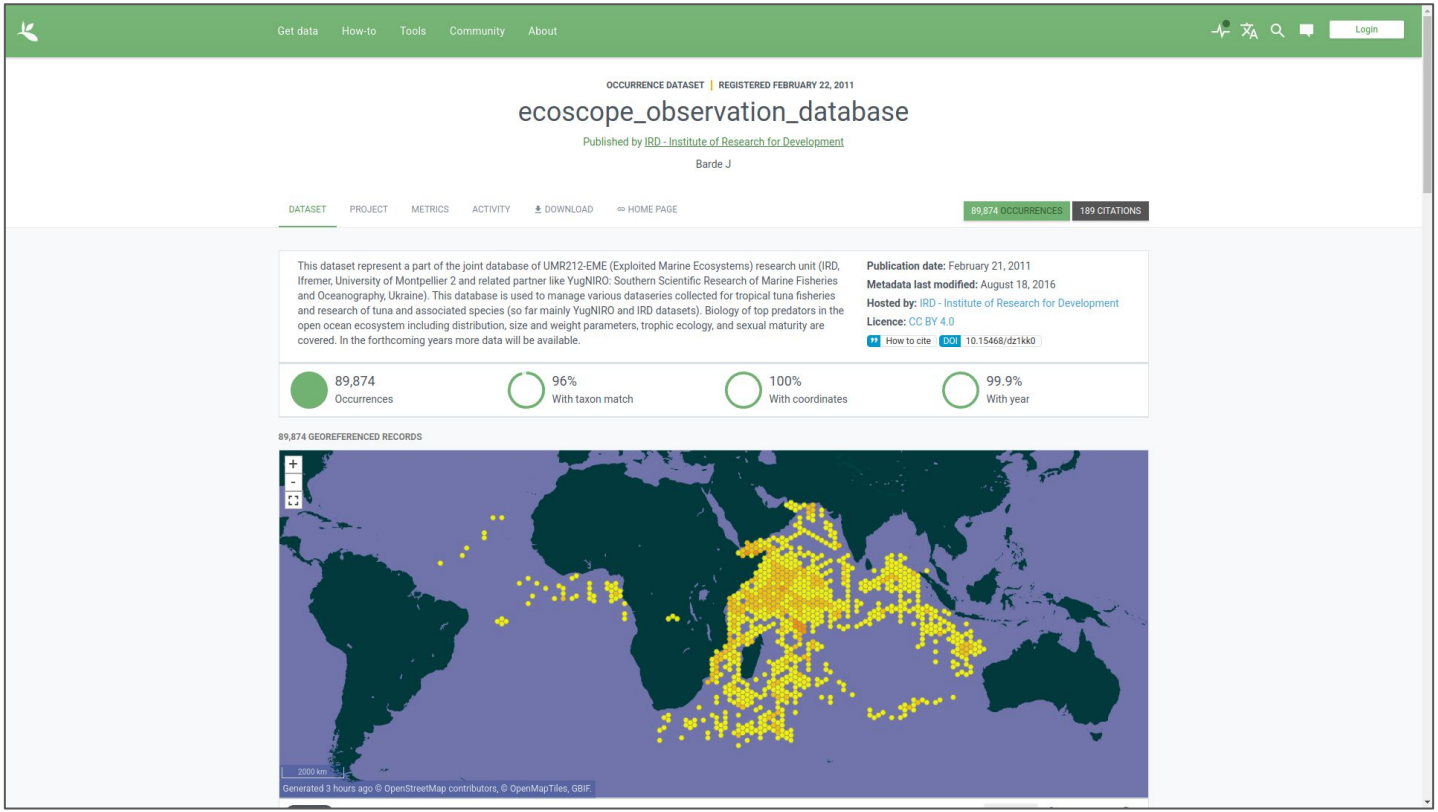
+ 110k DOIs

- + :
  - DOI sur les requêtes
  - Connexion de DOIs déjà attribués
  - Standardisation des formats : entrepôt métier
  - Média associées
- - :
  - Authentification nécessaire

Données de biodiversité : ex GBIF

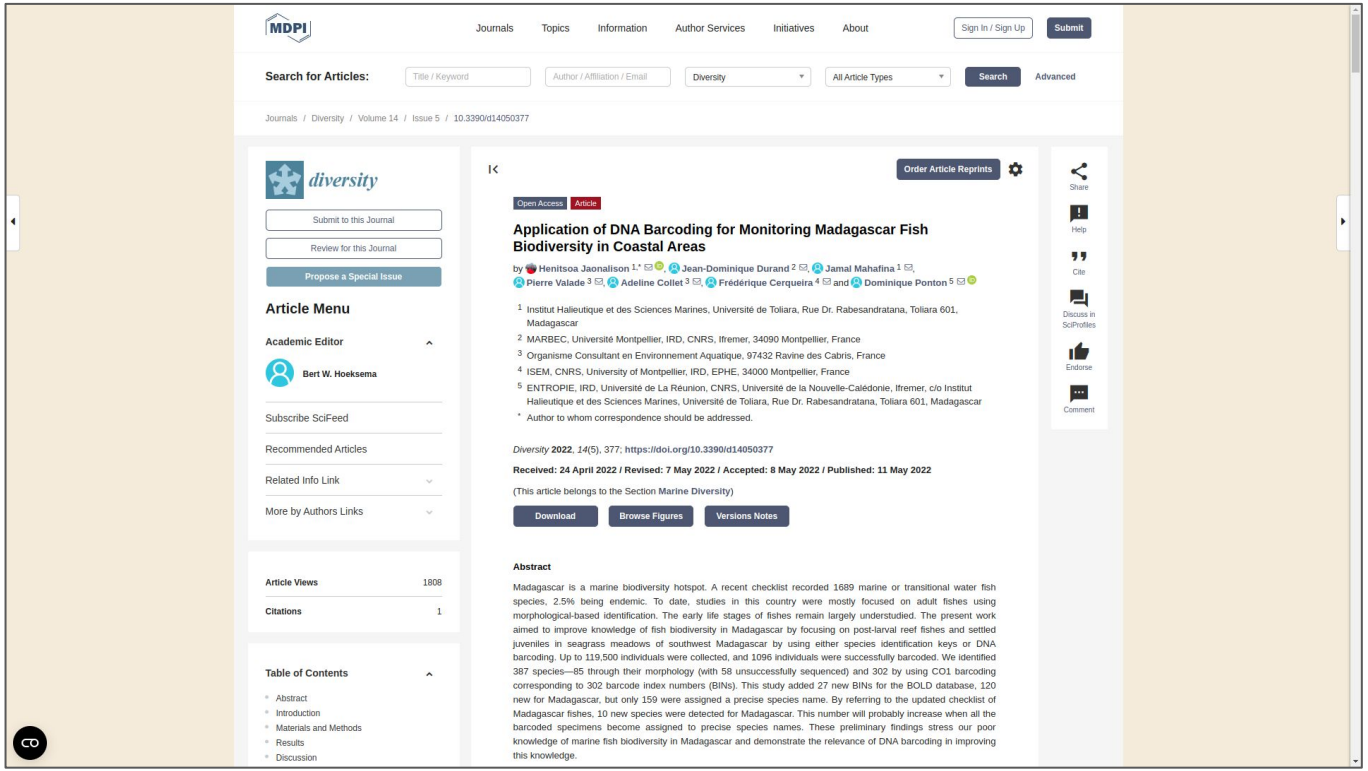


Données de biodiversité : ex GBIF



Données de Barcoding : ex BOLD

IHSM & CNDO : article barcoding



MDPI

Journals Topics Information Author Services Initiatives About Sign In / Sign Up Submit

Search for Articles: Title / Keyword Author / Affiliation / Email Diversity All Article Types Search Advanced

Journals / Diversity / Volume 14 / Issue 5 / 10.3390/d14050377

diversity

Submit to this Journal

Review for this Journal

Propose a Special Issue

Article Menu

Academic Editor

Bert W. Hoeksema

Subscribe SciFeed

Recommended Articles

Related Info Link

More by Authors Links

Article Views 1808

Citations 1

Table of Contents

- Abstract
- Introduction
- Materials and Methods
- Results
- Discussion

Order Article Reprints

Share Help Cite Discuss in SciProfiles Endorse Comment

Open Access Article

Application of DNA Barcoding for Monitoring Madagascar Fish Biodiversity in Coastal Areas

by Henitsoa Jaonalison [1,*], Jean-Dominique Durand [2], Jamal Mahafina [1], Pierre Valade [3], Adeline Collet [3], Frédérique Cerqueira [4] and Dominique Ponton [5]

[1] Institut Halieutique et des Sciences Marines, Université de Toliara, Rue Dr. Rabesandratana, Toliara 601, Madagascar

[2] MARBEC, Université Montpellier, IRD, CNRS, Ifremer, 34090 Montpellier, France

[3] Organisme Consultant en Environnement Aquatique, 97432 Ravine des Cabris, France

[4] ISEM, CNRS, University of Montpellier, IRD, EPHE, 34000 Montpellier, France

[5] ENTROPIE, IRD, Université de La Réunion, CNRS, Université de la Nouvelle-Calédonie, Ifremer, c/o Institut Halieutique et des Sciences Marines, Université de Toliara, Rue Dr. Rabesandratana, Toliara 601, Madagascar

[*] Author to whom correspondence should be addressed.

Diversity **2022**, *14*(5), 377; https://doi.org/10.3390/d14050377

Received: 24 April 2022 / Revised: 7 May 2022 / Accepted: 8 May 2022 / Published: 11 May 2022

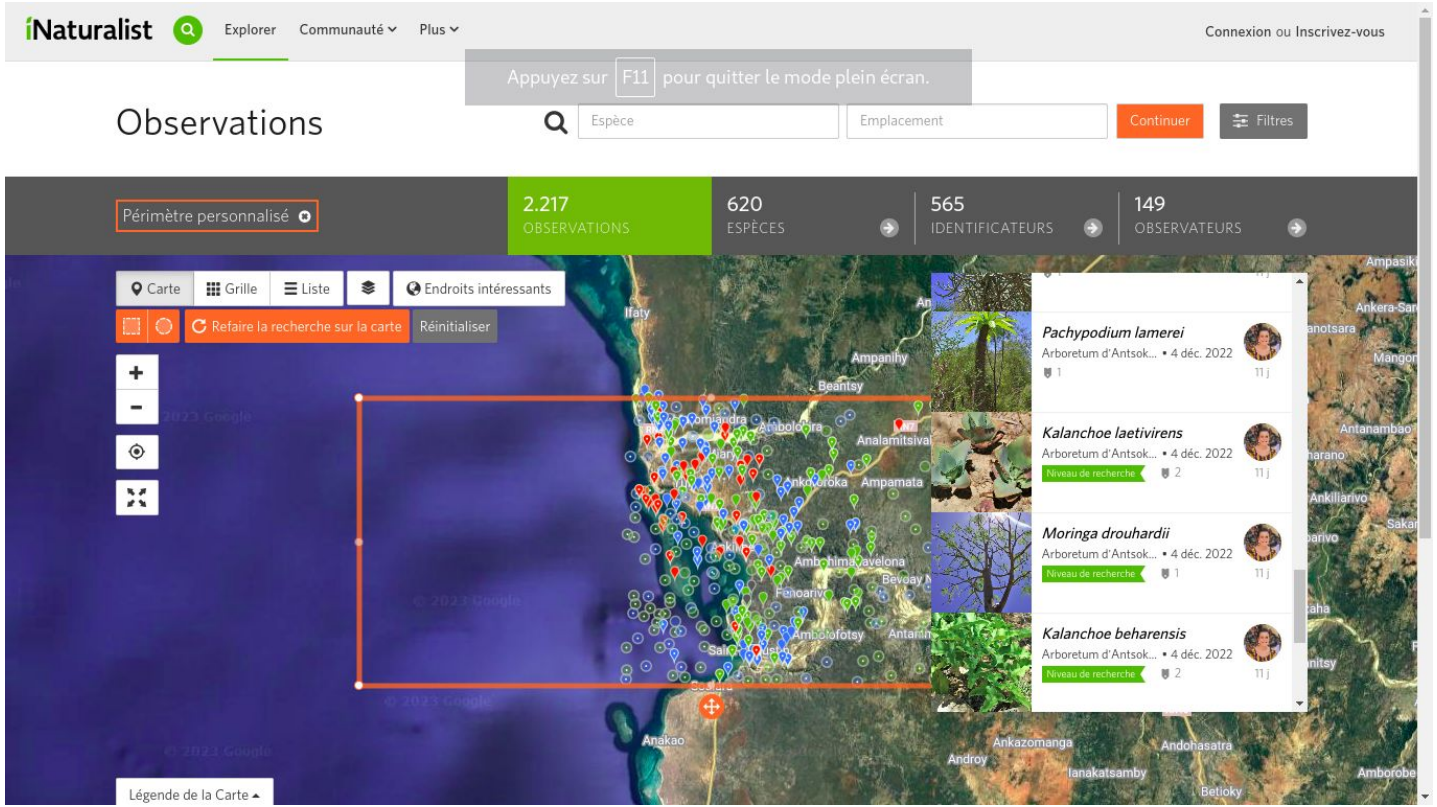
(This article belongs to the Section Marine Diversity)

Download Browse Figures Versions Notes

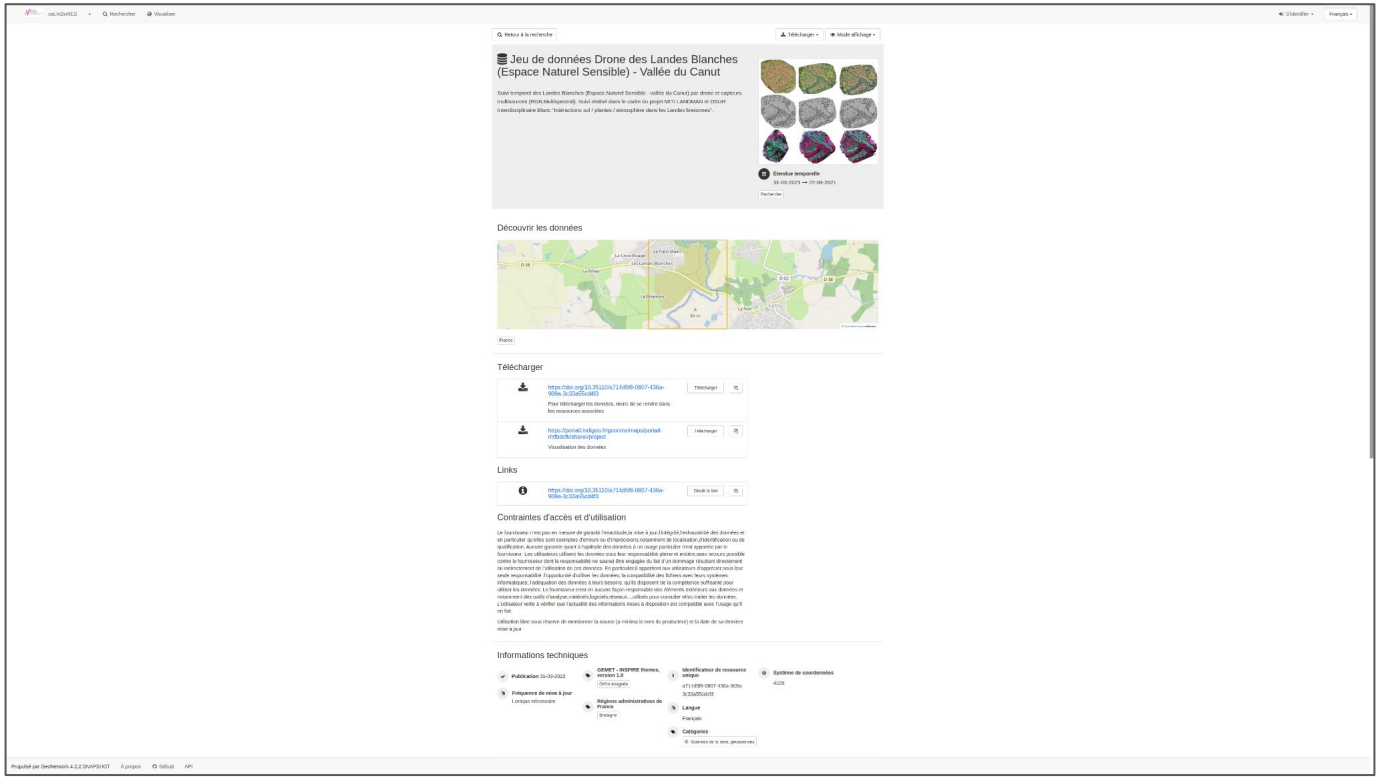
Abstract

Madagascar is a marine biodiversity hotspot. A recent checklist recorded 1689 marine or transitional water fish species, 2.5% being endemic. To date, studies in this country were mostly focused on adult fishes using morphological-based identification. The early life stages of fishes remain largely understudied. The present work aimed to improve knowledge of fish biodiversity in Madagascar by focusing on post-larval reef fishes and settled juveniles in seagrass meadows of southwest Madagascar by using either species identification keys or DNA barcoding. Up to 119,500 individuals were collected, and 1096 individuals were successfully barcoded. We identified 387 species—85 through their morphology (with 58 unsuccessfully sequenced) and 302 by using CO1 barcoding corresponding to 302 barcode index numbers (BINs). This study added 27 new BINs for the BOLD database, 120 new for Madagascar, but only 159 were assigned a precise species name. By referring to the updated checklist of Madagascar fishes, 10 new species were detected for Madagascar. This number will probably increase when all the barcoded specimens become assigned to precise species names. These preliminary findings stress our poor knowledge of marine fish biodiversity in Madagascar and demonstrate the relevance of DNA barcoding in improving this knowledge.

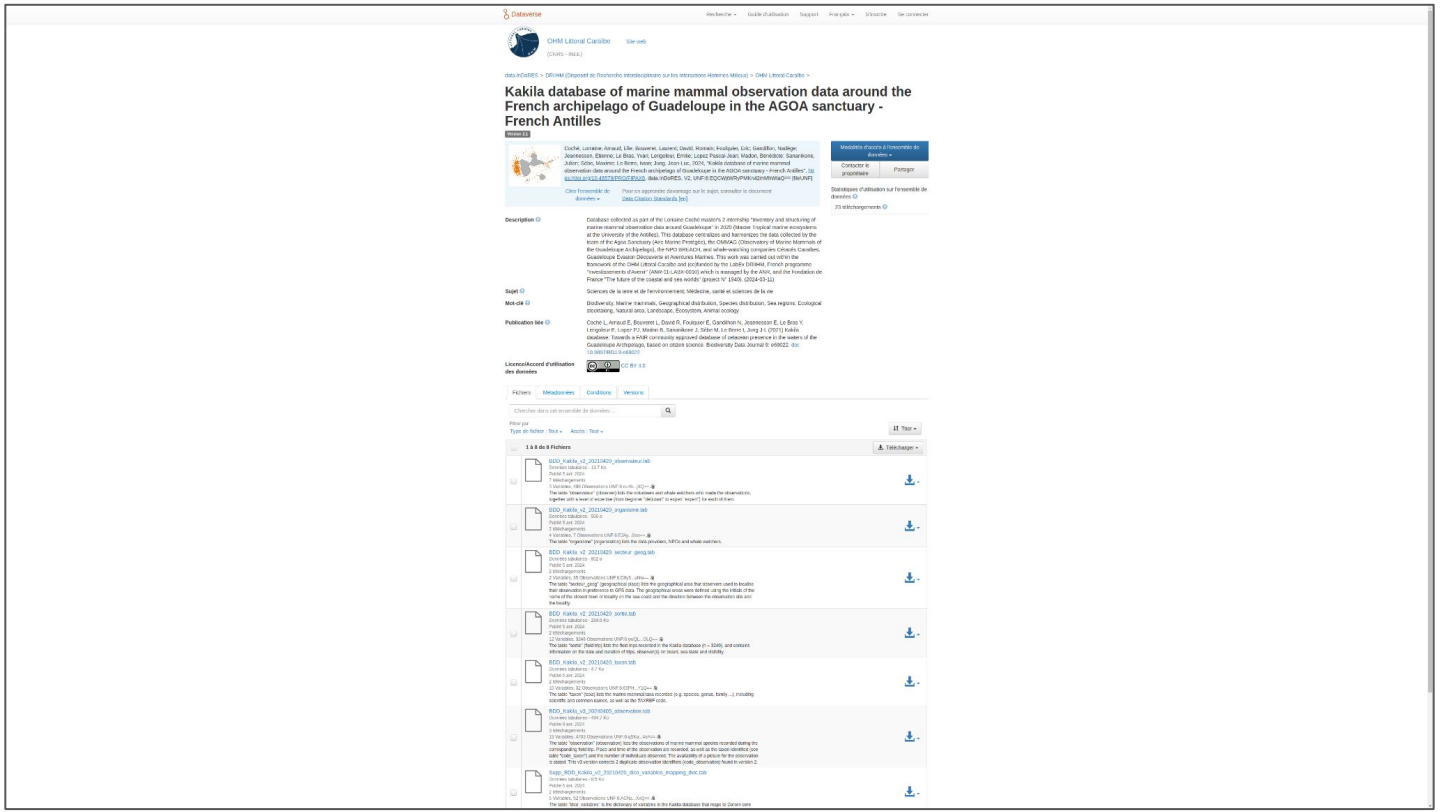
Sciences citoyennes : ex iNaturalist



CNRS [cat.InDoRES]

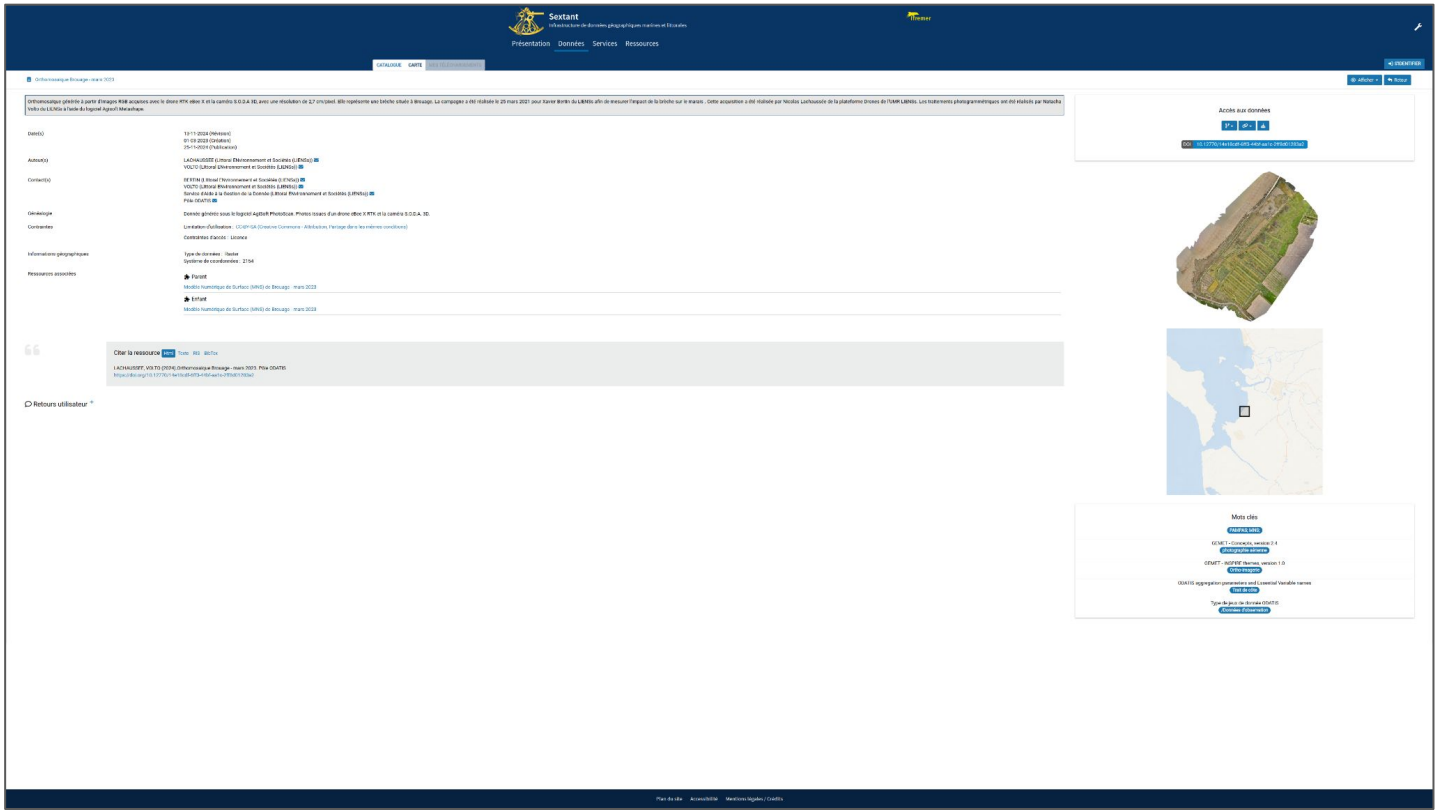


CNRS InDoRES : https://data.indores.fr/

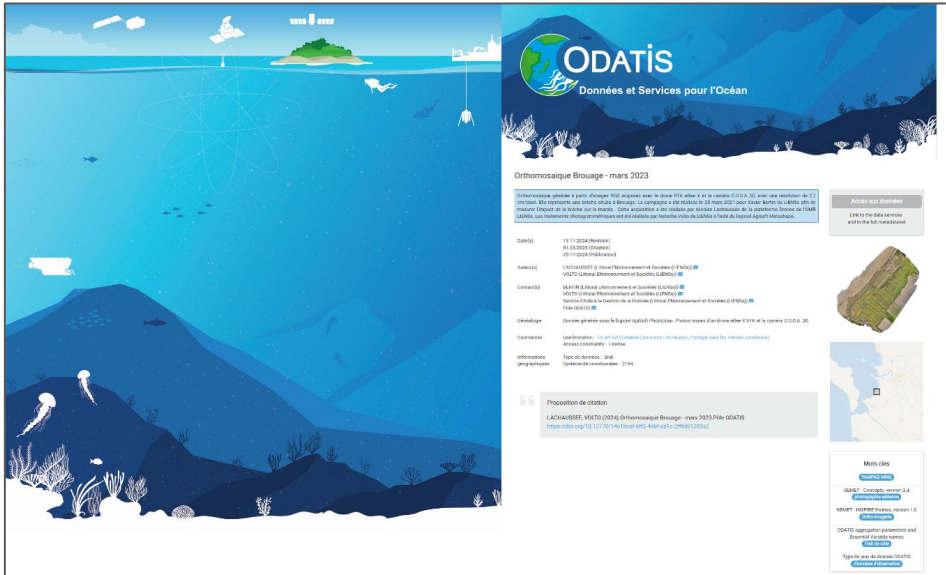


Données [Seanoe]

Données Sextant

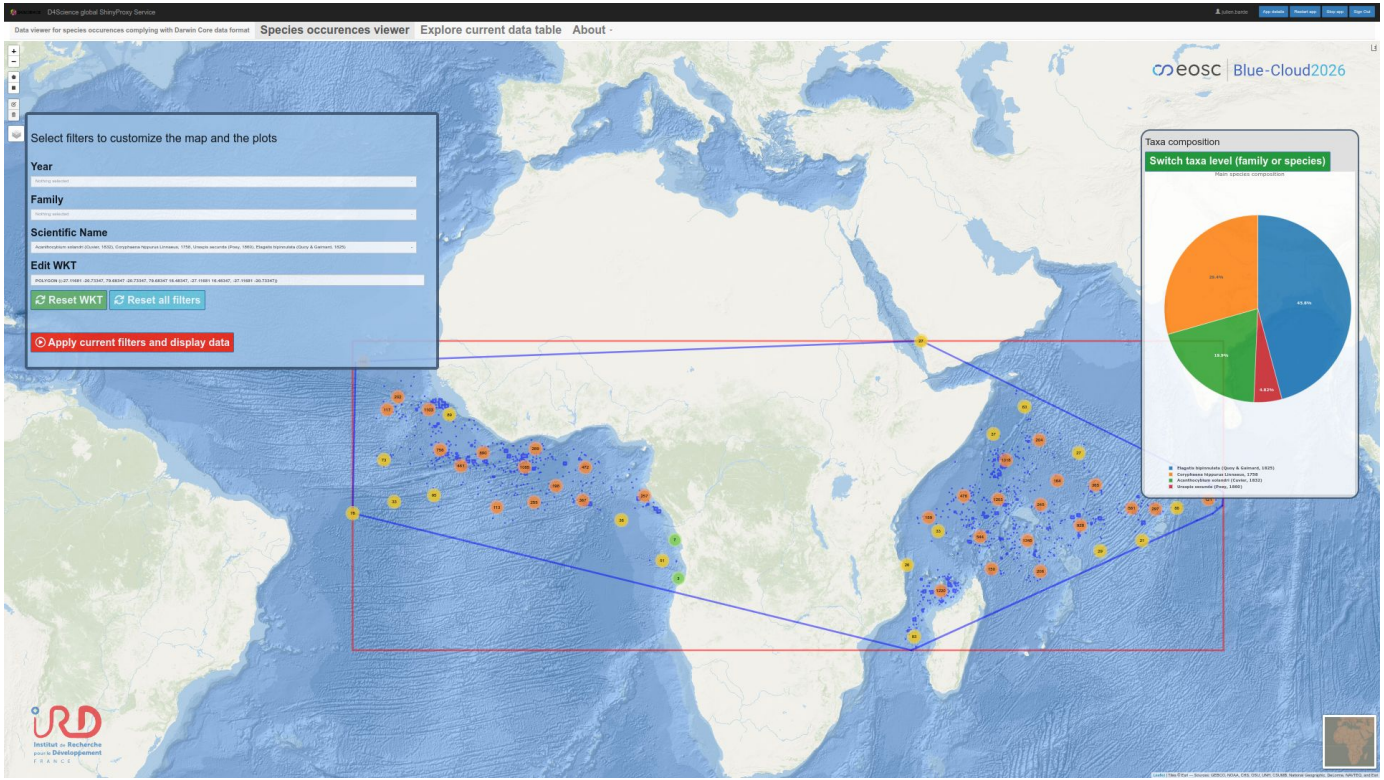


DOI ODATIS

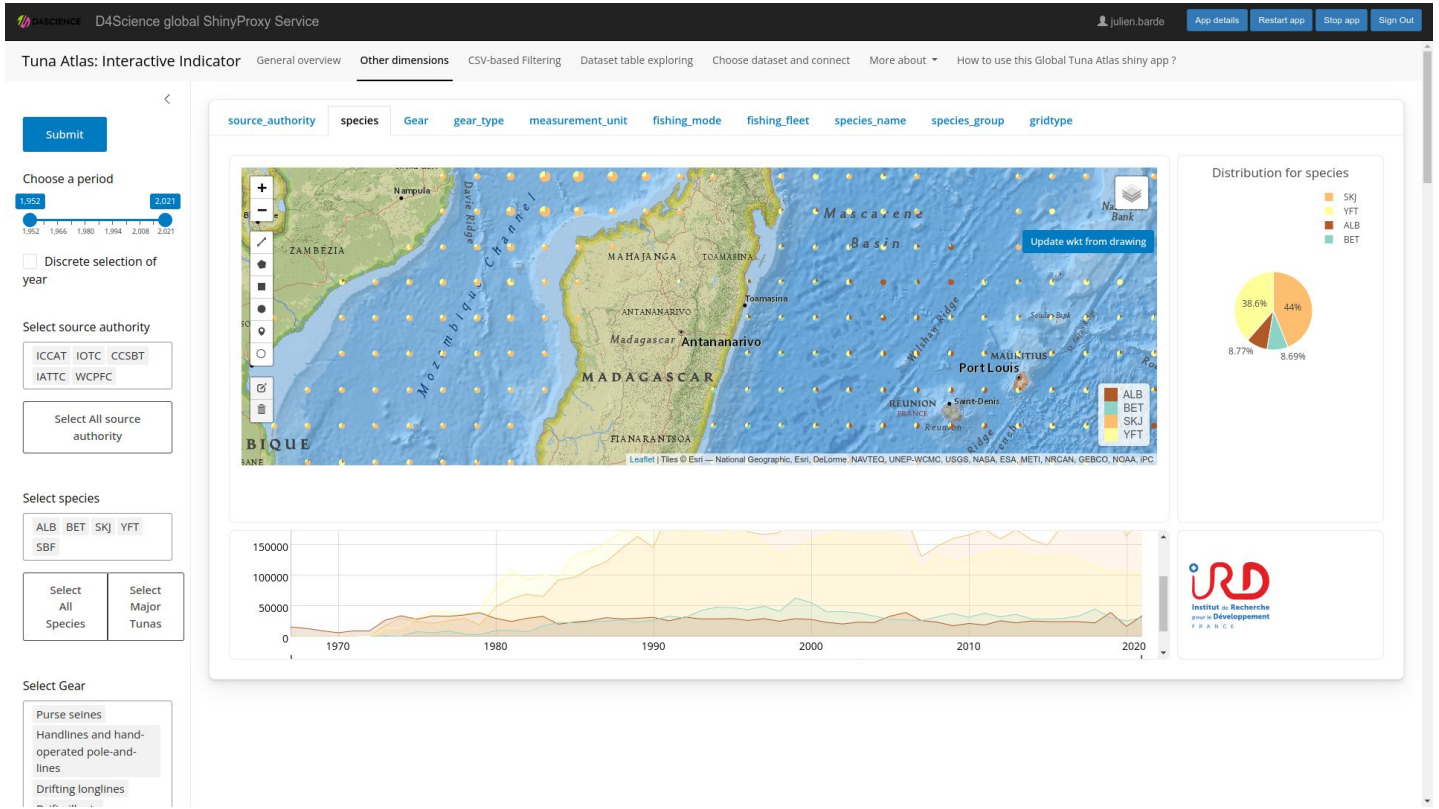


Exemple des données de drones CNRS / Ifremer / IRD…

Applis Shiny reproductibles



Applis Shiny reproductibles



Conclusion

Les DOIs….et les entrepôts de données…

DOI 10.5281/zenodo.14795560

- Au service de la pérennité et reproductibilité
- Le choix du bon entrepôt dépend du contexte :
  - projets, partenariats..
  - Type de données : biodiversité, géomatique, domaine marin..
- Obtention facile et plus ou moins longue selon la qualité de ce qu'on propose
- Suffisant pour mettre en oeuvre un Plan de Gestion de Données
- Faible participation des scientifiques faute de contraintes :
  - D'accord sur le principe mais pas de temps
  - Volonté affichée de ne pas partager
- Mettre des DOIs pour mieux archiver les données des stockages MARBEC ?
- Question fil rouge => bonnes raisons de ne pas le faire ?

Questions ?

- Ma question fil rouge => bonnes raisons de ne pas mettre des DOIs ?
- Questions des participants ?