

Galaxy pour le système terre – Perspectives



DATA
TERRA



eosc



Du monde de la biodiversité au système terre

Du PNDB au projet GAIA DATA

2022->



DATA
TERRA



**E-Infrastructure nationale distribuée et
fédérée**

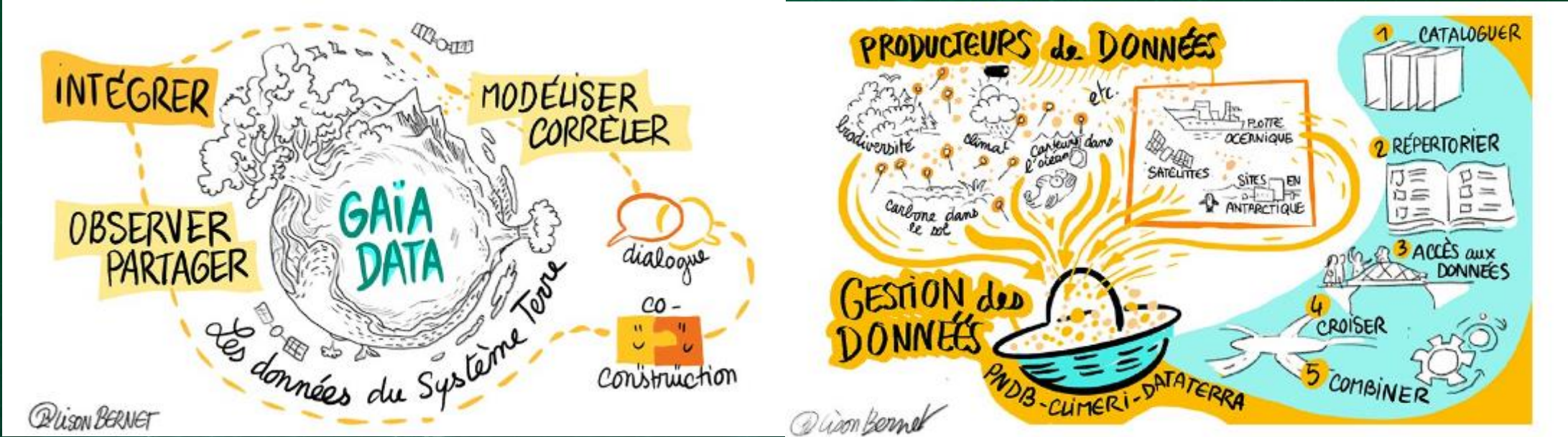
au service d'une science collaborative

→ “cloud pour la science”

Du monde de la biodiversité au système terre

Du PNDB au projet GAIA DATA

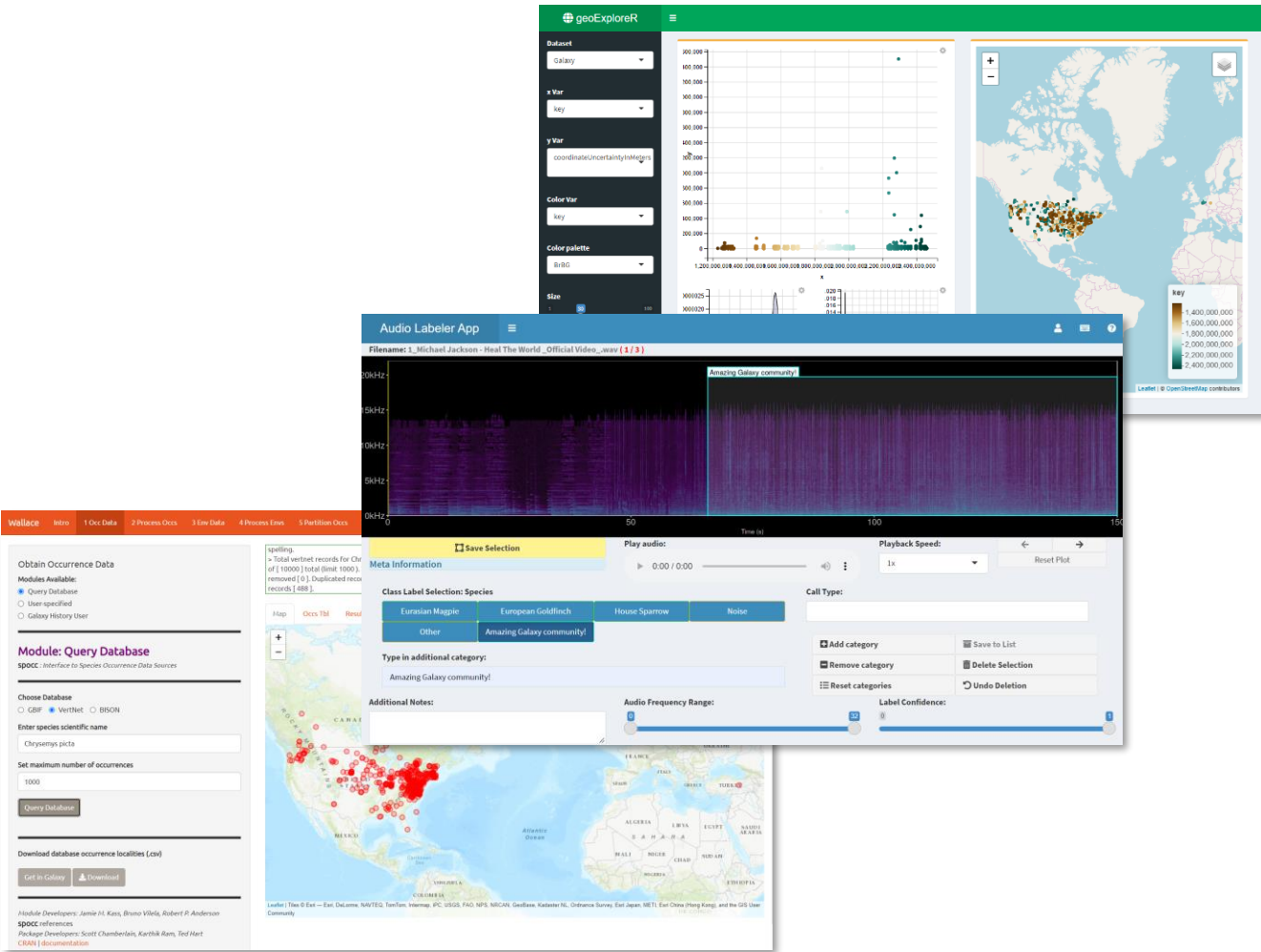
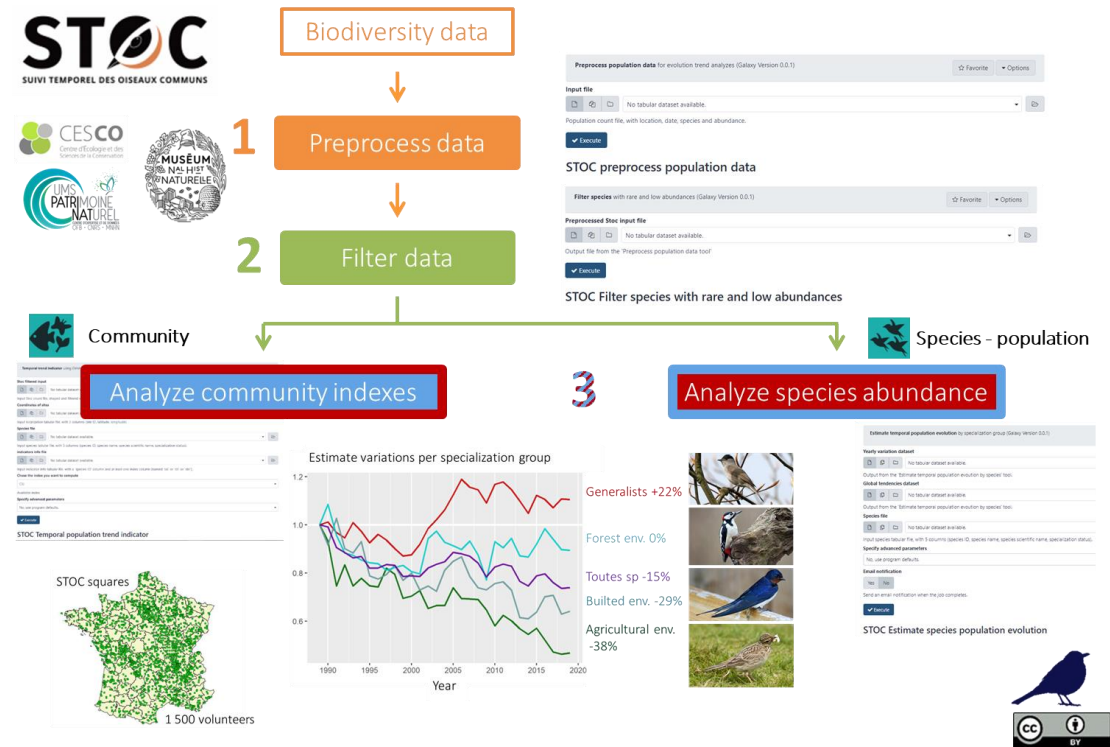
2022->



Partager des outils, mobiliser les utilisateurs

Galaxy-E comme un « démonstrateur »

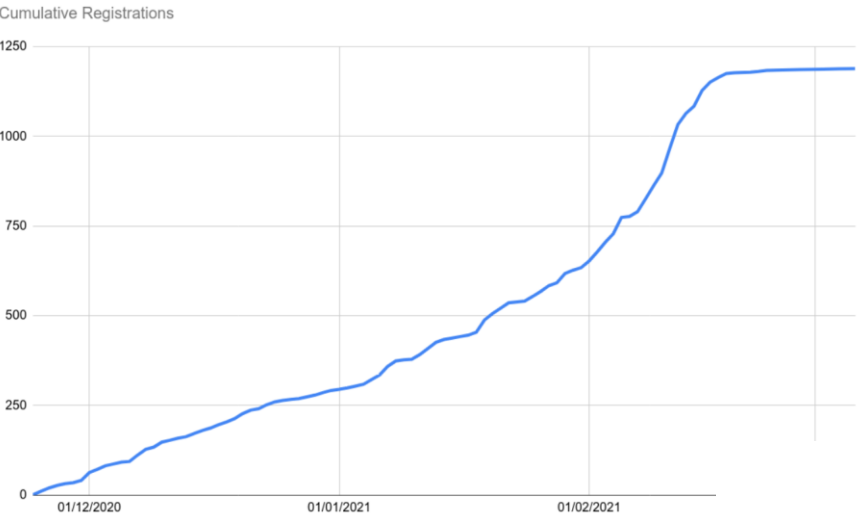
2023



Brique essentielle pour contribuer à des initiatives de grande ampleur

Projet européen sur la culture de la donnée et de son analyse

2019-2023





Gallantries

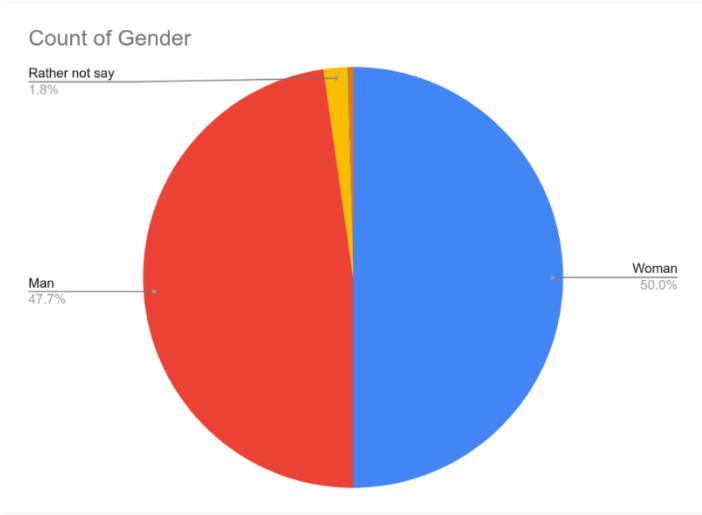
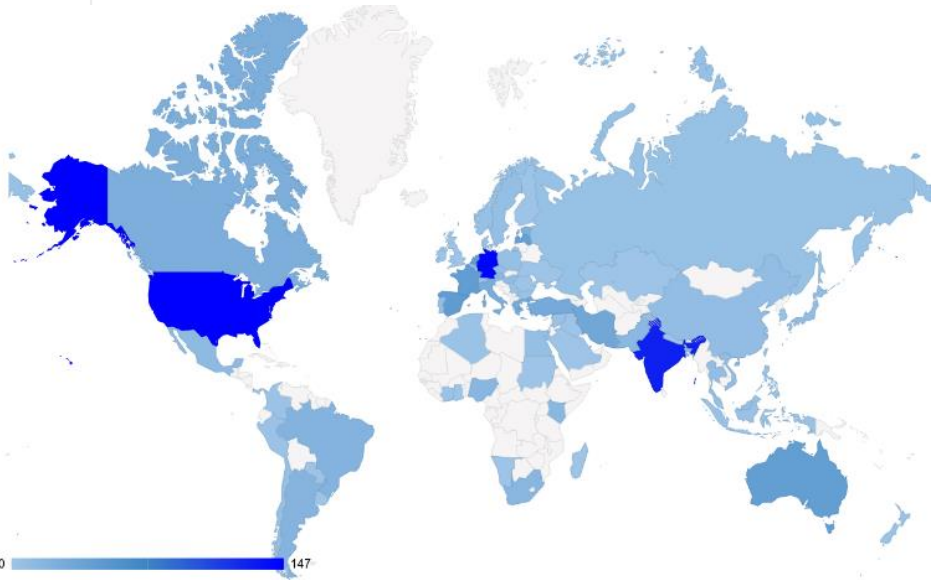


Good Practice



With the support of the Erasmus+ Programme of the European Union

This project (2020-1-NL01-KA203-064717) is funded with the support of the Erasmus+ programme of the European Union.



Massive training week

- 1200 registrations 2021
- 3300 registrations 2022!

Brique essentielle pour contribuer à des initiatives de grande ampleur

Des projets pour la mise en place d'un cloud Européen science ouverte

2022->2025



Multidisciplinary use cases

Earth & Environmental
Dynamics

Read more



Environmental Bio-
geochemical Assets

Read more



Biodiversity
Observation

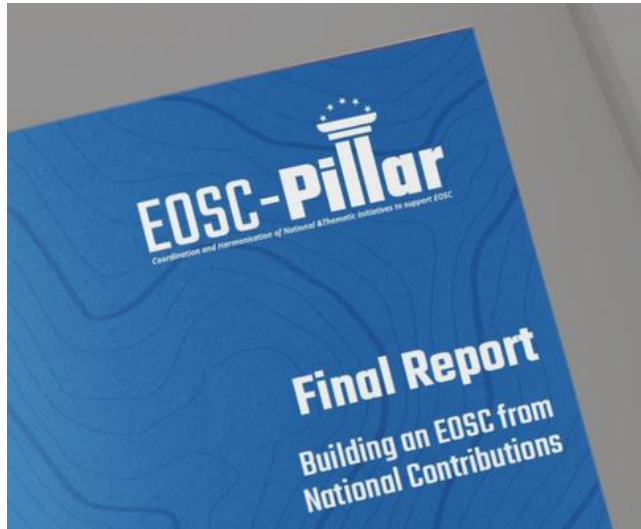
Read more



Brique essentielle pour contribuer à des initiatives de grande ampleur

Des projets pour la mise en place d'un cloud Européen science ouverte

2019->2025



Scientific Collaboration Agreement between
EOSC-Pillar and EOSC-Life



 EOSC + THE ASSOCIATION + NEWS & EVENTS + EOSC FOCUS

Stronger together: FAIR-EASE and EuroScienceGateway join forces

06 February 2023

  Two Horizon Europe projects directed toward the development and c Science Cloud (EOSC)—[FAIR-EASE](#) and [EuroScienceGateway](#) (ESG)— over the next three years in order to extend the open source [Galaxy](#) pl broader group of scientific user communities.

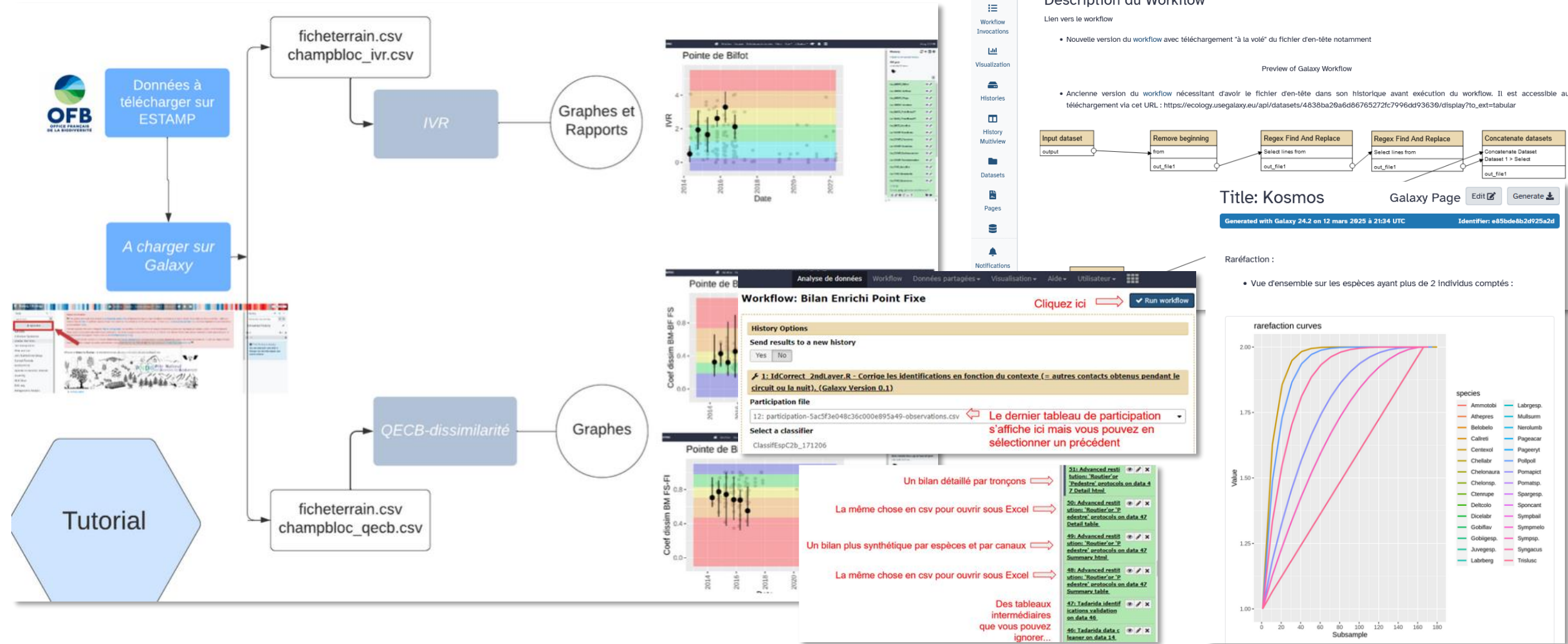
  The cooperation will include shared compute resources, joint work

Galaxy as a cross-EOSC platform

New scientific collaboration agreement between EOSC-Life and EOSC-Pillar

Science participatives

Support aux projets de sciences participatives



Science participatives

Crowdsourcing / annotation de données par la foule

2020-2022



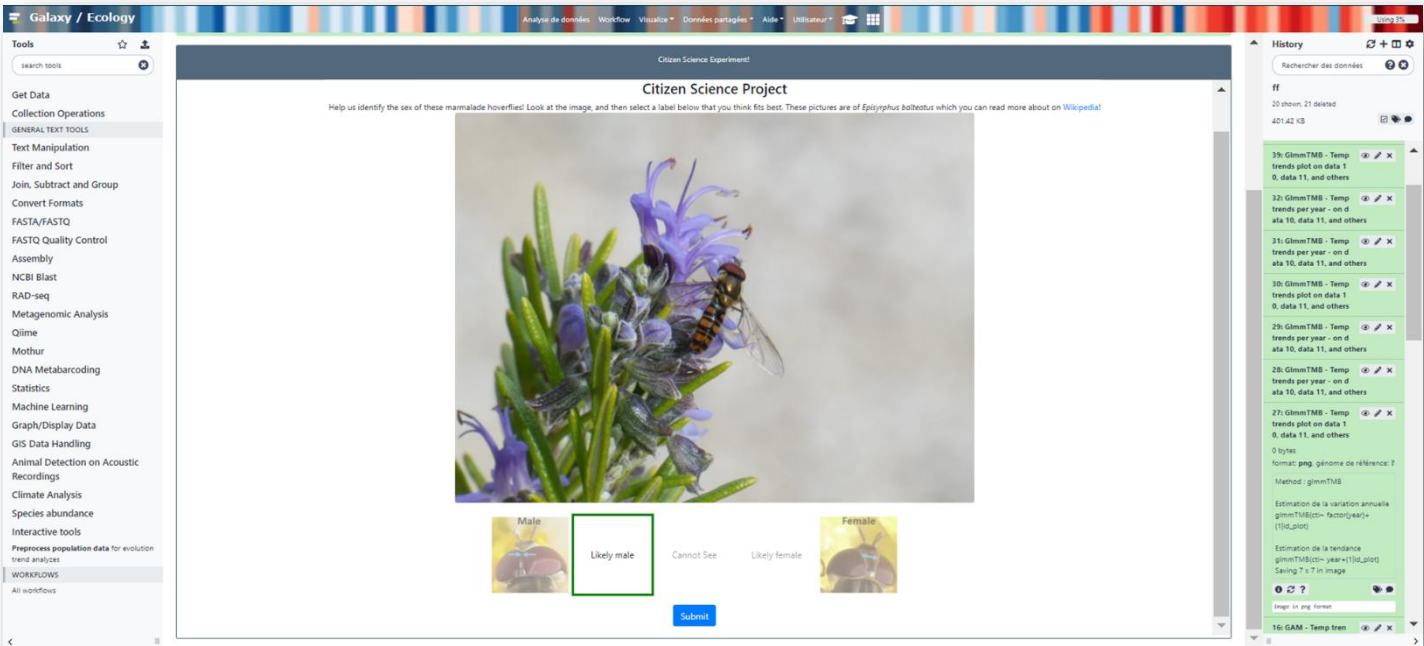
GAPARS project

Gamification of participatory science for training and education purposes

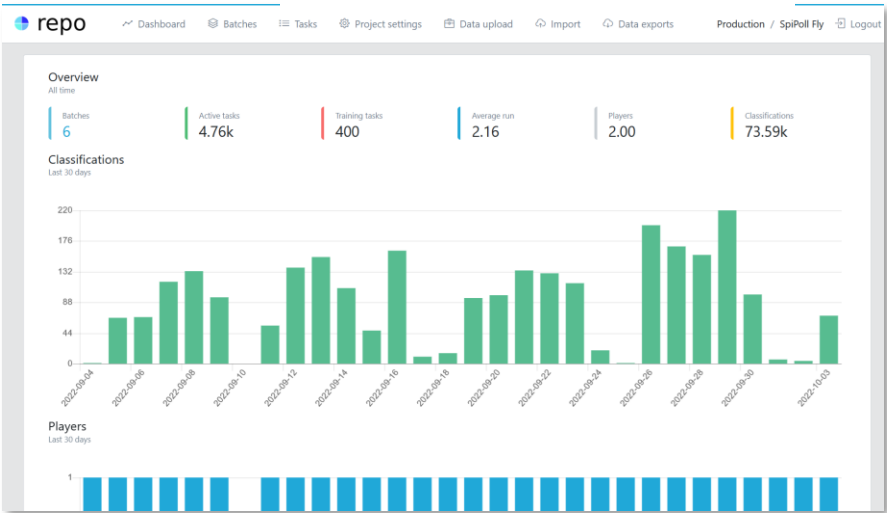
Citizen Science for Serious Gamers



MOODA concept (Massively Open Online Data Analysis)



Crowdsourcing with hoverflies (syphres)
images from SPIPOLL project



203 690 classifications



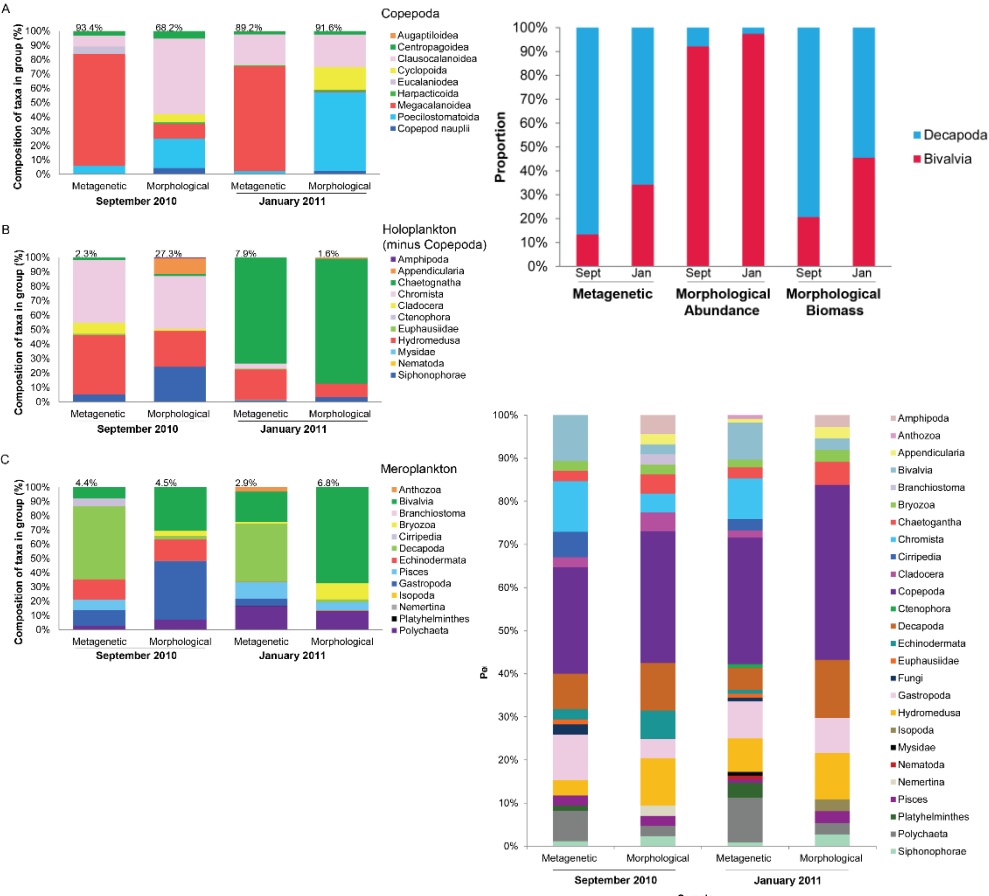
Science participatives

Analyse complexe de données par les citoyens

Next Generation Sequencing Reveals the Hidden Diversity of Zooplankton Assemblages

Penelope K. Lindeque, Helen E. Parry, Rachel A. Harmer, Paul J. Somerfield, Angus Atkinson

Published: November 7, 2013 • <https://doi.org/10.1371/journal.pone.0081327>



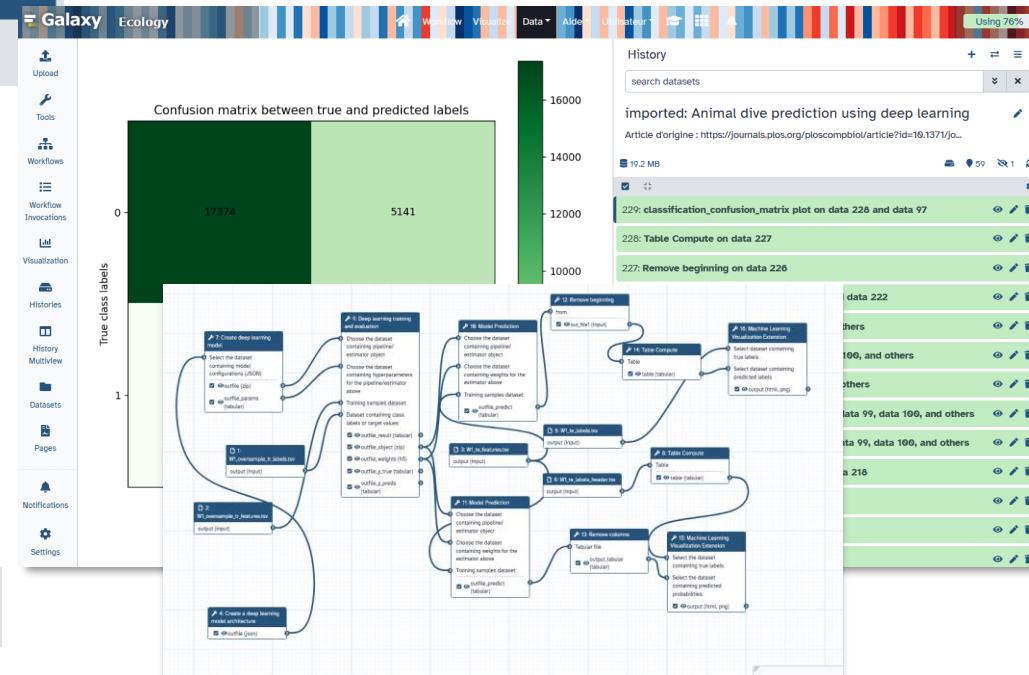
Annotations et IA

Adapter des outils existant pour nos cas d'usage

The diagram illustrates a three-step workflow for YOLO model training using AnyLabeling and Ultralytics:

- AnyLabeling Interactive**
Effortless data labeling with AI support from YOLO and Segment Anything!
Screenshot: The interface shows audio labeling for 'Hansel and Gretel' and image labeling for '17: Concarneau-station-crabe-blennie-crevettes.jpg'.
Run Tool
- Perform YOLO training with ultralytics**
Screenshot: The Ultralytics YOLO training interface.
Run Tool
- Convert AnyLabeling JSON to YOLO text**
Screenshot: The AnyLabeling JSON to YOLO text conversion interface.
Run Tool

Reproduire les analyses à partir d'article de recherche



PLOS COMPUTATIONAL BIOLOGY

 OPEN ACCESS  PEER-REVIEWED

RESEARCH ARTICLE

Deep inference of seabird dives from GPS-only records: Performance and generalization properties

Amédée Roy , Sophie Lanco Bertrand, Ronan Fablet

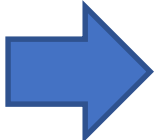
Version 2  Published: March 11, 2022 • <https://doi.org/10.1371/journal.pcbi.1009890>

Nouvelle méthode pour créer et partager des workflows

By Arthur BARREAU Supervised by Marc ELEAUME and Yvan LE BRAS

AUTOMATISER LE SUIVI DE L'ÉTAT DES ECOSYSTEM EN L'ANTARCTIQUE : AVEC UN WORKFLOW GALAXY

MUSEUM STATION MARINE CONCARNEAU PNDP Pôle National de Données de Biodiversité Université de Rennes



Un workflow pour automatiser la génération d'un rapport sur l'état des écosystèmes Antarctique

Galaxy Ecology

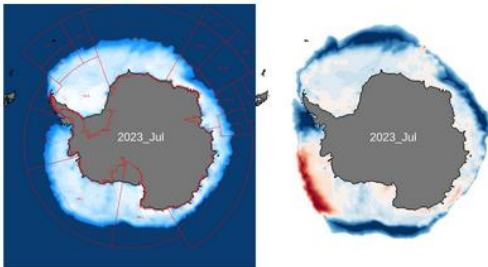
Workflow: fig3 (Version: 1) Run Workflow

edited 9 days ago Arthur Barreau never run

FIG3: MONTHLY SEA ICE CONCENTRATION AND ANOMALY MAP

☒ Set this option to Yes to include the Fig1 visualization in the output. If set to No, this figure will be skipped.

This figure, generates two maps for a selected month: one showing average sea ice concentration with the 1981–2010 median extent, and one showing anomalies relative to the climatological baseline.



Area Selection Fig3 * ✓

This input allows you to choose which zones will be used or displayed for the data. There are two methods for defining the zones:

1. Using Antarctic Subareas:

- 881, 882, 883, 481, 482, 483, 484, 485, 486, 5841, 5842, 5843a, 5843b, 5844a, 5844b, 5851, 5852, 586, 587

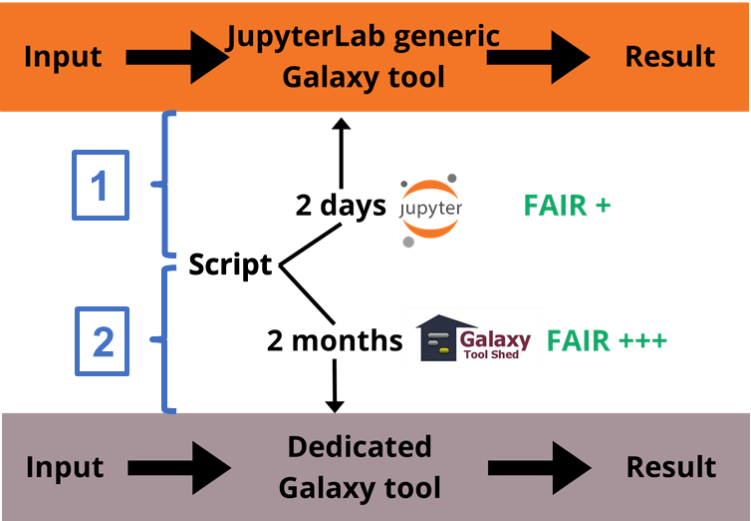
2. Using a Geographical Location (North, South, West, East coordinates):

- N-45, S-90, W-180, E180

The order of the letters and the number of zones are not important, but the values must be **separated by commas**.

881, 882, 883, 481, 482, 483, 484, 485, 486, 5841, 5842, 5843a, 5843b, 5844a, 5844b, 5851, 5852, 586, 587 ✓

Une nouvelle approche à 2 étapes pour proposer des workflows Galaxy à partir de scripts bruts / codes sources



Nouvelle méthode pour créer et partager des workflows

```
library(data.table)
library(stringr)
library(tidy)
library(dplyr)
library(mgcv)
library(gam4)

# ** dataset ----
data <- fread("galaxy_inputs/P10/P10_ER90_Final_COMPLETED.csv.txt")
data <- data[,1]
data <- as.data.frame(data)
data <- data[which(data$id_site_annee %in% c("81124-E04_2021", "81124-E03_2021")),]
colnames(data)[25] <- "diametre_rotor"
colnames(data)[26] <- "hauteur_nacelle"

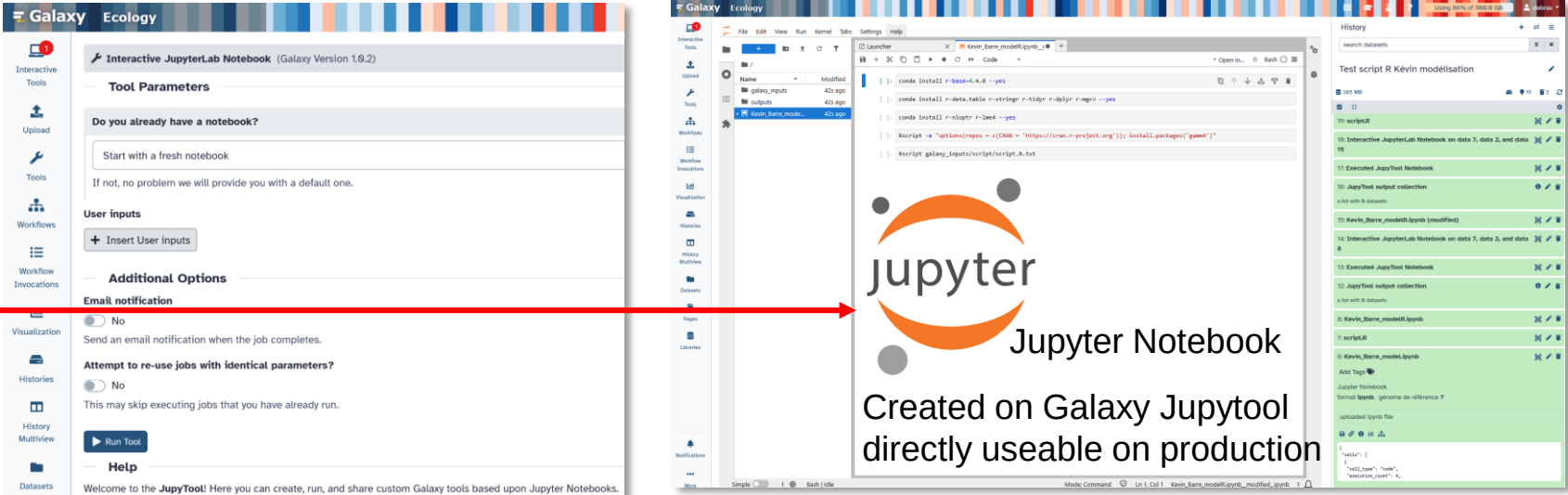
data$MRE <- as.numeric(data$Pippip_90)+
as.numeric(data$Pippy_90)+
as.numeric(data$Pipnat_90)+
as.numeric(data$Hypsav_90)+
as.numeric(data$Pipkuh_90)+
as.numeric(data$Minsch_90)

data$MRE_binom <- rep(NA,dim(data)[1])
data$MRE_binom[which(data$MRE > 0)] = 1
data$MRE_binom[which(data$MRE == 0)] = 0

data$materiel_tot <- paste0(data$materiel_corrige,"_",data$`trigger`db`)

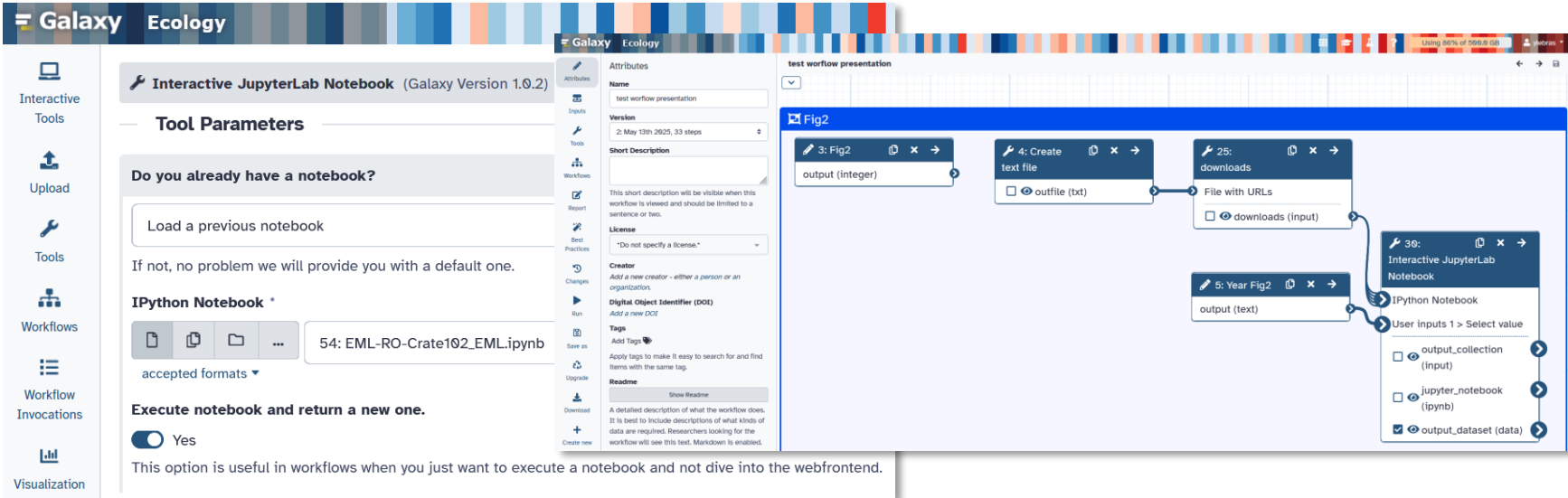
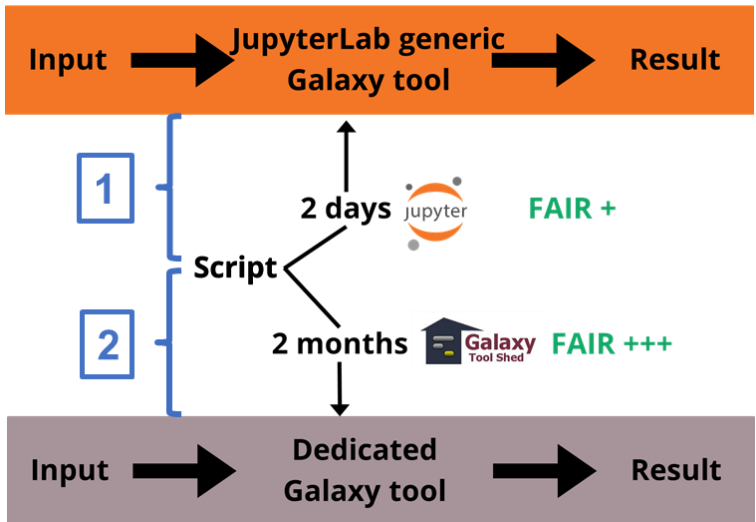
data_10 <- data %>%
  drop_na(jday,
    pourcentage_nuit,
    rpm_60min,
    vent_60min,
    temperature_60min,
    rr_60min,
    annee,
    Code_eolienne_meteo,
    materiel_tot)

# ** Model ----
m <- gamm(MRE_binom ~ s(jday, k=4), family = binomial, data = data_10, random=list(annee=~1))
m <- gamm4(MRE_binom ~ s(jday, k=4), family = binomial, data = data_10, random=~(1|annee))
```



Exécution de Jupyter Notebook

Exécution de Jupyter Notebook dans workflows



Tutoriels en ligne dédiés système terre



GTN:500127	Slides	An earth-system RO-Crate earth-system ro-crate ocean marine omics	FAIR Data, Workflows, and Research	1H
GTN:T00455	Hands-on	Ecoregionalization workflow tutorial ecology taxonomic data EBV workflow modeling gbif ocean earth-system interactive-tools biodiversity	Ecology	2H
GTN:T00450	Hands-on	Marine Omics identifying biosynthetic gene clusters earth-system ocean marine omics biodiversity	Ecology	3H
GTN:T00367	Hands-on	QGIS Web Feature Services earth-system GIS Geographical Information System WFS Spatial data Maps OGC biodiversity	Ecology	0H30M
GTN:T00400	Hands-on	From NDVI data with OpenEO to time series visualisation with Holoviews earth-system land degradation NDVI copernicus holoviews	Ecology	1H
GTN:T00332	Hands-on	Obis marine indicators earth-system ocean marine omics biodiversity	Ecology	1H
GTN:T00529	Hands-on	Adding file-sources to Galaxy earth-system ocean data work-in-progress		
GTN:T00473	Hands-on	Create a subdomain for your community on UseGalaxy.eu subdomain community earth-system interactive-tools		
Galaxy Server				
GTN:T00476	Hands-on	Make your tools available on your subdomain subdomain earth-system interactive-tools	Galaxy Community Building	30m
GTN:T00375	Hands-on	Ocean Data View (ODV) earth-system ocean geographical information system ODV netcdf data maps marine data	Climate	1H
GTN:T00457	Hands-on	Getting your hands-on earth data earth-system ocean land atmosphere biodiversity	Climate	1H
GTN:T00376	Hands-on	Ocean's variables study earth-system ocean netcdf data	Climate	1H
GTN:T00396	Hands-on	Sentinel 5P data visualisation earth-system volcano satellite data	Climate	1H
GTN:T00546	Hands-on	Ocean's variables study earth-system ocean biogeochemical work-in-progress	Climate	1H
GTN:T00426	Hands-on	Analyse Argo data earth-system ocean pangeo argo	Climate	1H

<https://tinyurl.com/galaxy-es>

Equipe Galaxy système terre



Alessandro Rizzo (IRD)

Marie Jossé (CNRS)

Erwan Bodéré (IFREMER)

Jérôme Détoc (IFREMER)

Sylvain Grellet (BRGM / One Water)

Eric Lecaude (BRGM / One Water)



Gwenaél Le Caër (IFREMER)

CDD BRGM One Water

Yvan Le Bras (MNHN)



Pauline Segueineau (MNHN)



Karim Ramage (IPSL)

Annexes



We,EXPLORE. **FDJ UNITED**



et pourquoi pas vous ?

Navigateurs, scientifiques, citoyens et décideurs
tous unis pour
l'Océan

EN MER

*Collecte, numérisation,
transmission*



We,EXPLORE.



FairScope



A TERRE

*Réception, traitement
et analyse*

EcoTaxa^{2.7}



FDJ UNITED

PlayforPlankton

EXPLORE.

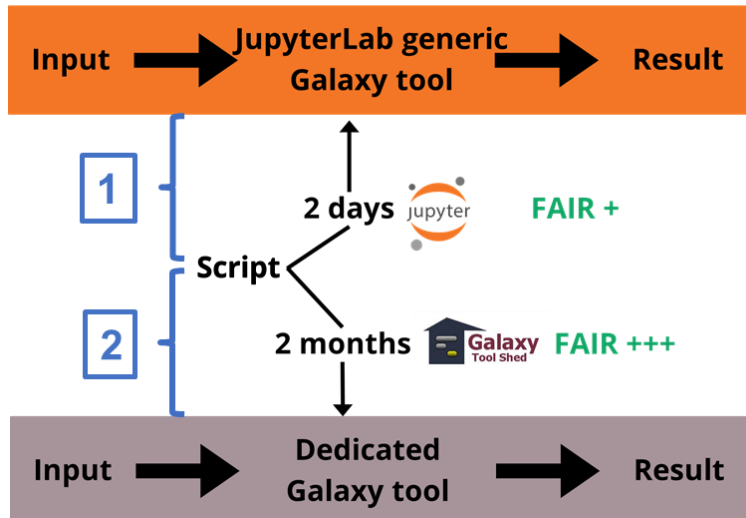
PROGRAMME SCIENCES CITOYENNES

Etapes du process & parties prenantes



Nouvelle méthode pour créer et partager des workflows

Une nouvelle approche à 2 étapes pour proposer
des workflows Galaxy à partir de scripts bruts /
codes sources



```
library(data.table)
library(stringr)
library(tidyr)
library(dplyr)
library(mgcv)
library(gamm4)

# ** dataset ----
data <- fread("galaxy_inputs/P10/P10_ER90_Final_COMPLETED.csv.txt")
data <- data[, -1]
data <- as.data.frame(data)
data <- data[which(data$id_site_annee %in% c("81124-E04_2021", "81124-E03_2021")), ]
colnames(data)[25] <- "diametre_rotor"
colnames(data)[26] <- "hauteur_nacelle"

data$MRE <- as.numeric(data$Pippip_90)+
  as.numeric(data$Pippig_90)+
  as.numeric(data$Pipnat_90)+
  as.numeric(data$Hypsav_90)+
  as.numeric(data$Pipkuh_90)+
  as.numeric(data$Hinsch_90)

data$MRE_binom <- rep(NA, dim(data)[1])
data$MRE_binom[which(data$MRE > 0)] = 1
data$MRE_binom[which(data$MRE == 0)] = 0

data$materiel_tot <- paste0(data$materiel_corrige, "_", data$trigger dB")

data_10 <- data %>%
  drop_na(jday,
    pourcentage_nuit,
    rpm_60min,
    vent_60min,
    temperature_60min,
    rr_60min,
    annee,
    Code_eolienne_meteo,
    materiel_tot)

# ** Model ----
m <- gamm(MRE_binom ~ s(jday, k=4), family = binomial, data = data_10, random=list(annee~1))
m <- gamm4(MRE_binom ~ s(jday, k=4), family = binomial, data = data_10, random=~(1|annee))
```

```
library(data.table)
library(stringr)
library(tidyr)
library(dplyr)
library(mgcv)
library(gamm4)
```

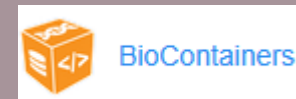
bioconda / packages / r-data.table 1.11.6
bioconda / packages / r-stringr 1.1.0
bioconda / packages / r-dplyr 0.5.0

No conda package!



conda recipe creation

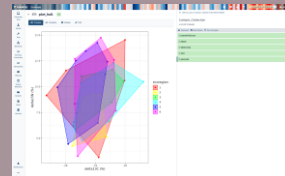
New conda package



Creation of a Biocontainer
with all dependencies

Galaxy tool xml
+
adapted R script

Tool in a
production
Galaxy instance



Nouvelle méthode pour créer et partager des workflows

```
library(data.table)
library(stringr)
library(tidy)
library(dplyr)
library(mgcv)
library(gamm4)

# ** dataset ----
data <- fread("galaxy_inputs/P10/P10_ER90_Final_COMPLETED.csv.txt")
data <- data[,1]
data <- as.data.frame(data)
data <- data[which(data$id_site_annee %in% c("81124-E04_2021", "81124-E03_2021")),]
colnames(data)[25] <- "diametre_rotor"
colnames(data)[26] <- "hauteur_nacelle"

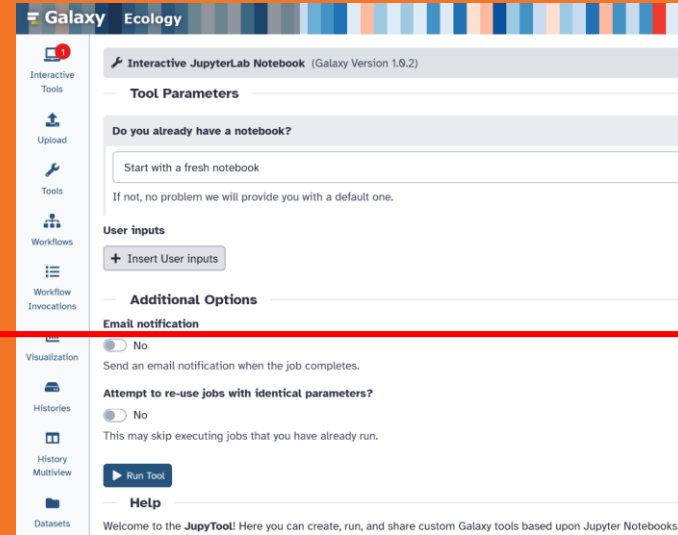
data$MRE <- as.numeric(data$Pippip_90)+
as.numeric(data$Pippyp_90)+
as.numeric(data$Pipnat_90)+
as.numeric(data$Hypsav_90)+
as.numeric(data$Pipkuh_90)+
as.numeric(data$Minsch_90)

data$MRE_binom <- rep(NA,dim(data)[1])
data$MRE_binom[which(data$MRE > 0)] = 1
data$MRE_binom[which(data$MRE == 0)] = 0

data$materiel_tot <- paste0(data$materiel_corrige,"_",data$`trigger`db`)

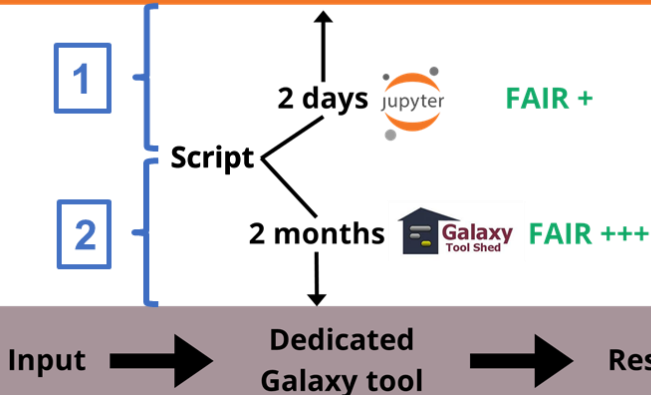
data_10 <- data %>%
  drop_na(jday,
    pourcentage_nuit,
    rpm_60min,
    vent_60min,
    temperature_60min,
    rr_60min,
    annee,
    Code_eolienne_meteo,
    materiel_tot)

# ** Model ----
m <- gamm(MRE_binom ~ s(jday, k=4), family = binomial, data = data_10, random=list(annee=~1))
m <- gamm4(MRE_binom ~ s(jday, k=4), family = binomial, data = data_10, random=~(1|annee))
```



jupyter Jupyter Notebook
Created on Galaxy Jupyter tool
directly useable on production

Input → JupyterLab generic Galaxy tool → Result



```
library(data.table)
library(stringr)
library(tidy)
library(dplyr)
library(mgcv)
library(gamm4)

# ** dataset ----
data <- fread("galaxy_inputs/P10/P10_ER90_Final_COMPLETED.csv.txt")
data <- data[,1]
data <- as.data.frame(data)
data <- data[which(data$id_site_annee %in% c("81124-E04_2021", "81124-E03_2021")),]
colnames(data)[25] <- "diametre_rotor"
colnames(data)[26] <- "hauteur_nacelle"

data$MRE <- as.numeric(data$Pippip_90)+
as.numeric(data$Pippyp_90)+
as.numeric(data$Pipnat_90)+
as.numeric(data$Hypsav_90)+
as.numeric(data$Pipkuh_90)+
as.numeric(data$Minsch_90)

data$MRE_binom <- rep(NA,dim(data)[1])
data$MRE_binom[which(data$MRE > 0)] = 1
data$MRE_binom[which(data$MRE == 0)] = 0

data$materiel_tot <- paste0(data$materiel_corrige,"_",data$`trigger`db`)

data_10 <- data %>%
  drop_na(jday,
    pourcentage_nuit,
    rpm_60min,
    vent_60min,
    temperature_60min,
    rr_60min,
    annee,
    Code_eolienne_meteo,
    materiel_tot)

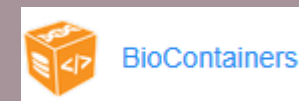
# ** Model ----
m <- gamm(MRE_binom ~ s(jday, k=4), family = binomial, data = data_10, random=list(annee=~1))
m <- gamm4(MRE_binom ~ s(jday, k=4), family = binomial, data = data_10, random=~(1|annee))
```

bioconda / packages / r-data.table 1.11.6
bioconda / packages / r-stringr 1.1.0
bioconda / packages / r-dplyr 0.5.0

No conda package!



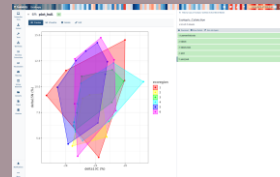
conda recipe creation → New conda package



Creation of a Biocontainer with all dependencies

Galaxy tool xml
+ adapted R script

Tool in a production Galaxy instance



Besoins utilisateurs vs outils « analyse / traitement de données »

Opérationnalisation des EBVs

Mise à disposition workflows de production de métriques de biodiversité

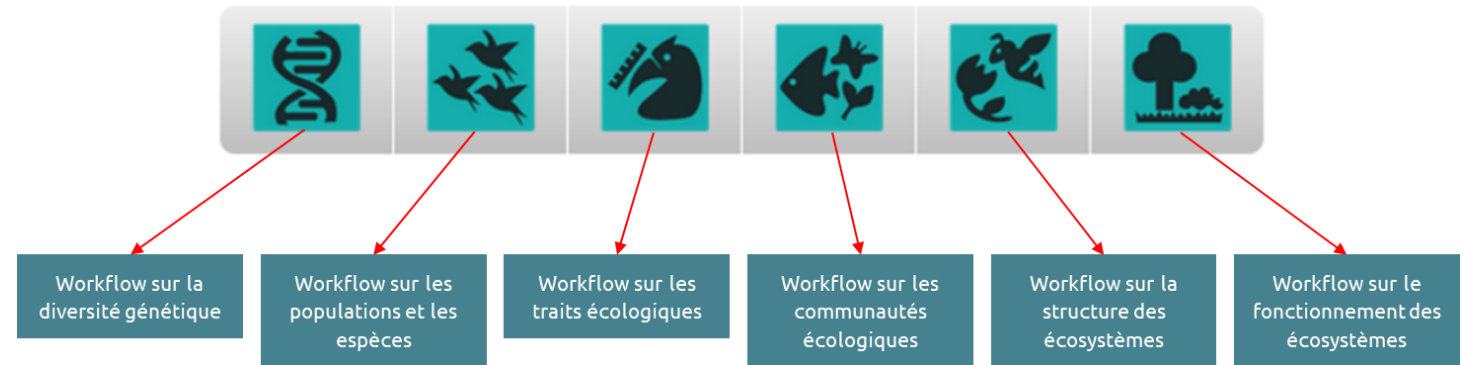
Accès banques de données internationales (occurrence, génétiques, satellites, climat..)

Lien recherche - expertise

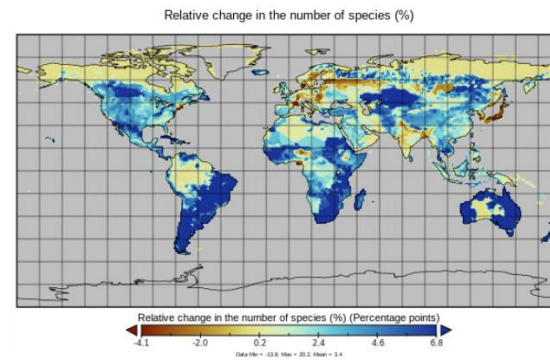
Formation

Visualisation de données

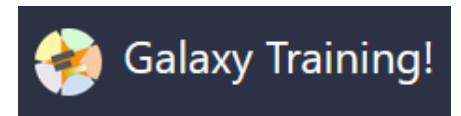
Analyse scientifique de données



<https://www.pndb.fr/fr/activites/projets-techniques-et-scientifiques/ebvosc#Workflow%20examples>



Ecosystème « Galaxy Ecology »



Besoins utilisateurs vs outils « analyse / traitement de données »

Opérationnalisation des EBVs

Mise à disposition workflows de production de métriques de biodiversité

Accès banques de données internationales (occurrence, génétiques, satellites, climat..)

Lien recherche - expertise

Formation

Visualisation de données

Analyse scientifique de données

Get species occurrences data from GBIF, OBIS, ALA, iNAT and others

OBIS occurrences retrieve data

Argo data access for global ocean in situ observing system

SdmPredictors List Layers from environmental predictors for species distribution modeling

Adds environment variables From Copernicus and etopo given geolocalized and timestamped observations

Retrieve bold Functions to search in Bold and download the available sequences of each subtaxa (get_fasta)

Download and Extract Reads in FASTQ format from NCBI SRA

■ Sentinel-5P Level 2

Observations from the Sentinel-5 Precursor satellite of the Copernicus Earth Observation Programme. It contains a polar orbiting satellite that completes 14 orbits of the Earth a day.

■ Coupled Model Intercomparison Project 6

The sixth phase of global coupled ocean-atmosphere general circulation model ensemble

■ CMIP6 GCMs downscaled using WRF

High-resolution historical and future climate simulations from 1980-2100

■ GBIF European region public datasets

The Global Biodiversity Information Facility is an international network and data infrastructure aimed at providing anyone, anywhere, open access to data about all types of life on Earth.

■ NOAA Global Forecast System (GFS)

The Global Forecast System (GFS) is a weather forecast model produced by the National Centers for Environmental Prediction (NCEP).

■ NOAA Unified Forecast System Subseasonal to Seasonal Prototype 5

The Unified Forecast System Subseasonal to Seasonal prototype 5 (UFS S2Sp5) dataset is reforecast data from the UFS atmosphere-ocean.

■ Copernicus Digital Elevation Model (DEM)

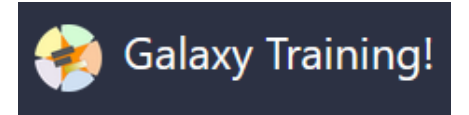
The Copernicus DEM is a Digital Surface Model (DSM) which represents the surface of the Earth including buildings, infrastructure and vegetation.

■ ECMWF ERA5 Reanalysis

ERA5 is the fifth generation of ECMWF atmospheric reanalyses of the global climate created with the best available observation data from satellites, in-situ stations and ECMWF's Integrated Forecast System. The dataset provides atmospheric meteorological parameters.

■ Atmospheric Models from Météo-France

Global and high-resolution regional atmospheric models from Météo-France.



Besoins utilisateurs vs outils « analyse / traitement de données »

Opérationnalisation des EBVs

Mise à disposition workflows de production de métriques de biodiversité

Accès banques de données internationales (occurrence, génétiques, satellites, climat..)

Lien recherche - expertise

Formation

Visualisation de données

Analyse scientifique de données

Data and Metadata Management
These tutorials are focusing on data and metadata management in Ecology.

Lesson

Creating FAIR Quality assessment reports and draft of Data Papers from EML metadata with MetaShRIMPS

Metadata EML FAIR Data Paper

Data preprocessing
These lessons focus on manners to preprocess data used in Ecology

Lesson

Biodiversity data exploration
taxonomic data data quality

Checking expected species and contamination in bacterial isolate
illumina bacteria microgalaxy

Cleaning GBIF data for the use in Ecology
gbif data management data cleaning

Data access
These lessons focus on ways to access data classically used in Ecology

Lesson

QGIS Web Feature Services
earth-system GIS Geographical Information System WFS Spatial data Maps OGC

Data analysis
These lessons focus on ways to analyse data in Ecology

Lesson

Champs blocs indicators

Compute and analyze biodiversity metrics with PAMPA toolsuite
Species population EBV class Community composition EBV class EBV dataset EBV workflow modeling

From NDVI data with OpenEO to time series visualisation with Holoviews
earth-system land degradation NDVI copernicus holoviews

Metabarcoding/eDNA through Obitoools
Genetic composition EBV class Community composition EBV class EBV dataset EBV workflow eDNA
Metabarcoding

Obis marine indicators
earth-system ocean marine omics biodiversity

Preparing genomic data for phylogeny reconstruction
phylogeny data handling functional annotation

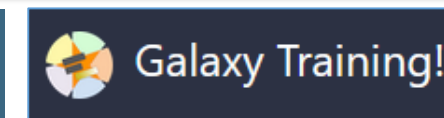
RAD-Seq Reference-based data analysis
RAD-seq Genetic composition EBV class Species population EBV class EBV dataset EBV workflow

RAD-Seq de-novo data analysis
RAD-seq Genetic composition EBV class Species populations EBV class EBV dataset EBV workflow

RAD-Seq to construct genetic maps
RAD-seq Genetic composition EBV class EBV dataset EBV workflow

Regional GAM
Species populations EBV class Species traits EBV class EBV dataset EBV workflow

Sentinel 2 biodiversity
Remote sensing



<https://training.galaxyproject.org/training-material/topics/ecology/>

Data visualization
These tutorials showcase data visualization in Ecology

Lesson

Visualization of Climate Data using NetCDF xarray Map Plotting
pangeo

Visualize EBV cube data with Panoply netCDF viewer
interactive-tools EBV cube Data visualization

Besoins utilisateurs vs outils « analyse / traitement de données »

Opérationnalisation des EBVs

Mise à disposition workflows de production de métriques de biodiversité

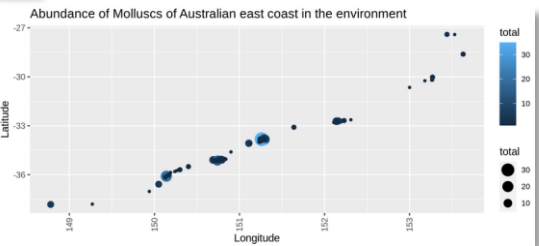
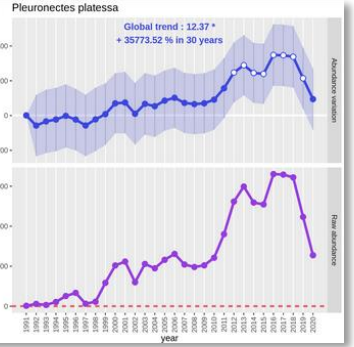
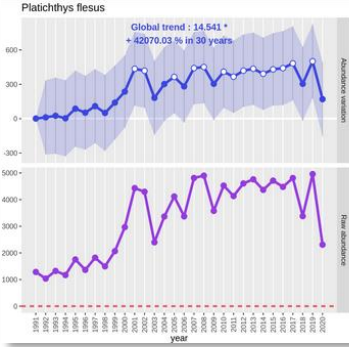
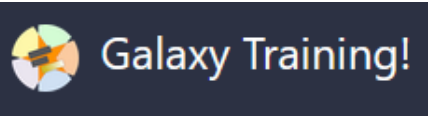
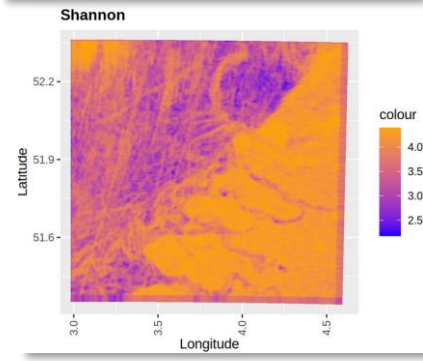
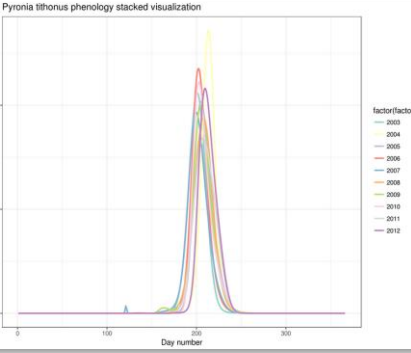
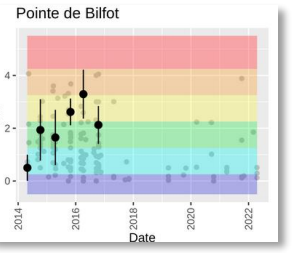
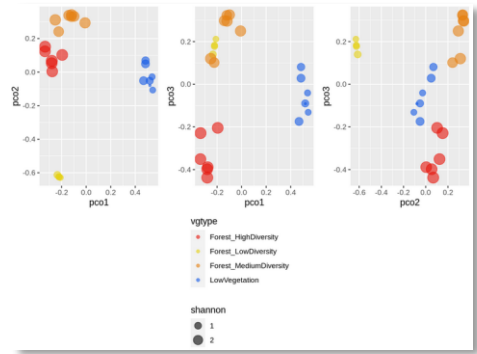
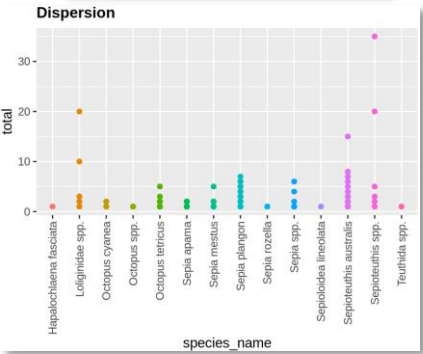
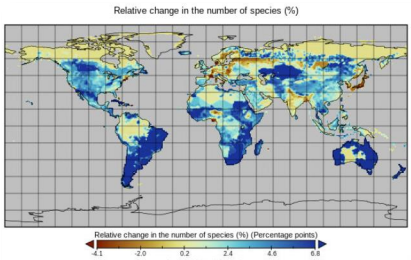
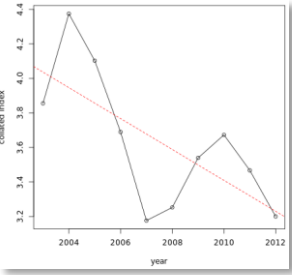
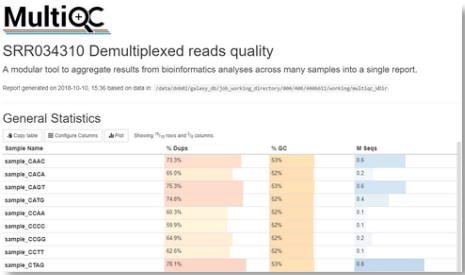
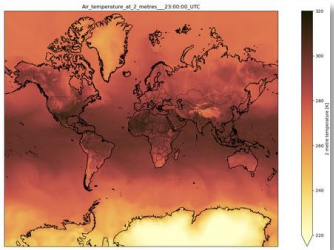
Accès banques de données internationales (occurrence, génétiques, satellites, climat..)

Lien recherche - expertise

Formation

Visualisation de données

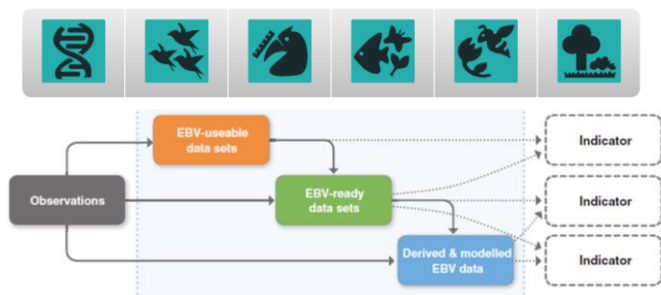
Analyse scientifique de données



MOTTO

Rely on what already exists and has proven efficiency

Concepts and principles



Adapt and apply
to help identify gaps and
bias

Data



Use and reuse
to exploit full
potential

Analytical processes



Atomize and generalize
to get plasticity and
reproducibility

Technical solutions
& related networks



Galaxy
PROJECT



DataONE

Share and contribute
to enhance communication
and networks

Global Open Science Cloud EBVOSC case study

Build National BONs → DigitalBON → GBiOS