0% song guaranted

Pôle National de Données de Biodiversité

Galaxy-E: Ecological data analysis, citizen science and biodiversity indicators production!

@ColineRoyaux #PNDB @Yvan2935

Coline Royaux Yvan Le Bras

yvan.le-bras@mnhn.fr

eosc | FAIR-EASE

IA-BIODIV CHALLENGE

EOSC-Pillar

GO FAIR BiodiFAIRse

GEO BON Group on Earth Observations Biodiversity Observation Network French BON

Gallantries

Erasmus+

Context – We need Atomization

Currently, in ecology ...

One R script for one input datafile

```
                ,direction= wide )
    tab[is.na(tab)] <- 0
                                        #  filename <- "touverUnNom"
                                        #  chemin <- paste(rep,filename,sep="/")
                                        #  write.table(tab, chemin)
  colnames(tab) <- sub("nombre.","",colnames(tab))

    return(tab)
}

## sous jeux de donnees si choix d espece d annee ou d un pourcentage de carres
makeSousTab <- function(tab,vecSp=NULL,echantillon=1,
                        methodeEchantillon="carre",vecannees=NULL) {
    cat("  -- Fabrication du sous jeu de donnees --\n")
    flush.console()
   ## reduction de la table à certaine espèces
   if(!is.null(vecSp)) {
        cat("     selection",length(vecSp),"espece(s):\n -> ")
        cat(vecSp)
        cat("\n")
        tab <- data.frame(carre = tab$carre,annee = tab$annee,tab[,vecSp])
        colnames(tab) <- c("carre","annee",vecSp)
    }
  ## reduction de la table pour certaines annees
    if(!is.null(vecannees)) {
        tab <- subset(tab,annee>=vecannees[1] & annee <= vecannees[2])

    }
  ## reduction de la table par une proportion de carre suivie
    if(echantillon != 1) {
        if(echantillon < 1 & echantillon >0) {
            nbinit <- nrow(tab)
            if(methodeEchantillon == "global") {
                nb <- round(nrow(tab)*echantillon)
                cat("     echantillonage",echantillon*100,
                    "% des donnees par la methode",methodeEchantillon,"\n")
                cat(" -> conservation de",nb,"lignes sur",nbinit,"\n")
                flush.console()
                tab <- tab[sample(1:nrow(tab))[1:nb],]
            } else {
                if (methodeEchantillon =="carre") {
                    cat("     echantillonage",echantillon*100,
                        "% des carrees par la methode",methodeEchantillon,"\n")
                    nbcarreinit <- length(unique(tab$carre))
                    chat=sample(unique(tab$carre),
                        length(unique(tab$carre))*echantillon,replace=F)
                    cat(" -> conservation de",length(chat),"carrees sur",
                        nbcarreinit)
                    tab=subset(tab, subset = carre %in% chat)
                    cat(" (",nrow(tab)," lignes sur ",nbinit,")\n",sep="")
                } else {

                    stop("Methode d echantillonnage non reconnue")
```

Context – We need Atomization

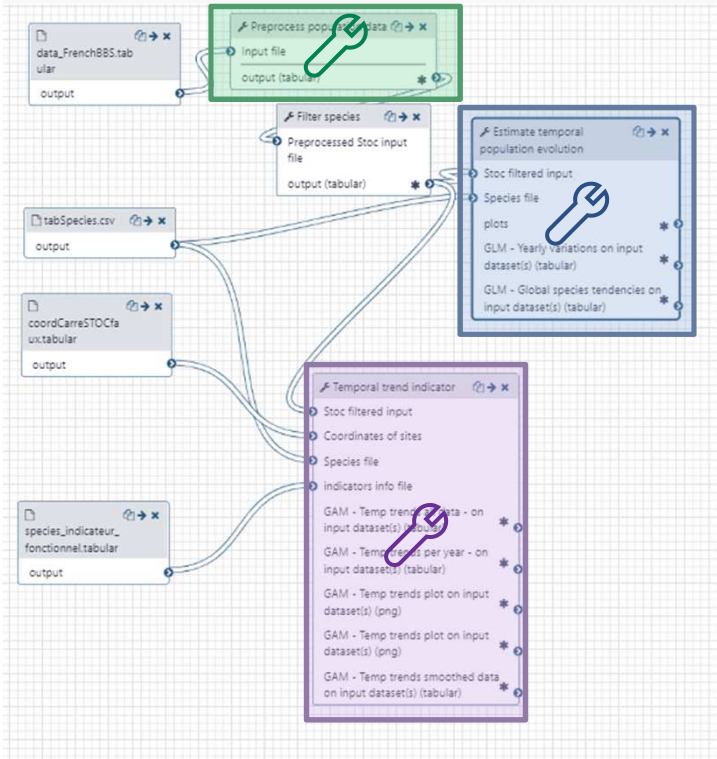
Currently, in ecology ...

One R script for one input datafile

```
                ,direction= wide )
  tab[is.na(tab)] <- 0
                                  # filename <- "touverUnNom"
                                  # chemin <- paste(rep,filename,sep="/")
                                  # write.table(tab, chemin)
colnames(tab) <- sub("nombre.","",colnames(tab))

  return(tab)
}

## sous jeux de donnees si choix d espece d annee ou d un pourcentage de carres
makeSousTab <- function(tab,vecSp=NULL,echantillon=1,
                        methodeEchantillon="carre",vecannees=NULL) {
  cat("  -- Fabrication du sous jeu de donnees --\n")
  flush.console()
 ## reduction de la table à certaine espèces
 if(!is.null(vecSp)) {
      cat("     selection",length(vecSp),"espece(s):\n -> ")
      cat(vecSp)
      cat("\n")
      tab <- data.frame(carre = tab$carre,annee = tab$annee,tab[,vecSp])
      colnames(tab) <- c("carre","annee",vecSp)
  }
## reduction de la table pour certaines annees
  if(!is.null(vecannees)) {
      tab <- subset(tab,annee>=vecannees[1] & annee <= vecannees[2])

  }
## reduction de la table par une proportion de carre suivie
  if(echantillon != 1) {
      if(echantillon < 1 & echantillon >0) {
          nbinit <- nrow(tab)
          if(methodeEchantillon == "global") {
              nb <- round(nrow(tab)*echantillon)
              cat("     echantillonage",echantillon*100,
                  "% des donnees par la methode",methodeEchantillon,"\n")
              cat(" -> conservation de",nb,"lignes sur",nbinit,"\n")
              flush.console()
              tab <- tab[sample(1:nrow(tab))[1:nb],]
          } else {
              if (methodeEchantillon =="carre") {
                  cat("     echantillonage",echantillon*100,
                      "% des carrees par la methode",methodeEchantillon,"\n")
                  nbcarreinit <- length(unique(tab$carre))
                  chat=sample(unique(tab$carre),
                      length(unique(tab$carre))*echantillon,replace=F)
                  cat(" -> conservation de",length(chat),"carrees sur",
                      nbcarreinit)
                  tab=subset(tab, subset = carre %in% chat)
                  cat(" (",nrow(tab)," lignes sur ",nbinit,")\n",sep="")
              } else {

                  stop("Methode d echantillonnage non reconnue")
```

With Galaxy...

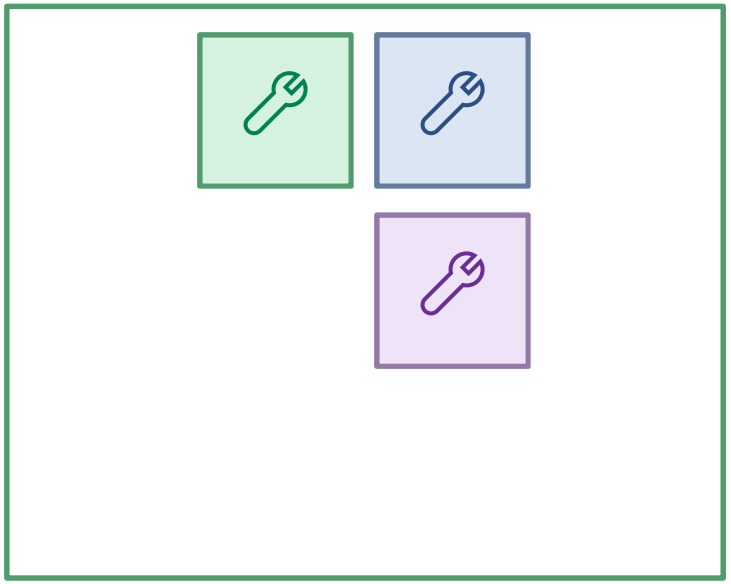
Several atomized R scripts for several input datafiles

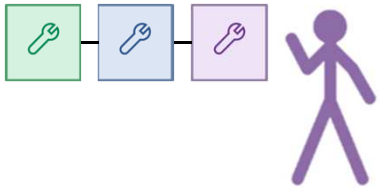


Context – We need Sharing & Generalization

	participation	Nuit	num_micro	groupe	espece	nb_contacts
1	55de2cd52121b1000d27430e	2015-07-26	0	bat	Barbar	1
2	55de2cd52121b1000d27430e	2015-07-26	0	bush-cricket	Barfis	1
3	55de2cd52121b1000d27430e	2015-07-26	0	noise	noise	5022
4	55de2cd52121b1000d27430e	2015-07-26	0	bush-cricket	Decalb	5
5	55de2cd52121b1000d27430e	2015-07-26	0	bush-cricket	Tyllil	18
6	55de2cd52121b1000d27430e	2015-07-26	0	bat	Nyclei	1
7	55de2cd52121b1000d27430e	2015-07-26	0	bush-cricket	Phanan	269

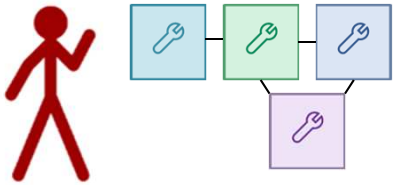
Toolshed





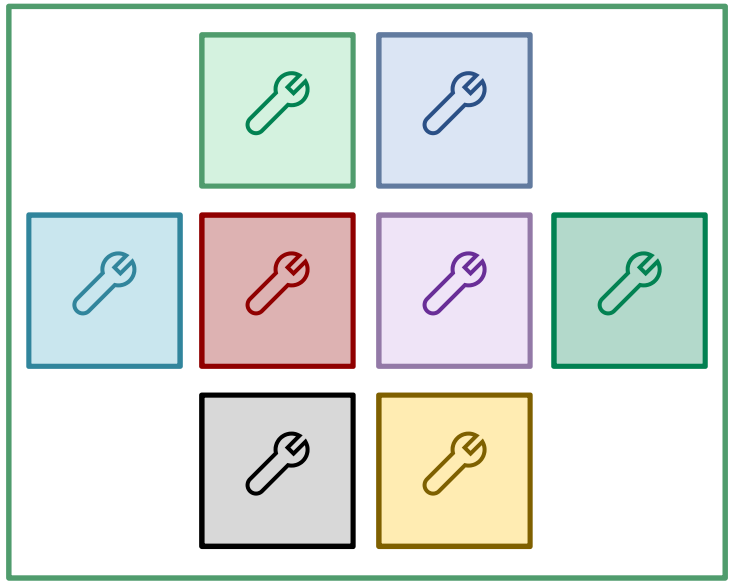
Context – We need Sharing & Generalization

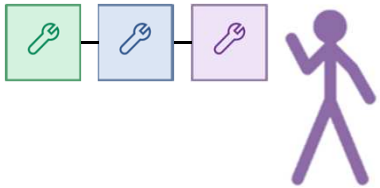
	carre	annee	espece	abond
1	2	2016	ACCGEN	0
2	2	2017	ACCGEN	0
3	2	2018	ACCGEN	0
4	2	2019	ACCGEN	0
5	183	2016	ACCGEN	0
6	183	2017	ACCGEN	0
7	183	2018	ACCGEN	0
8	183	2019	ACCGEN	0

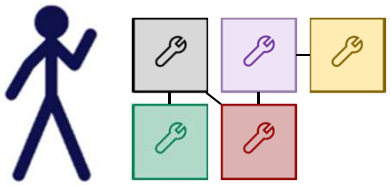


	participation	Nuit	num_micro	groupe	espece	nb_contacts
1	55de2cd52121b1000d27430e	2015-07-26	0	bat	Barbar	1
2	55de2cd52121b1000d27430e	2015-07-26	0	bush-cricket	Barfis	1
3	55de2cd52121b1000d27430e	2015-07-26	0	noise	noise	5022
4	55de2cd52121b1000d27430e	2015-07-26	0	bush-cricket	Decalb	5
5	55de2cd52121b1000d27430e	2015-07-26	0	bush-cricket	Tyllil	18
6	55de2cd52121b1000d27430e	2015-07-26	0	bat	Nyclei	1
7	55de2cd52121b1000d27430e	2015-07-26	0	bush-cricket	Phanan	269

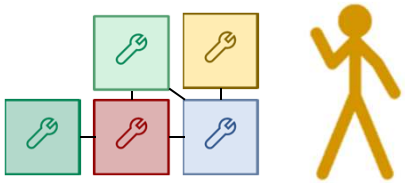
Toolshed







	Unit	obs	rotation	codeSp	sexe	taille	classe_taille	poids	nb_ind
1	AS140155		3	Hemifasc	-999	-999	P	-999	1
2	AS140159		1	Nasosp.	-999	-999	P	-999	3
3	AS140159		3	Gompvari	-999	-999	P	-999	1
4	AS140160		3	Gompvari	-999	-999	P	-999	1
5	AS140099		2	Parumult	-999	-999	P	-999	1
6	AS140088		1	Varilout	-999	-999	P	-999	1
7	AS140088		2	Gompvari	-999	-999	P	-999	2
8	AS140041		1	Nasosp.	-999	-999	P	-999	5



	Survey	Year	Quarter	Area	AphiaID	Species	LngtClass	CPUE_number_per_hour
1	BITS	1991	1	22	126281	Anguilla anguilla	0	0.000000
2	BITS	1991	1	22	126281	Anguilla anguilla	720	0.009160
3	BITS	1991	1	22	126417	Clupea harengus	0	0.000000
4	BITS	1991	1	22	126417	Clupea harengus	80	0.075785
5	BITS	1991	1	22	126417	Clupea harengus	85	0.088277
6	BITS	1991	1	22	126417	Clupea harengus	95	0.037892
7	BITS	1991	1	22	126417	Clupea harengus	100	0.063293
8	BITS	1991	1	22	126417	Clupea harengus	105	0.012492
9	BITS	1991	1	22	126417	Clupea harengus	110	0.618357

And some trainings! => https://training.galaxyproject.org/

Galaxy Training! Contributors Languages Help Extras Search Tutorials

Climate

Learn to analyze climate data through Galaxy.

You can view the tutorial materials in different languages by clicking the dropdown icon next to the slides ( ) and tutorial ( ) buttons below.

Requirements

Before diving into this topic, we recommend you to have a look at:

- Introduction to Galaxy Analyses

Material

Search

Lesson	Slides	Hands-on	Recordings	Input dataset	Workflows
Introduction to climate data					
Functionally Assembled Terrestrial Ecosystem Simulator (FATES) interactive-tools					
Functionally Assembled Terrestrial Ecosystem Simulator (FATES) with Galaxy Climate JupyterLab interactive-tools					
Getting your hands-on climate data					
Pangeo ecosystem 101 for everyone - Introduction to Xarray Galaxy Tools pangeo					
Pangeo Notebook in Galaxy - Introduction to Xarray pangeo interactive-tools jupyter-notebook jupyter-notebook					
Visualize Climate data with Panoply netCDF viewer interactive-tools					

And some trainings! => https://training.galaxyproject.org/

Galaxy Training! Contributors Languages Help Extras Search Tutorials

Climate

Learn to analyze climate data through Galaxy.

You can view the tutorial materials in different languages by clicking the dropdown icon next t

Requirements

Before diving into this topic, we recommend you to have a look at:

- Introduction to Galaxy Analyses

Material

Lesson	Slic
Introduction to climate data	
Functionally Assembled Terrestrial Ecosystem Simulator (FATES) interactive-tools	
Functionally Assembled Terrestrial Ecosystem Simulator (FATES) with Galaxy Climate JupyterLab interactive-tools	
Getting your hands-on climate data	
Pangeo ecosystem 101 for everyone - Introduction to Xarray Galaxy Tools pangeo	
Pangeo Notebook in Galaxy - Introduction to Xarray pangeo interactive-tools jupyter-notebook jupyter-notebook	
Visualize Climate data with Panoply netCDF viewer interactive-tools	

Galaxy Training! Contributors Languages Help Extras Search Tutorials

Ecology

Learn to analyse Ecological data through Galaxy.

You can view the tutorial materials in different languages by clicking the dropdown icon next to the slides ( ) and tutorial ( ) buttons below.

Requirements

Before diving into this topic, we recommend you to have a look at:

- Introduction to Galaxy Analyses

Material

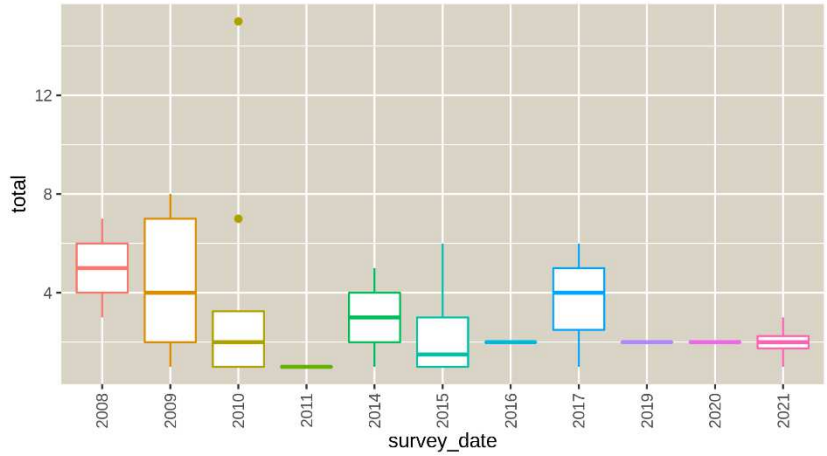
Search

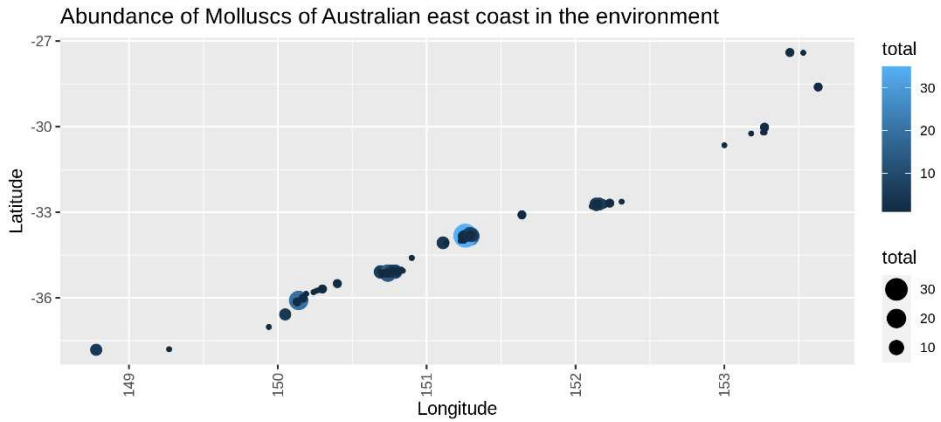
Lesson	Slides	Hands-on	Recordings	Input dataset	Workflows
Biodiversity data exploration					
Compute and analyze biodiversity metrics with PAMPA toolsuite					
Metabarcoding/eDNA through Obitools					
RAD-Seq de-novo data analysis RAD-seq					
RAD-Seq Reference-based data analysis RAD-seq					
RAD-Seq to construct genetic maps RAD-seq					
Regional GAM					
Species distribution modeling interactive-tools					

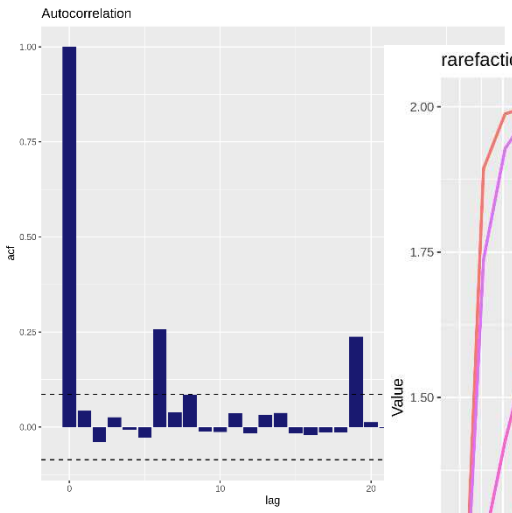
Galaxy-E Killer workflows

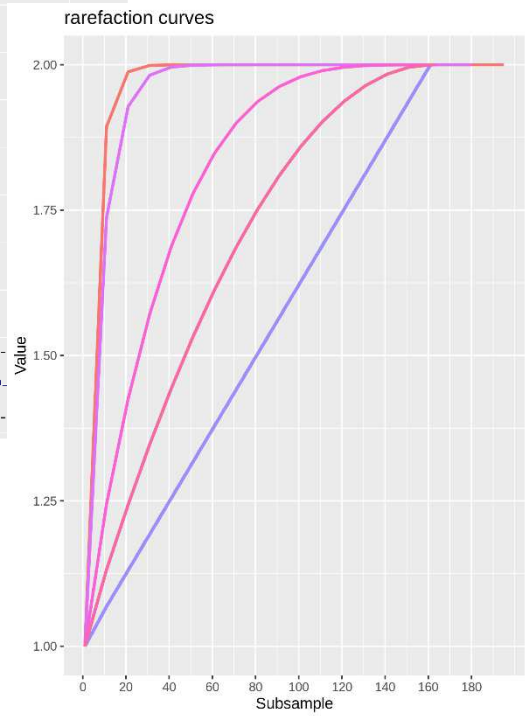
- Biodiversity exploration tools
- Biodiversity metrics & indicators production
- Dealing with GIS and netcdf files on Galaxy-E

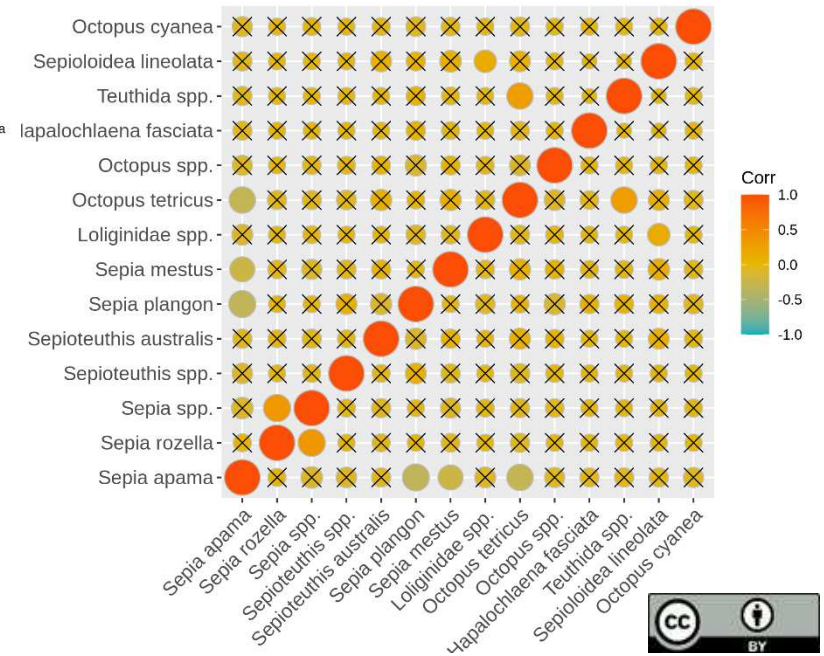
Biodiversity exploration tools



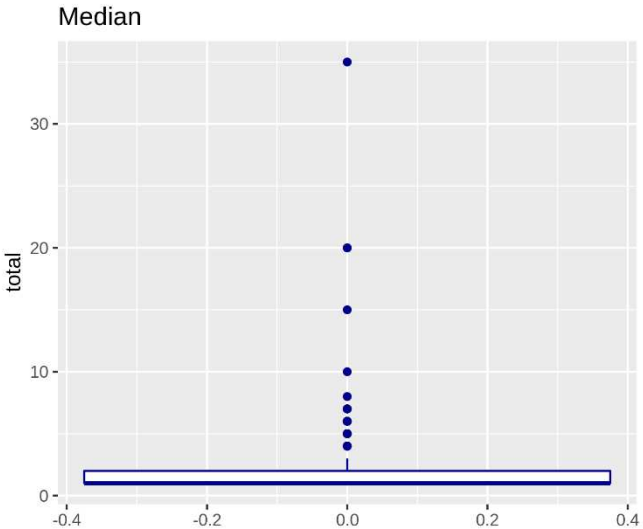


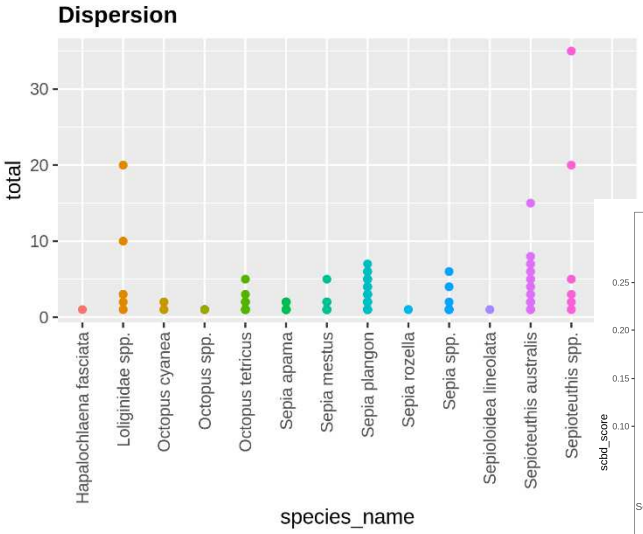


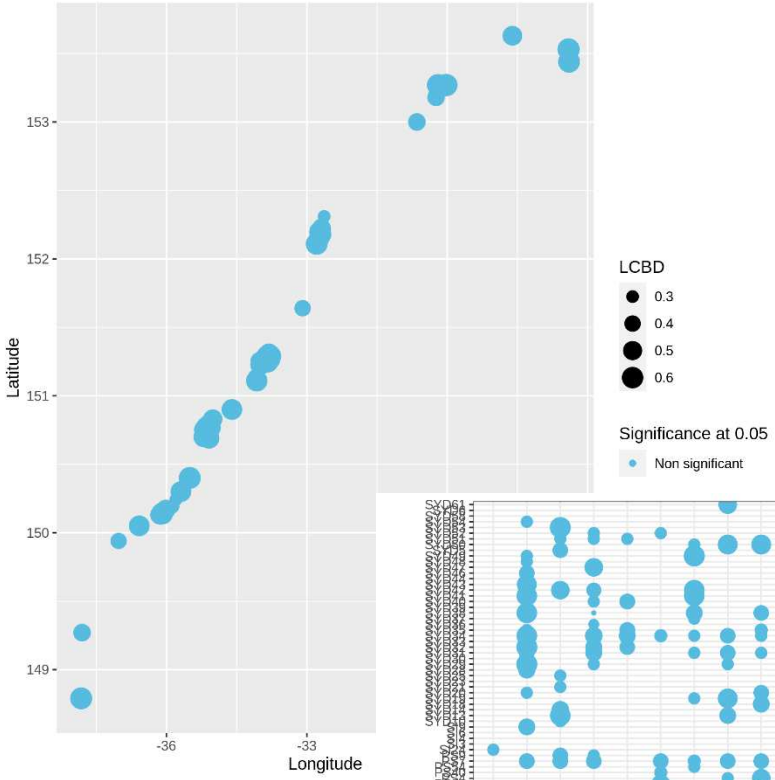


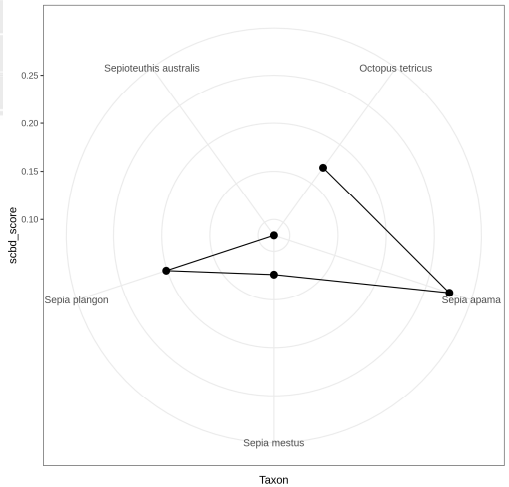


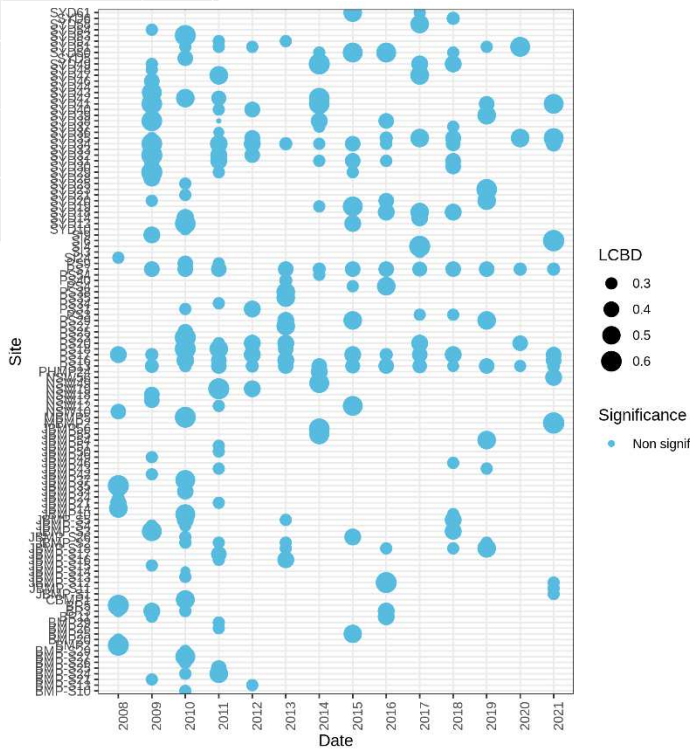
Biodiversity exploration tools









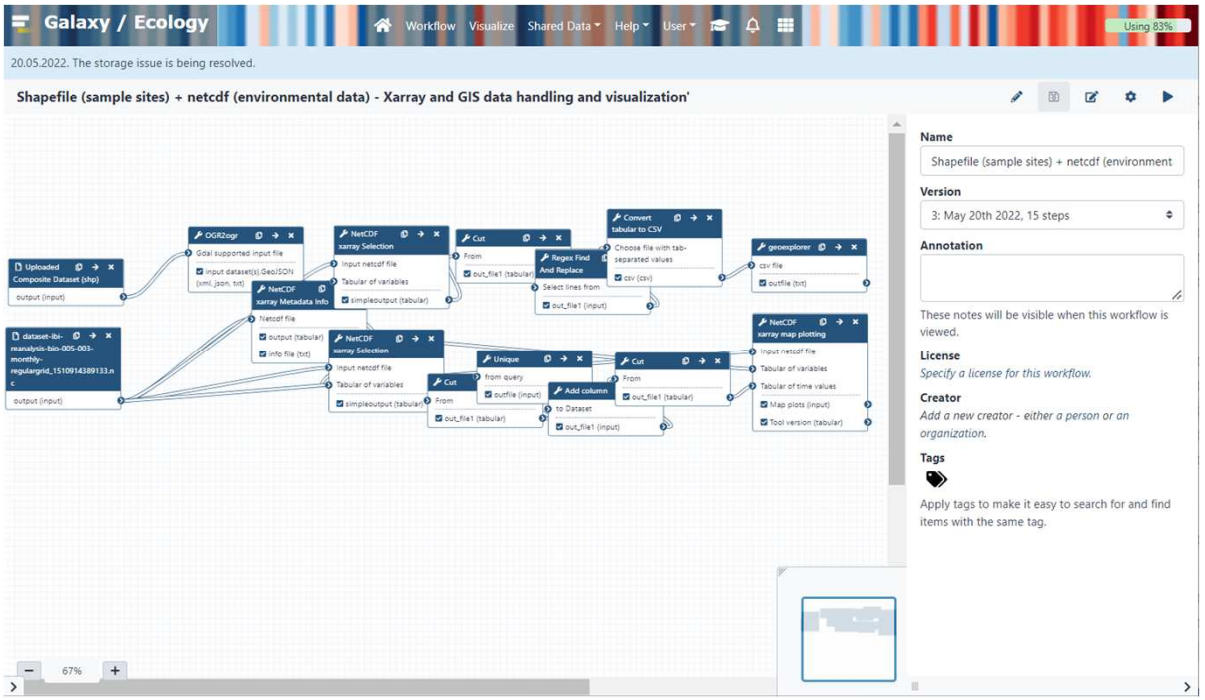


Dealing with GIS and netcdf files

Fouilloux *et al.* EGU22 **Pangeo for everyone with Galaxy**

A "Classical" data processing:
Sampling sites information in GIS data file
(often shapefile)
Environmental information in netCDF file

Create a file with environmental information on sampling sites!
Visualize maps of environmental parameters on sampling sites

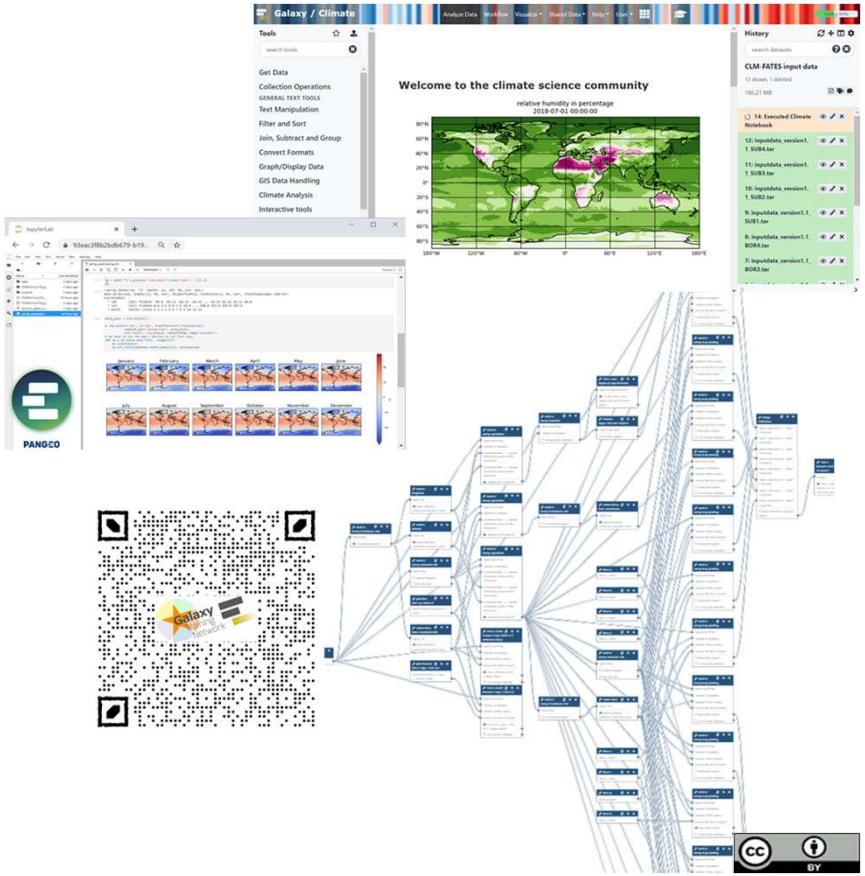


Until now: R + QGIS + a lot of manual manipulation
Now: a **Galaxy workflow** mixes scripts, GDAL & Xarray tools making it easily accessible and (re)-runnable.

Pangeo for everyone through Galaxy

Galaxy open-source platform for FAIR data analysis offers:

- ➔ Pangeo notebook deployment (local dask) **available to everyone (free registration);**
- ➔ Pangeo Galaxy Tools for fully **automated workflows**;
- ➔ **GUI** for users with **no programming skills**;
- ➔ Self-Paced **Learning material** and organisation of online training events with the Galaxy Training Network;
- ➔ **Training Infrastructure as a Service** is a free and ready to use with private queues where only training's jobs run.

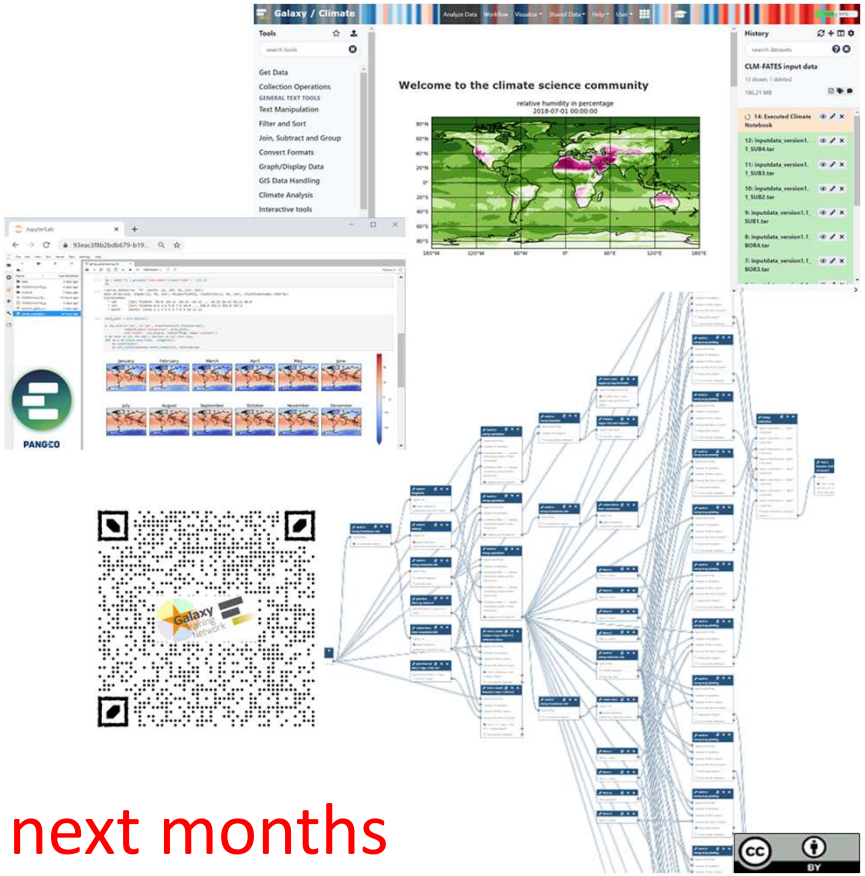


Slides from Fouilloux et al 2022 From EGU22 https://meetingorganizer.copernicus.org/EGU22/EGU22-5709.html?reply

Pangeo for everyone through Galaxy

Galaxy open-source platform for FAIR data analysis offers:

- ➔ Pangeo notebook deployment (local dask) **available to everyone (free registration);**
- ➔ Pangeo Galaxy Tools for fully **automated workflows**;
- ➔ **GUI** for users with **no programming skills**;
- ➔ Self-Paced **Learning material** and organisation of online training events with the Galaxy Training Network;
- ➔ **Training Infrastructure as a Service** is a free and ready to use with private queues where only training's jobs run.

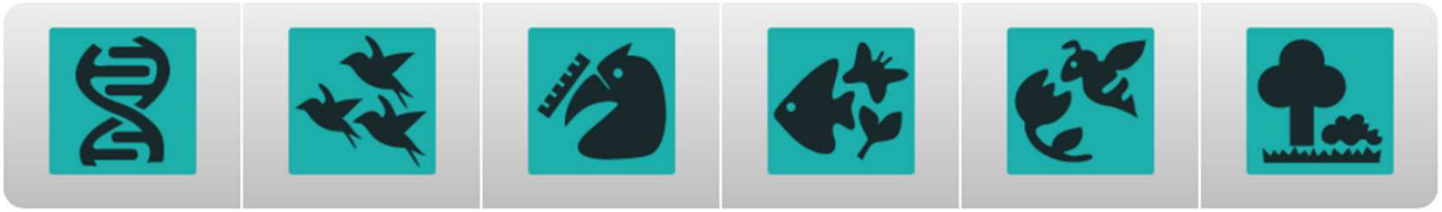


Amazing basis for eosc | FAIR-EASE

PNDB / Galaxy-E work for next months

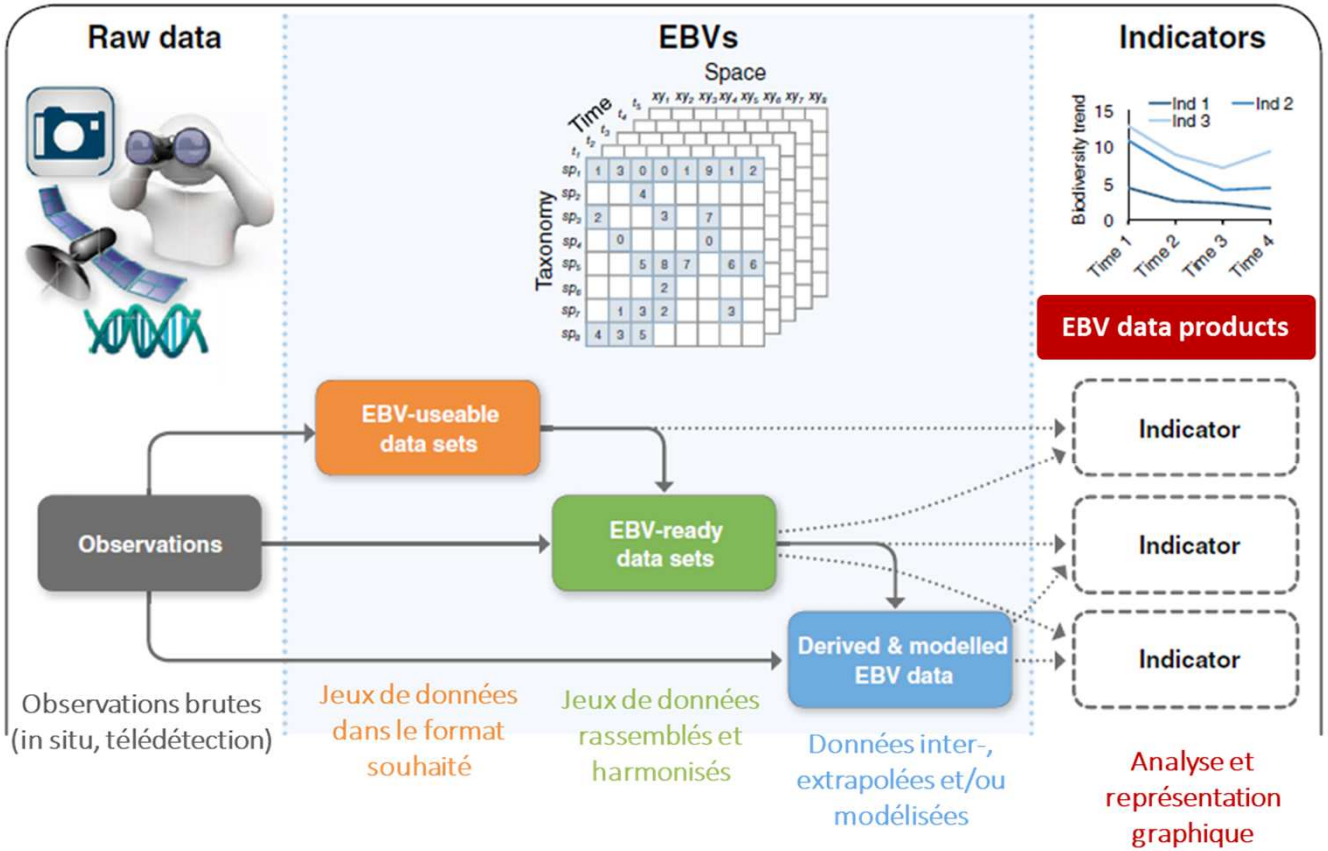
Slides from Fouilloux et al 2022 From EGU22 https://meetingorganizer.copernicus.org/EGU22/EGU22-5709.html?reply

Essential Biodiversity Variables workflows



Composition génétique | Espèces - Populations | Espèces - Traits | Composition de la communauté | Fonction de l'écosystème | Structure de l'écosystème

Kissling *et al.* 2017



EBV workflows: STOC

https://bit.ly/3AdFDFK

STOC
SUIVI TEMPOREL DES OISEAUX COMMUNS

CESCO Centre d'Écologie et des Sciences de la Conservation

UMS PATRIMOINE NATUREL

MUSÉUM NAL HIST NATURELLE

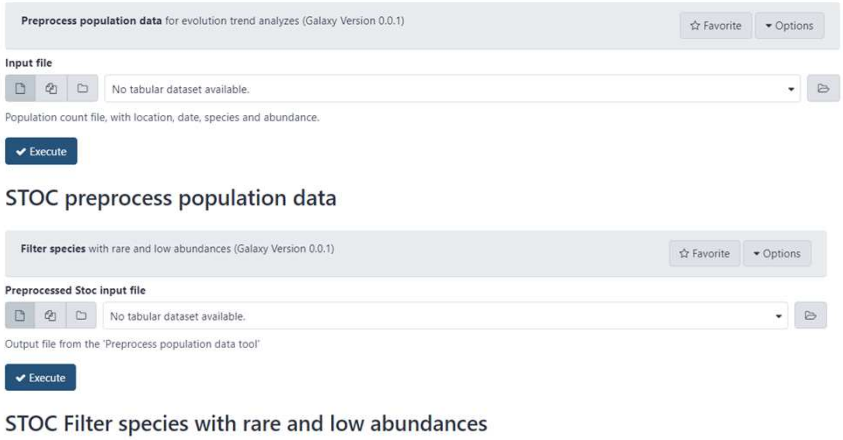
Biodiversity data

↓

1 Preprocess data

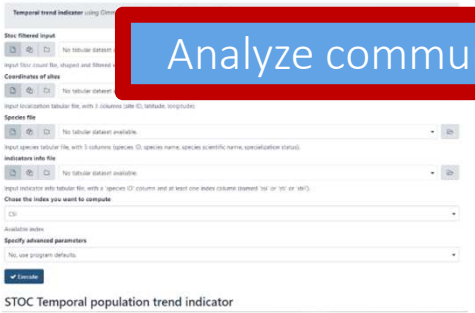
↓

2 Filter data



Community

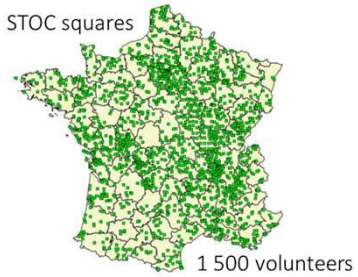
Species - population

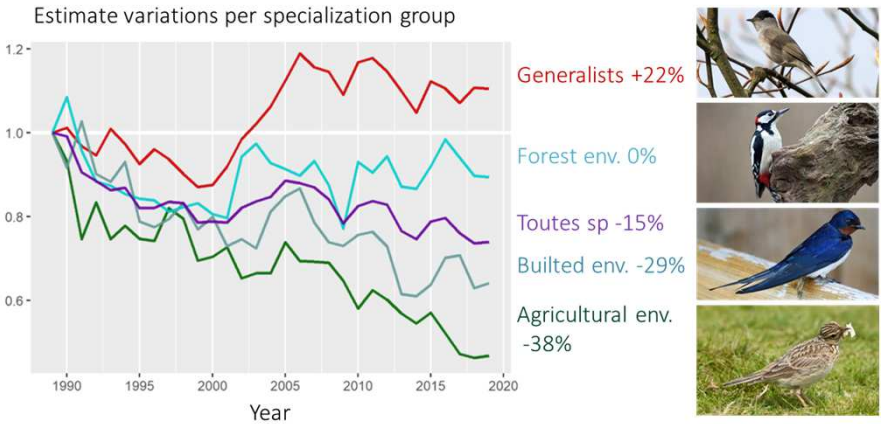


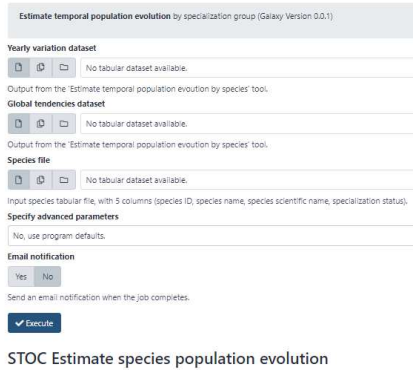
Analyze community indexes

3

Analyze species abundance





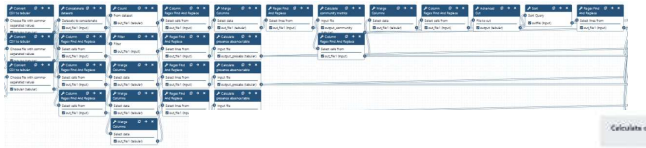


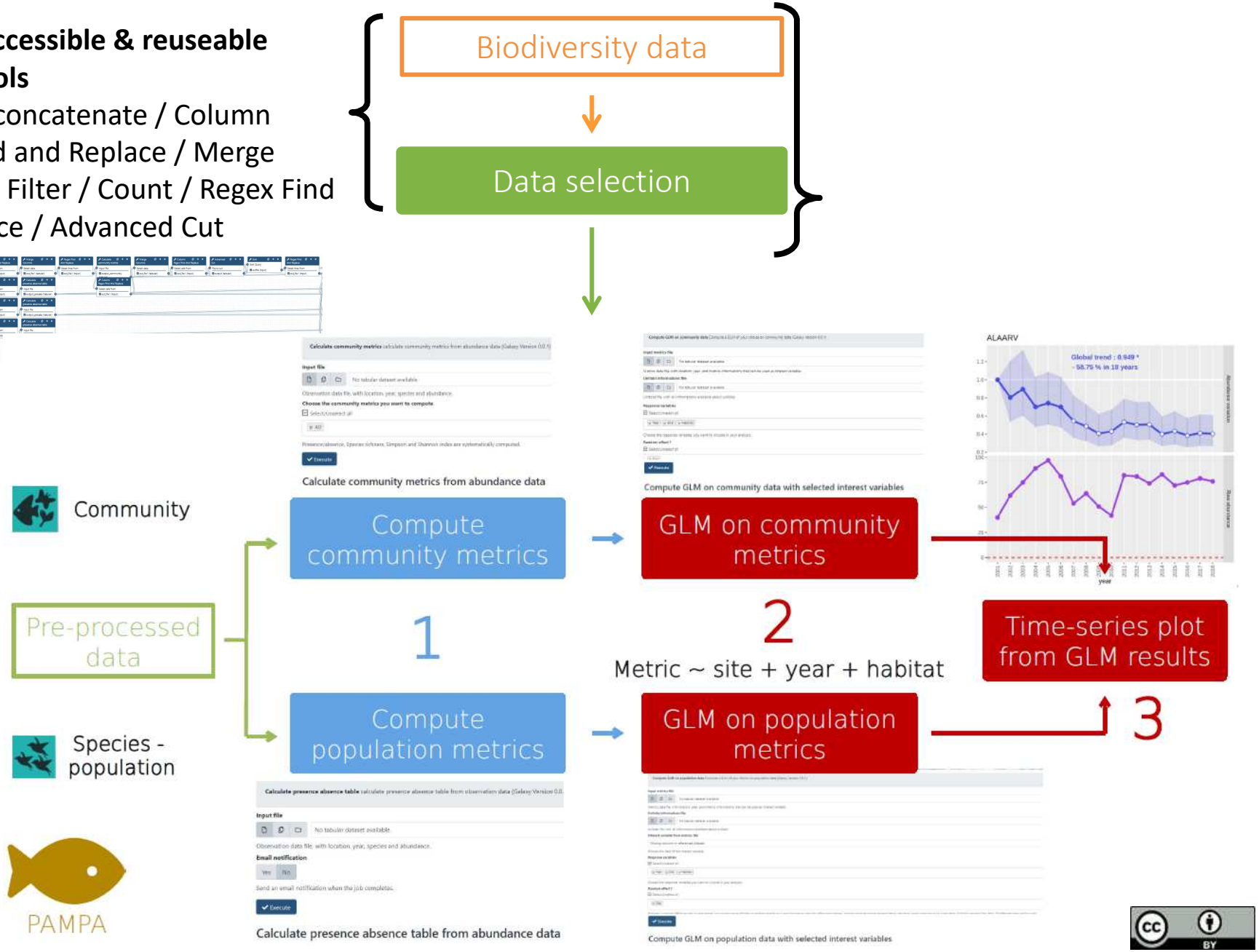
EBV workflows: PAMPA

wwz.ifremer.fr/pampa

Existing accessible & reuseable Galaxy tools

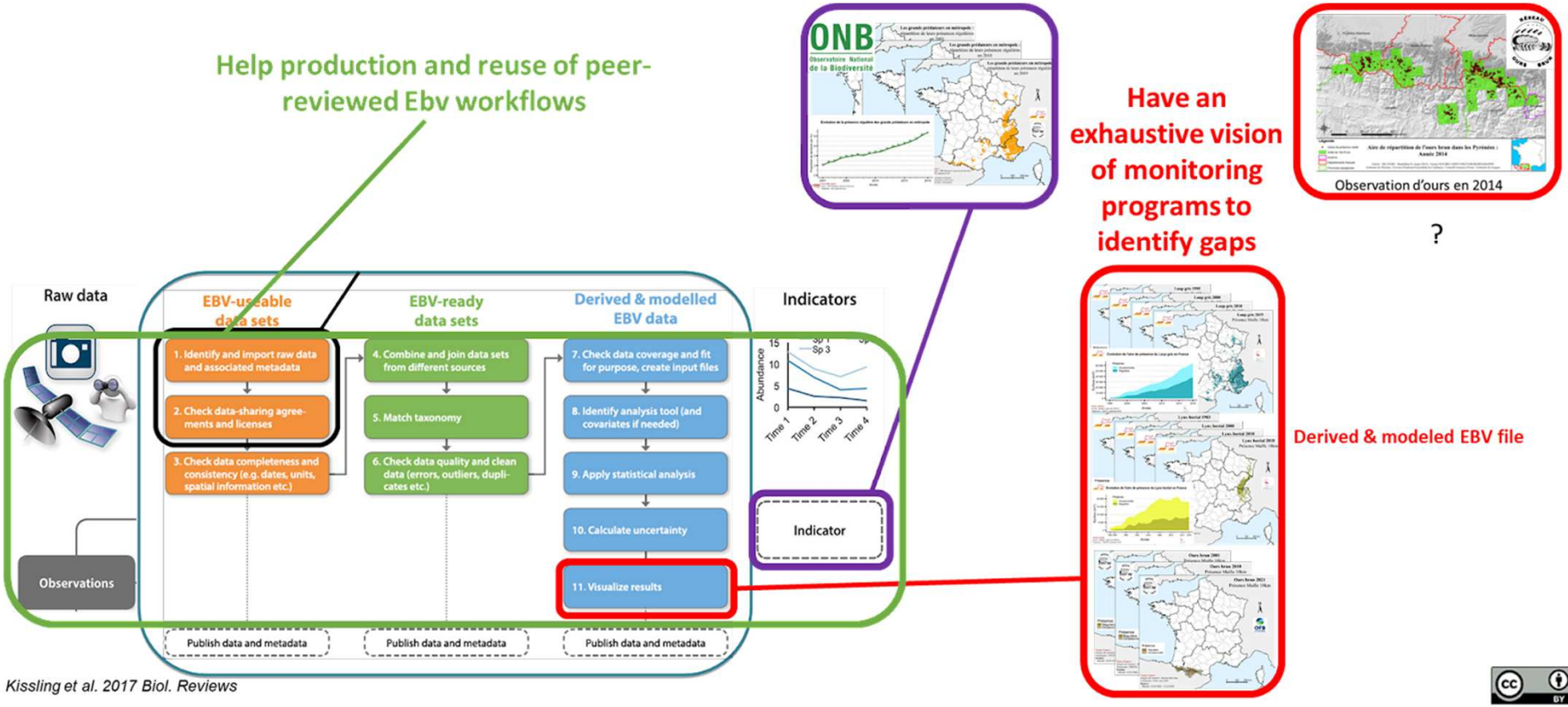
convert / concatenate / Column Regex Find and Replace / Merge Columns / Filter / Count / Regex Find and Replace / Advanced Cut





Essential Biodiversity Variables workflows

Help BONs to identify gaps & reuse EBV workflows



Kissling et al. 2017 Biol. Reviews

Essential Biodiversity Variables workflows

Global Open Science Cloud (GOSC)

Case Studies

Help BONs to identify gaps & reuse EBV workflows

Amazing basis for **EBVOSC**

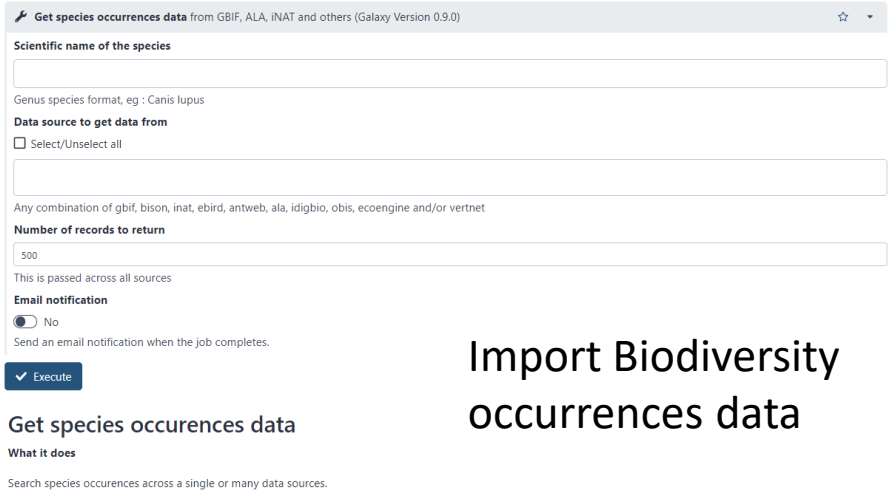
PNDB / Galaxy-E work for next months

Do you think this can help create a national biodiversity network on your country (Germany, Australian, ..) ? Contact us!

Not only for data analysis

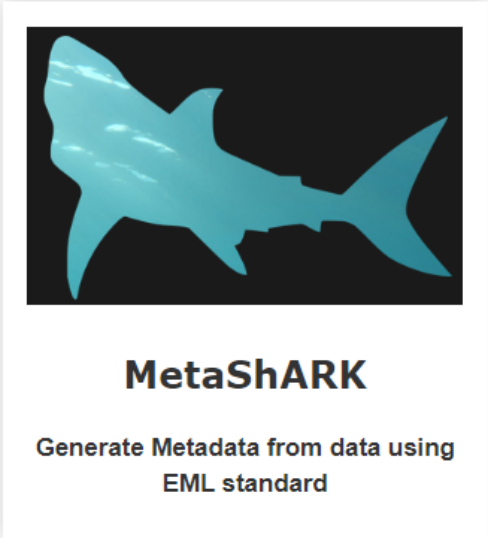
=> Also for Research data management

Ecological research data management



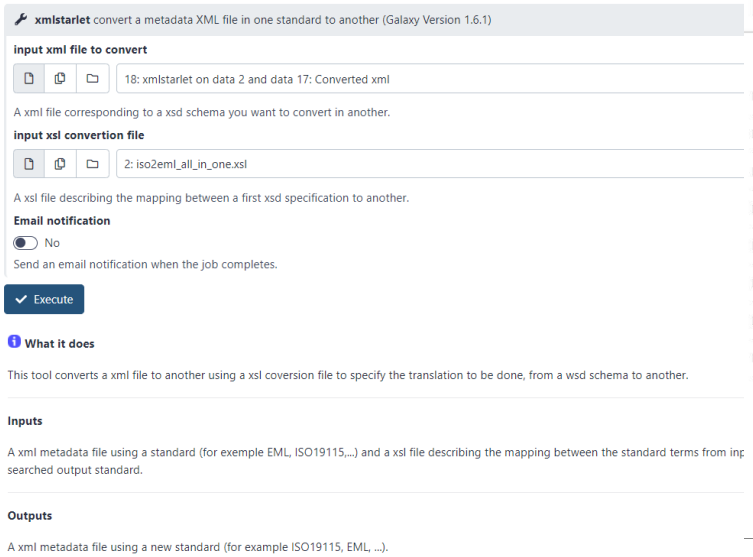
Import Biodiversity occurrences data

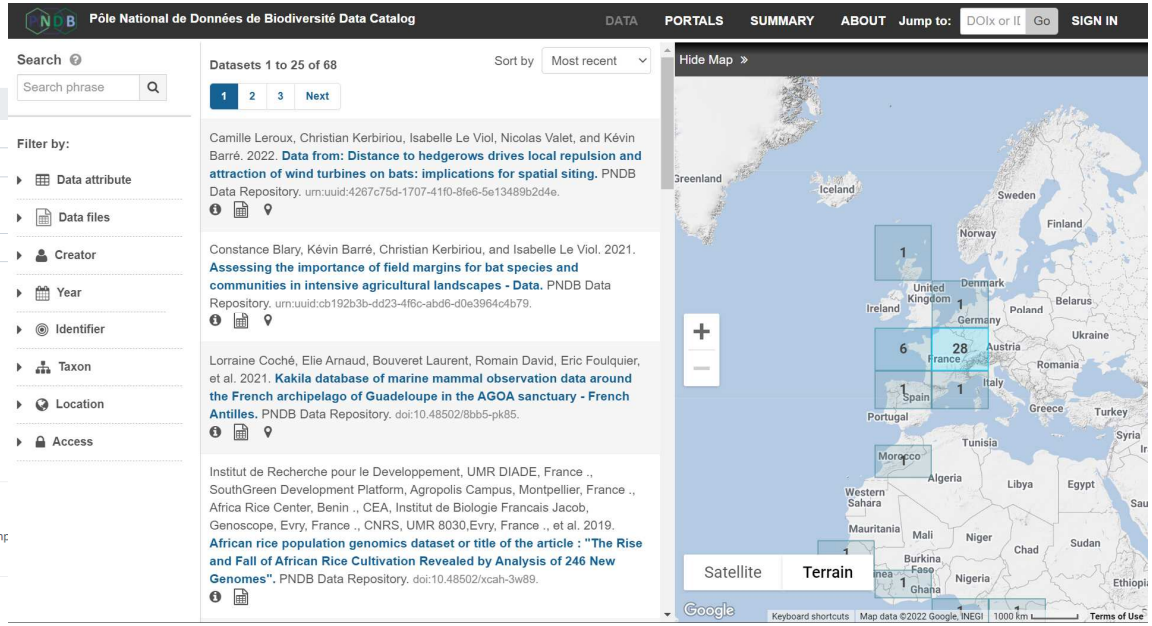
Generate metadata



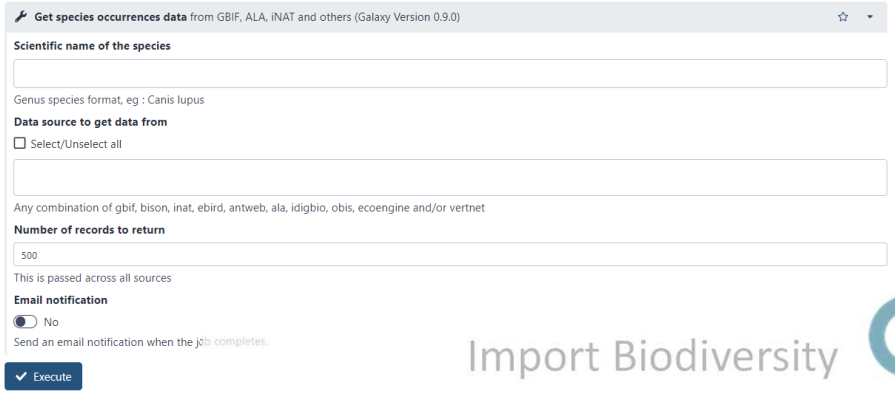
Upload data an metadata

Convert metadata





Ecological research data management



Import Biodiversity occurrences data

Generate metadata

Get species occurences data

What it does

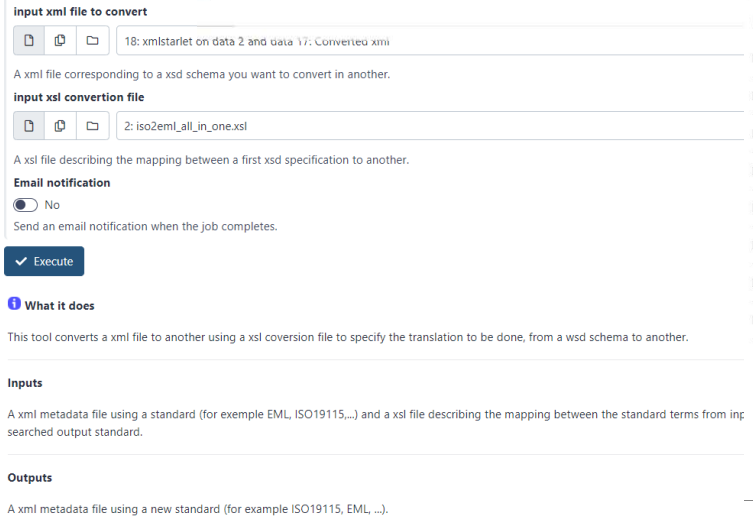
Search species occurences across a single or many data sources.

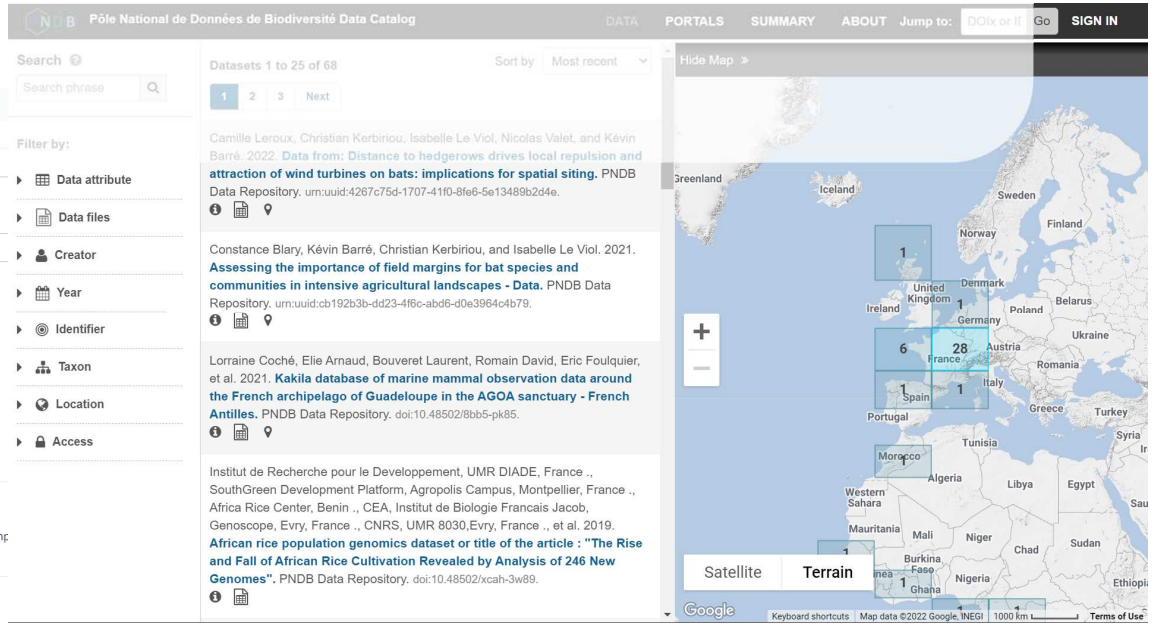
Amazing basis for

PNDB / Galaxy-E work for next months

Upload data an metadata

Convert metadata





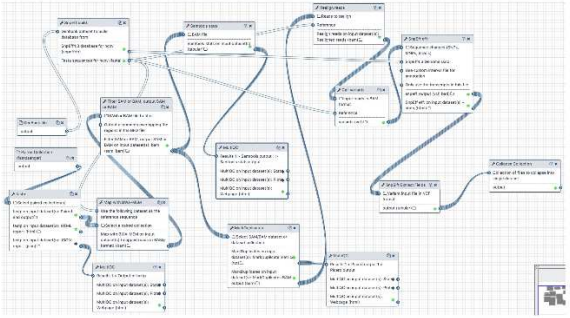
Not only asynchronously

=> Thanks to Galaxy interactive tools GxIT

One specific workflow goal

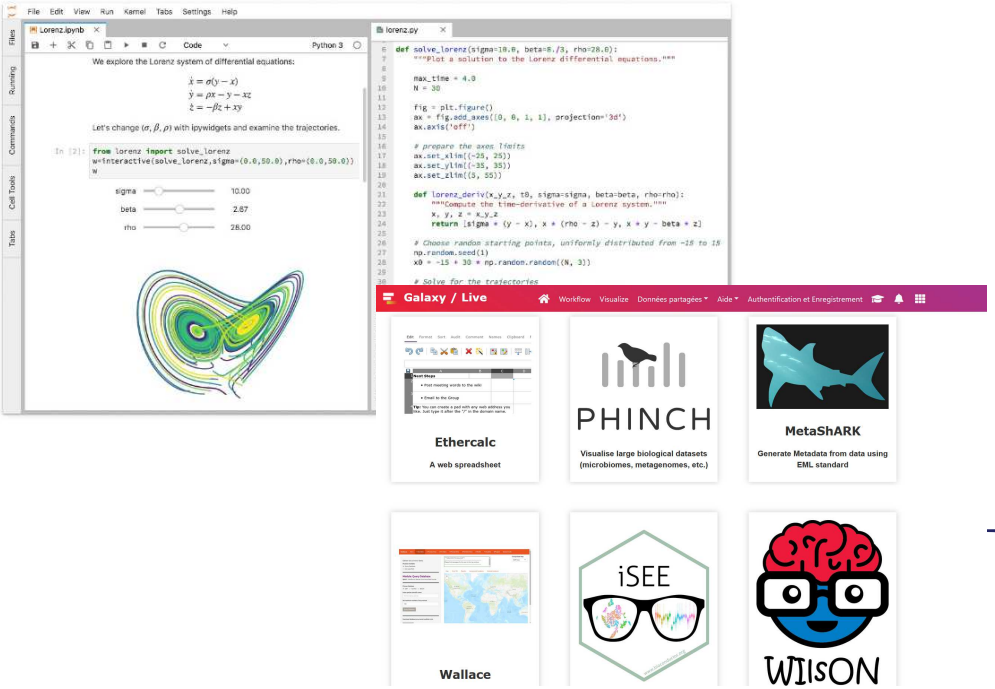
Galaxy as workflow engine

workflows (the async ones)



```
class: GalaxyWorkflow
doc: |
  Simple workflow that no-op cats a file
inputs:
  the_input:
      type: File
      doc: input doc
outputs:
  the_output:
      outputSource: cat/out_file1
steps:
  cat:
      tool_id: cat1
      doc: cat doc
      in:
      input1: the_input
```

workflows (the ~~a~~sync ones)



Interactive tools

Interactive Jupyter Notebook

GPU enabled Interactive Jupyter Notebook for Machine Learning

Interactive Climate Notebook

Interactive Pangeo Notebook

RStudio

Pylron Interactive Jupyter Notebook

HiGlass an interactive Hi-C data visualizer.

OpenRefine Working with messy data

Ubuntu XFCE Desktop

Panoply interative plotting tool for geo-referenced data

AskOmics a visual SPARQL query builder

Interactive CellXgene Environment

bam.iobio visualisation

VCF (iobio) Visualisation

Neo4j (Graph Database)

Phinch Visualisation

Paraview

Wilson Webbased Interactive Omics visualization

Wallace Webbased Interactive modeling of species niches and distributions

geoexplorer An interactive spatial analysis platform using ggvis and Leaflet

radiant Data analytics using Radiant R Shiny app

EtherCalc

VRM Editor interative tool for creating Variable Resolution Mesh for NorESM/CESM

SimText Interactive shiny app to explore SimText output data

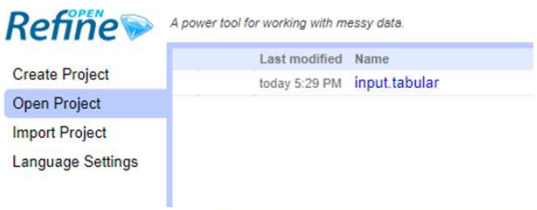
iSEE

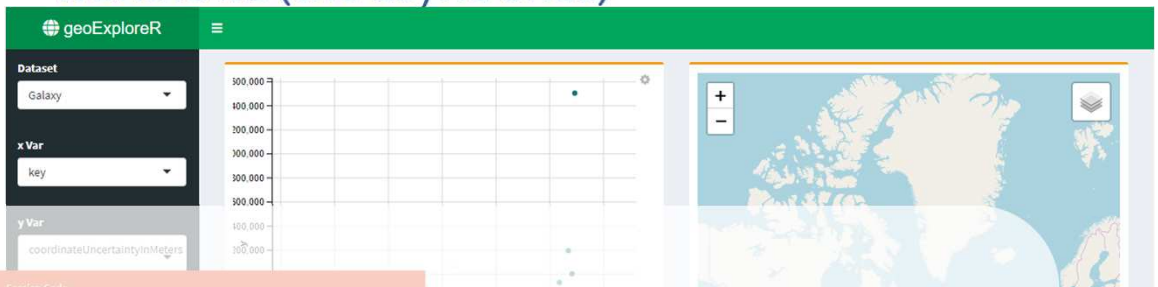
metashark Metadata Shiny Automated Resource and Knowledge

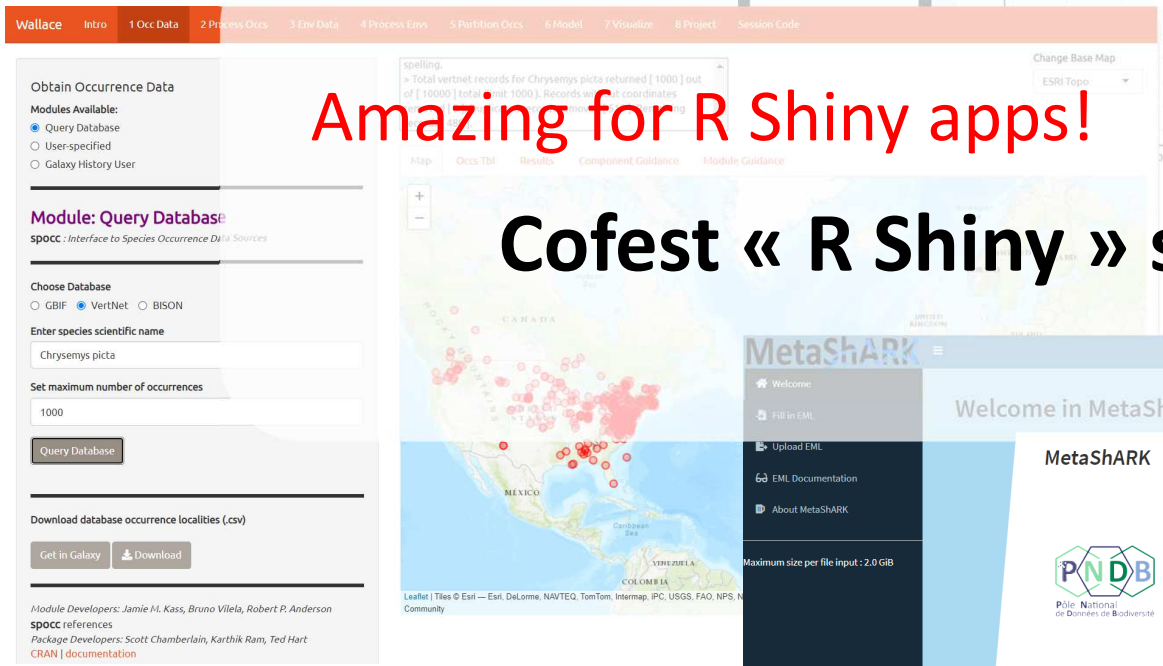
One specific workflow goal

Galaxy as workflow engine

workflows (the ~~a~~sync ones)

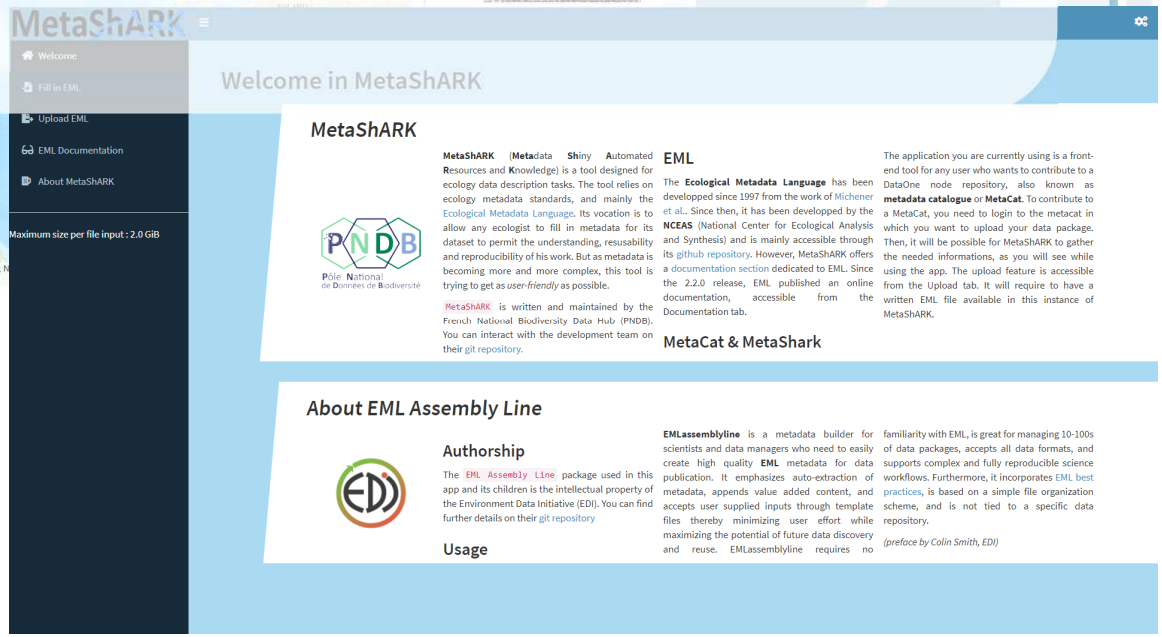






Amazing for R Shiny apps!

Cofest « R Shiny » subject?

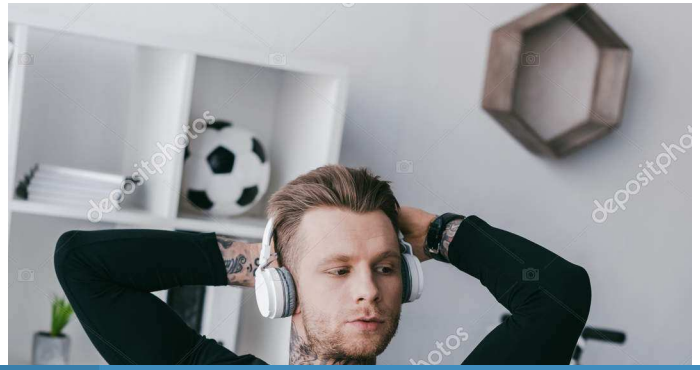


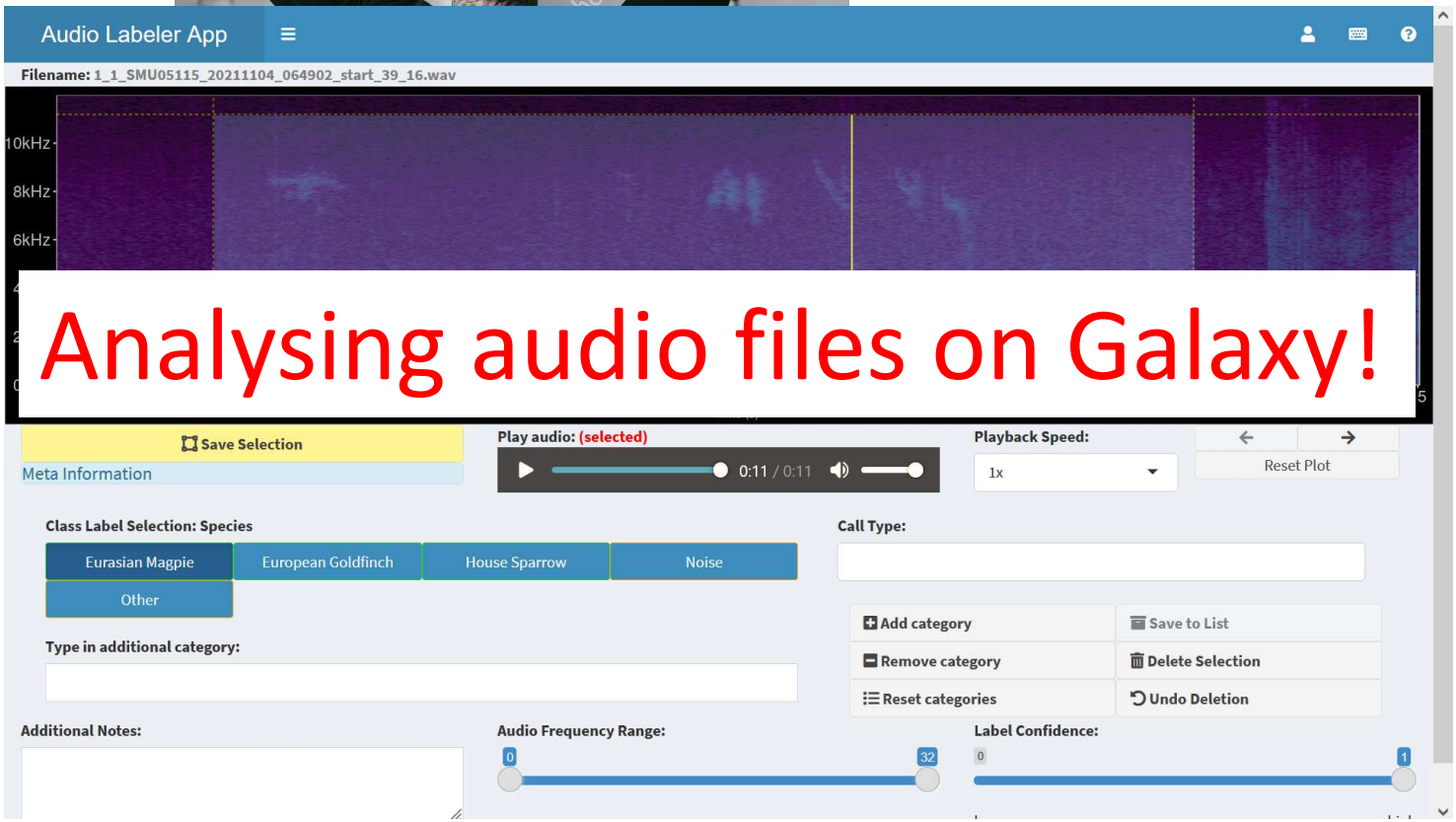
One specific workflow goal

Galaxy as workflow eng What is making this guy ?

One specific workflow goal

Galaxy as workflow eng **What is making this guy ?**





Analysing audio files on Galaxy!

Not only for scientists

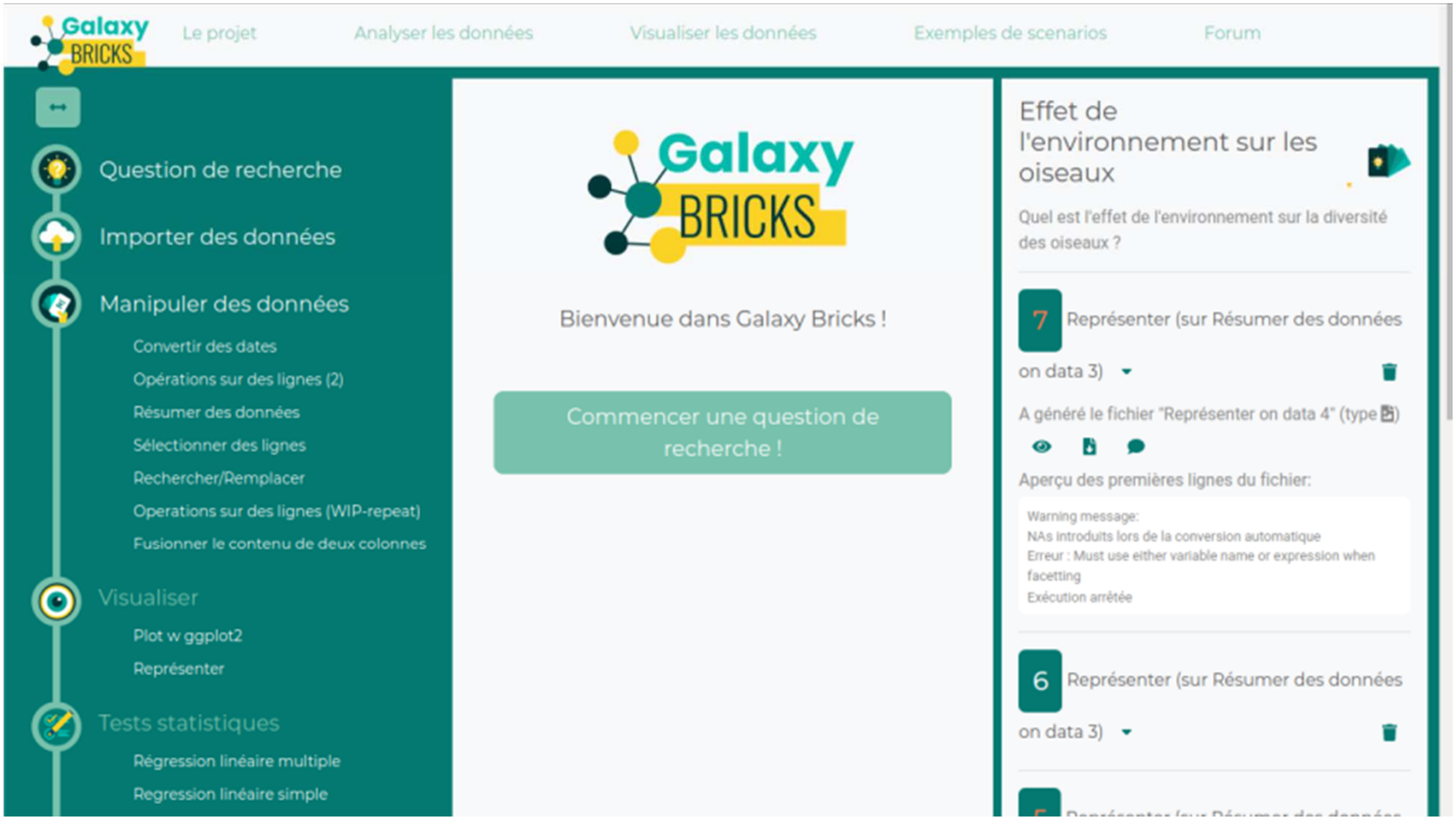
- Crowdsourcing through Galaxy webhooks
- Data / Biodiveristy literacy through Galaxy-Bricks

Galaxy-Bricks

VIGIENATURE École

Galaxy for pupils !

bricks.vigienature-ecole.fr



Galaxy-Bricks

VIGIENATURE École

Galaxy for pupils !

bricks.vigienature-ecole.fr

Amazing basis for

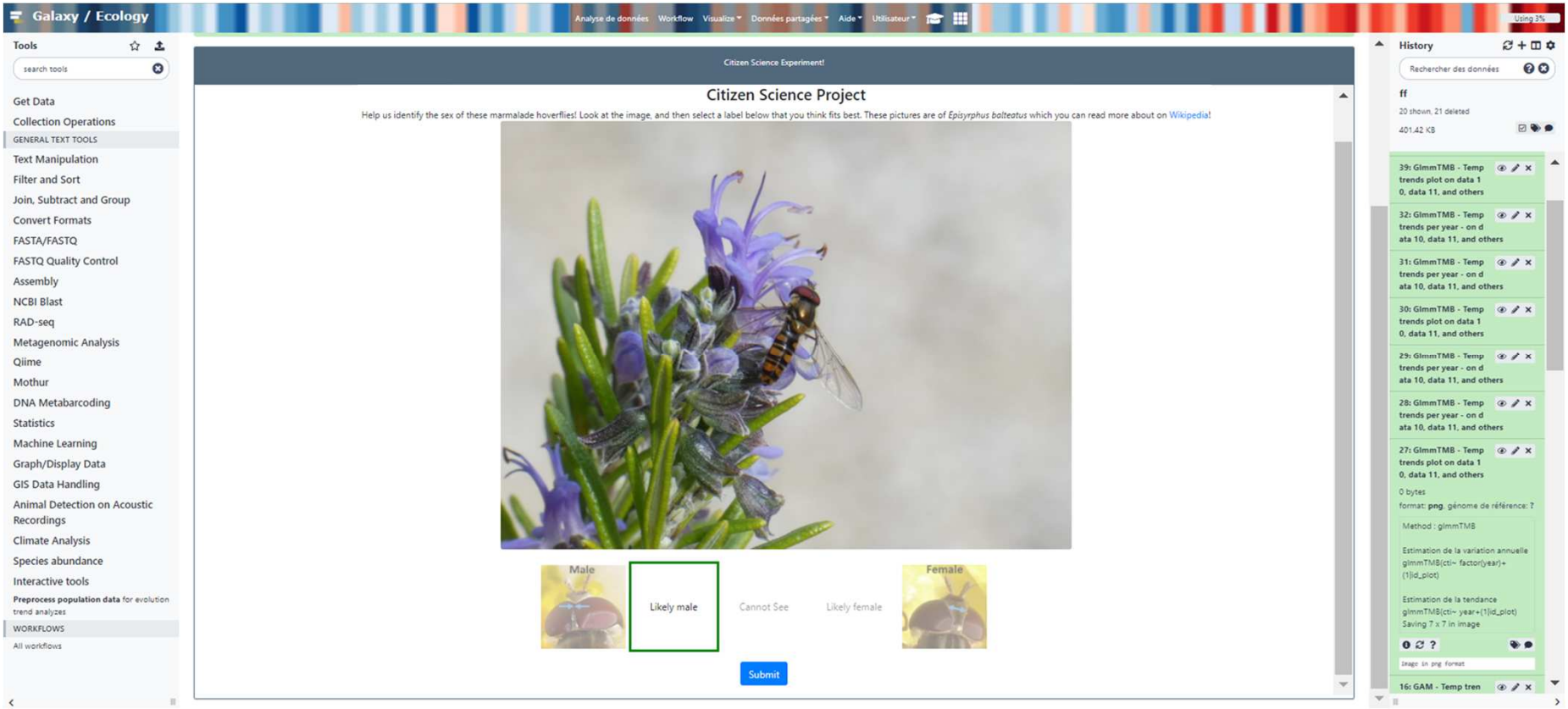
Cofest « vue.js » GUI?

Crowd sourcing through Galaxy

GAPARS project

MOODA concept (Massively Open Online Data Analysis)

Crowdsourcing with hoverflies (syphres) images from **SPIPOLL** project



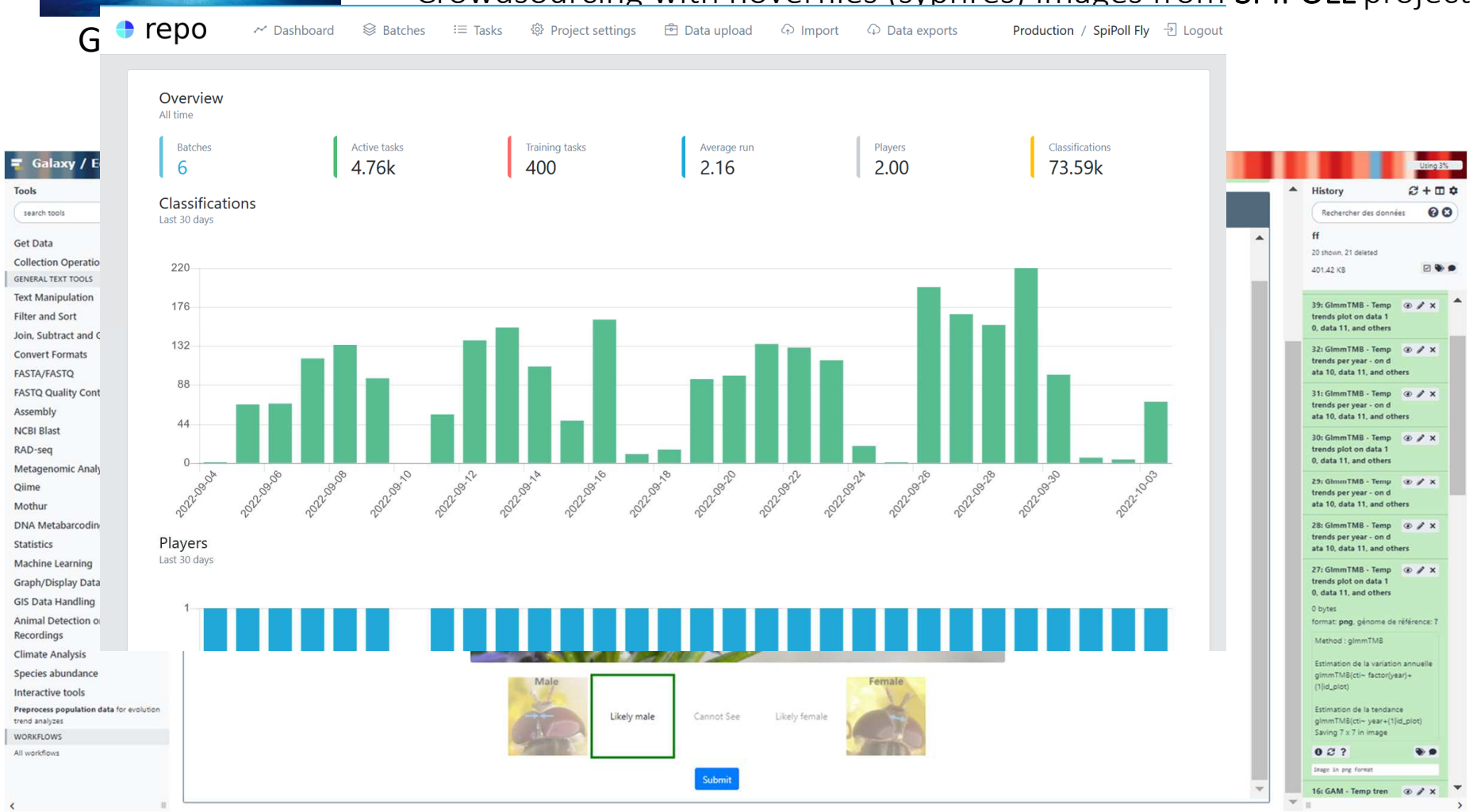
63 320 classifications in 2,5 years !

Crowd sourcing through Galaxy

MOODA concept (Massively Open Online Data Analysis)

Crowdsourcing with hoverflies (syphres) images from **SPIPOLL** project

G



63 320 classifications in 2,5 years !

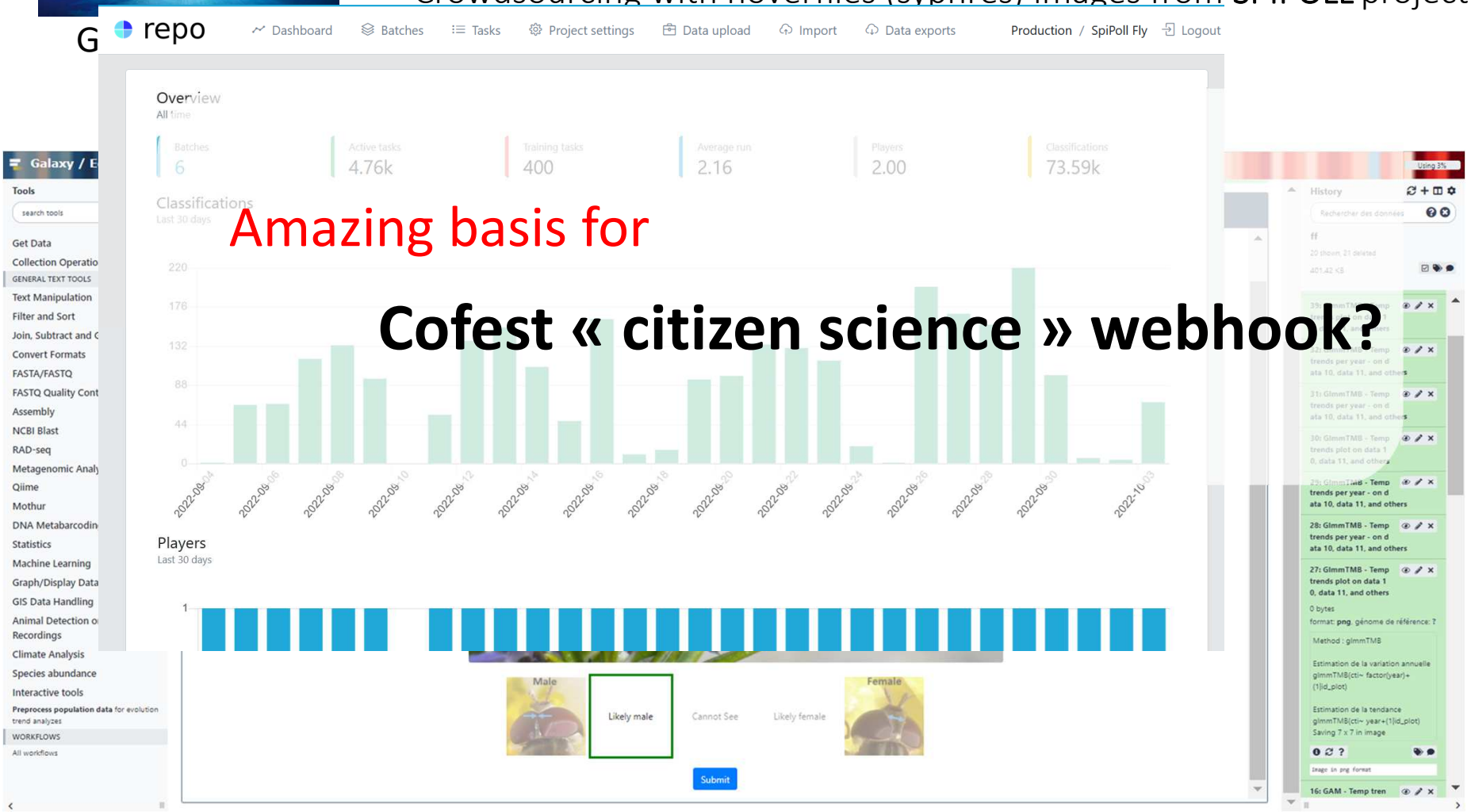
Crowd sourcing through Galaxy

THE FRAMEWORK PROGRAMME FOR RESEARCH AND INNOVATION
HORIZON 2020

MOODA concept (Massively Open Online Data Analysis)

Crowdsourcing with hoverflies (syphres) images from **SPIPOLL** project

G



Amazing basis for

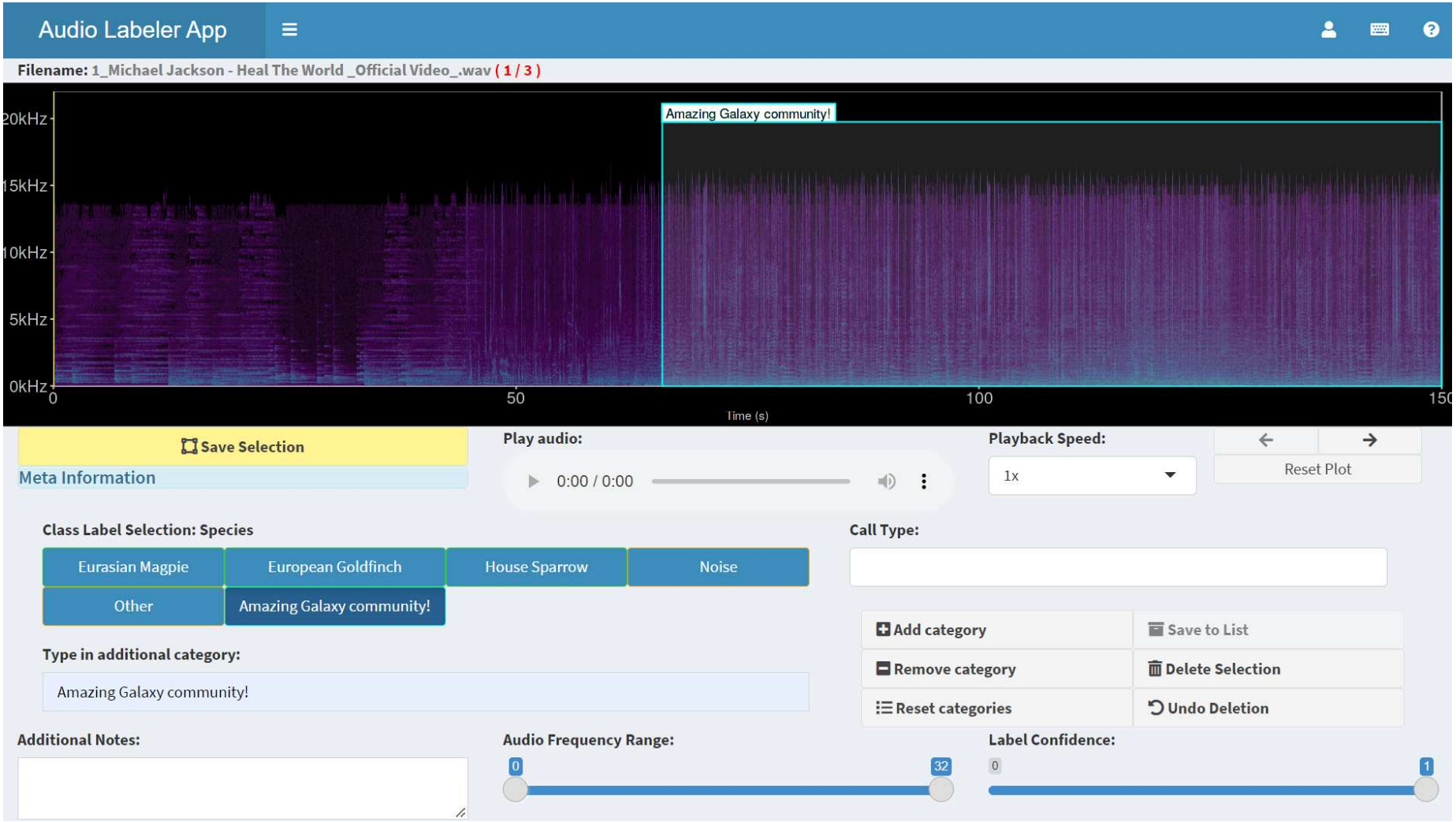
Cofest « citizen science » webhook?

63 320 classifications in 2,5 years !

Crowd sourcing through Galaxy

MOODA concept (Massively Open Online Data Analysis)

https://tinyurl.com/galaxymooda



Thank you !

PNDB team

Coline Royaux – engineer R / Galaxy dev (workflows to compute biodiversity indicators)

Elie Arnaud – engineer R Shiny / knowledge – metadata dev

Marie Jossé– engineer R / Galaxy dev

Julien Sananikone – engineer DevOps / sys admin / web dev

Olivier Norvez – animation coordinator

Yvan Le Bras – Beta tester *yvan.le-bras@mnhn.fr*

https://www.pndb.fr/

PNDB « bricks »:

MetaShARK Metadata work:
https://youtu.be/OVViSMzRGtw

Data metadata portal:
https://youtu.be/STwsYDHEt2A

Galaxy Europe demo:
- https://youtu.be/HelAHggX6D4
- Essential biodiversity variables on Galaxy: implementing PAMPA
- Producing biodiversity indicators from citizen science projects