

0% song guaranteed



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



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,
                  "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(" -> conservation de",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
# write.table(tab, filename)
# write.table(tab, chemin)
colnames(tab) <- sub("nombre.", "", colnames(tab))
return(tab)
}

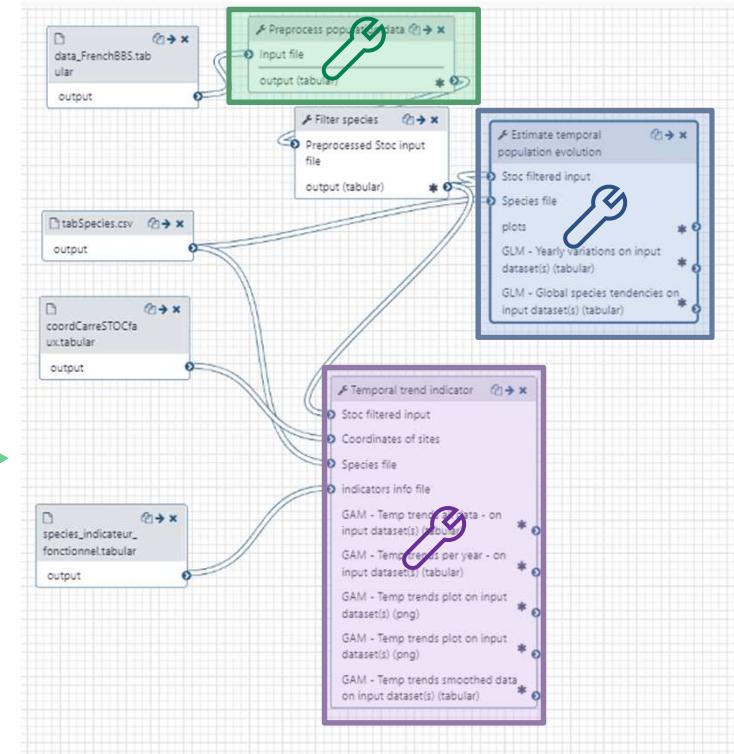
## sous jeux de donnees si choix d espece d annee ou d un pourcentage de carres
makesousTab <- function(tab, vecsp=NULL, echantillon=1,
                         methodeEchantillon="carre", vecannees=NULL) {
  cat(" -- Fabrication du sous jeu de donnees --\n")
  flush.console()
  ## reduction de la table 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 années
  if(!is.null(vecannees)) {
    tab <- subset(tab,annee>=vecannees[1] & annee <= vecannees[2])
  }

  ## calcul des proportions de carres utilisés
  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...

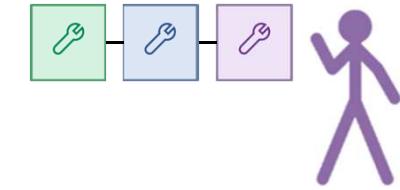
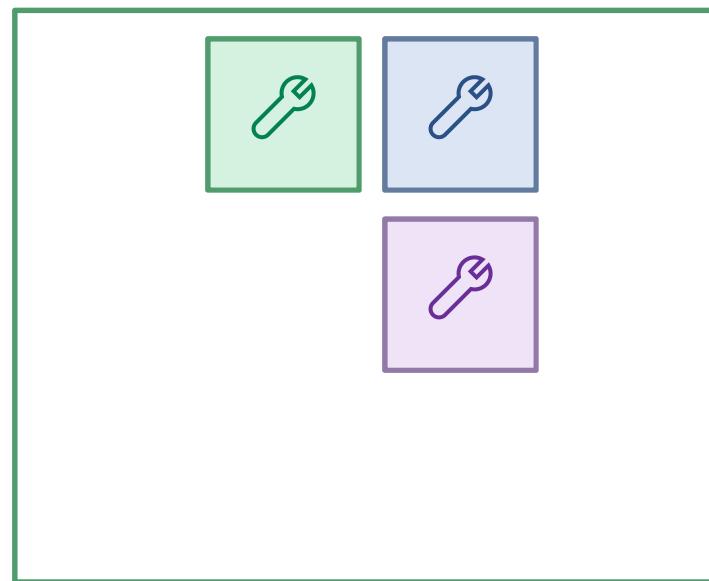
Several atomized R scripts for several input datafiles



Context – We need Sharing & Generalization

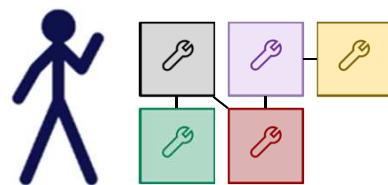
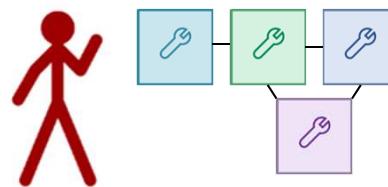
participation	Nuit	num_micro	groupe	espece	nb_contacts
1 55de2cd52121b1000d27430e	2015-07-26	0	bat	Barbar	1
2 55de2cd52121b1000d27430e	2015-07-26	0	bush-cricket	Barfis	1
3 55de2cd52121b1000d27430e	2015-07-26	0	noise	noise	5022
4 55de2cd52121b1000d27430e	2015-07-26	0	bush-cricket	Decalb	5
5 55de2cd52121b1000d27430e	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



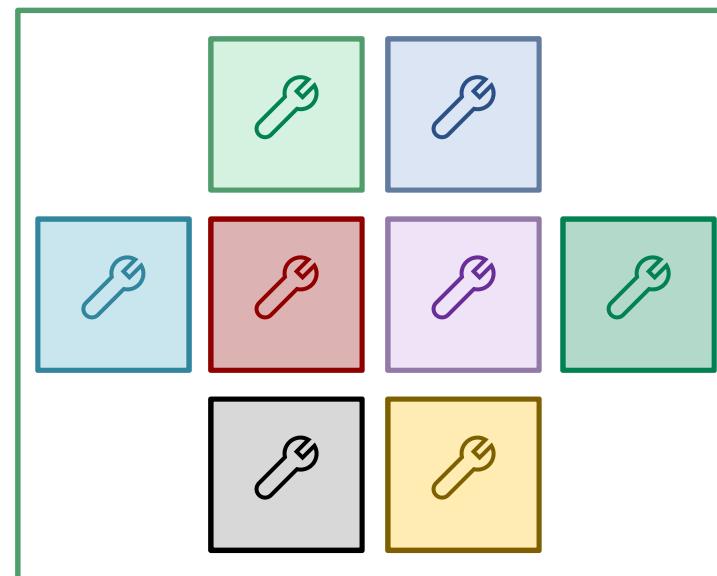
Context – We need Sharing & Generalization

	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
-	-	2010	ACCGEN	-

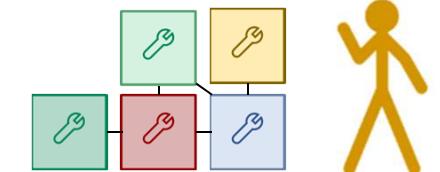
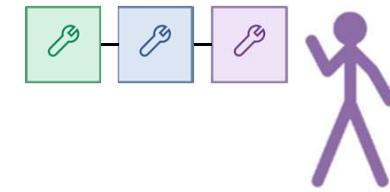


	Unité	obs	rotation	codeSp	sexe	taille	classe_taille	poids	nb_Ind
1	AS140155	3	Hemifasc	-999	-999	P	-999	-999	1
2	AS140159	1	Nasosp.	-999	-999	P	-999	-999	3
3	AS140159	3	Gompvari	-999	-999	P	-999	-999	1
4	AS140160	3	Gompvari	-999	-999	P	-999	-999	1
5	AS140099	2	Parumult	-999	-999	P	-999	-999	1
6	AS140088	1	Varilout	-999	-999	P	-999	-999	1
7	AS140088	2	Gompvari	-999	-999	P	-999	-999	2
8	AS140041	1	Nasosp.	-999	-999	P	-999	-999	5
9	AS140044	1	Parumult	.000	.000	n	.000	.000	4

Toolshed

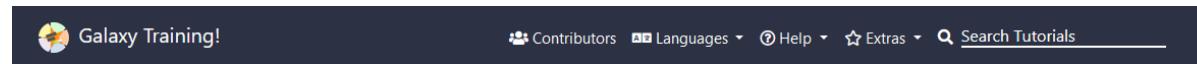


	participation	Nuit	num_micro	groupe	espece	nb_contacts
1	55de2cd52121b1000d27430e	2015-07-26	0	bat	Barbar	1
2	55de2cd52121b1000d27430e	2015-07-26	0	bush-cricket	Barfis	1
3	55de2cd52121b1000d27430e	2015-07-26	0	noise	noise	5022
4	55de2cd52121b1000d27430e	2015-07-26	0	bush-cricket	Decalb	5
5	55de2cd52121b1000d27430e	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



	Survey	Year	Quarter	Area	AphidID	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/>

A screenshot of a web browser displaying the Galaxy Training! website. The title bar says "Galaxy Training!". The main content area shows a "Climate" tutorial. At the top, there are links for "Contributors", "Languages", "Help", "Extras", and a search bar. Below that, a section titled "Climate" with the subtitle "Learn to analyze climate data through Galaxy." includes a note about viewing materials in different languages. A "Requirements" section lists a link to "Introduction to Galaxy Analyses". The main content area is titled "Material" and lists several lessons with their respective tool icons: "Introduction to climate data", "Functionally Assembled Terrestrial Ecosystem Simulator (FATES)", "Functionally Assembled Terrestrial Ecosystem Simulator (FATES) with Galaxy Climate JupyterLab", "Getting your hands-on climate data", "Pangeo ecosystem 101 for everyone - Introduction to Xarray Galaxy Tools", "Pangeo Notebook in Galaxy - Introduction to Xarray", and "Visualize Climate data with Panoply netCDF viewer".

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

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



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

The screenshot shows two separate sections of the Galaxy Training website:

Climate Section:

- Header:** Galaxy Training!, Contributors, Languages, Help, Extras, Search Tutorials.
- Title:** Climate
- Description:** Learn to analyze climate data through Galaxy.
- Note:** You can view the tutorial materials in different languages by clicking the dropdown icon next to the title.
- Requirements:** Before diving into this topic, we recommend you to have a look at:
 - Introduction to Galaxy Analyses
- Material:** A table with columns for Lesson, Slides, Hands-on, Recordings, Input dataset, and Workflows. It lists several tutorials:
 - 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

Ecology Section:

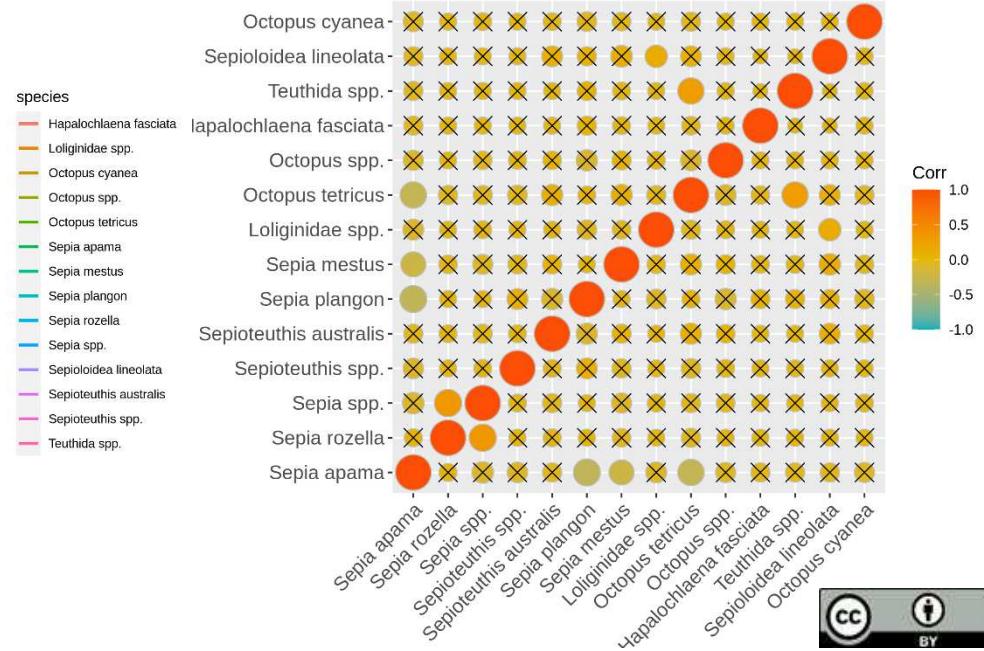
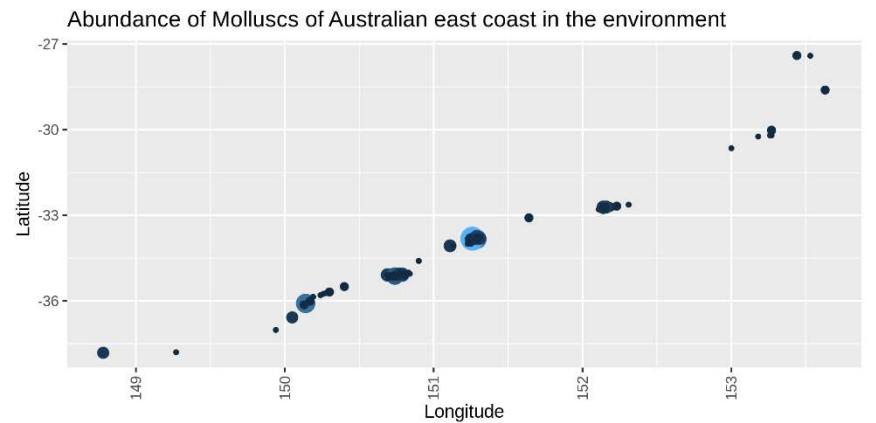
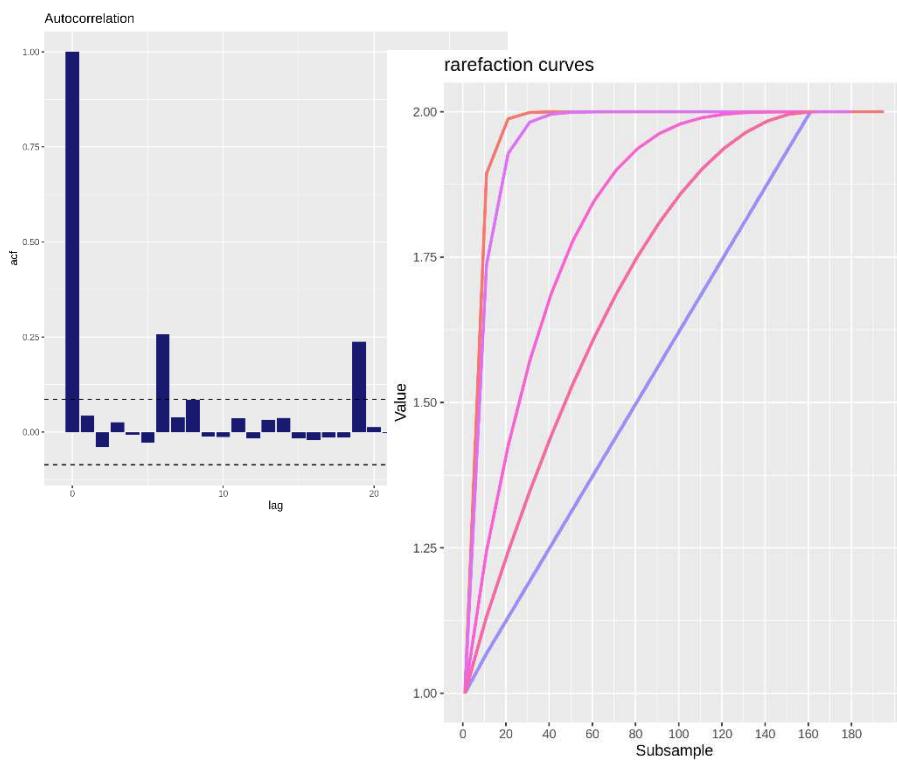
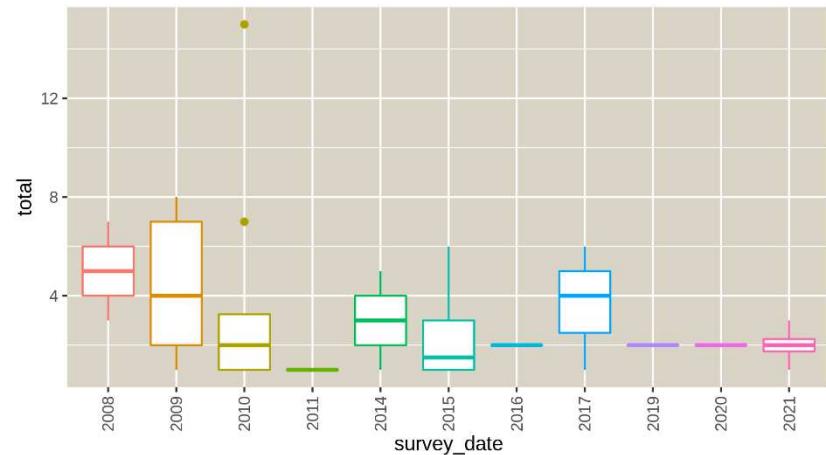
- Header:** Galaxy Training!, Contributors, Languages, Help, Extras, Search Tutorials.
- Title:** Ecology
- Description:** Learn to analyse Ecological data through Galaxy.
- Note:** You can view the tutorial materials in different languages by clicking the dropdown icon next to the slides (Slides) and tutorial (Tutorial) buttons below.
- Requirements:** Before diving into this topic, we recommend you to have a look at:
 - Introduction to Galaxy Analyses
- Material:** A table with columns for Lesson, Slides, Hands-on, Recordings, Input dataset, and Workflows. It lists several tutorials:
 - Biodiversity data exploration
 - Compute and analyze biodiversity metrics with PAMPA toolsuite
 - Metabarcoding/eDNA through Obitoools
 - 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

Bottom Right: CC BY license logo

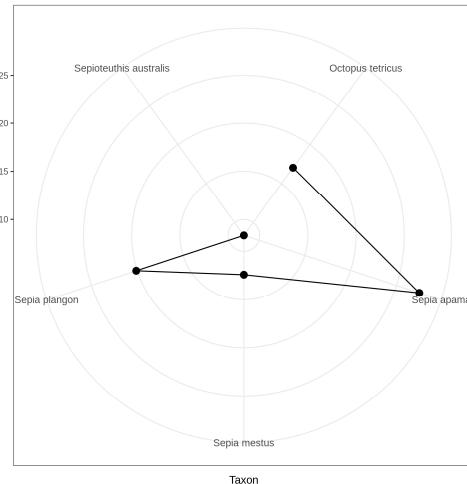
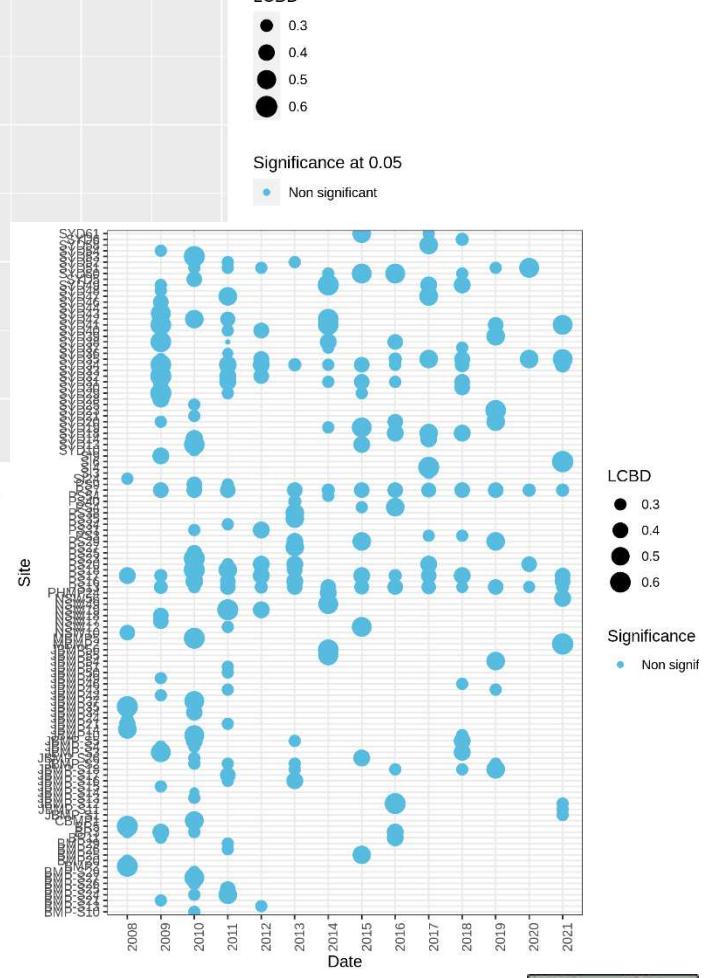
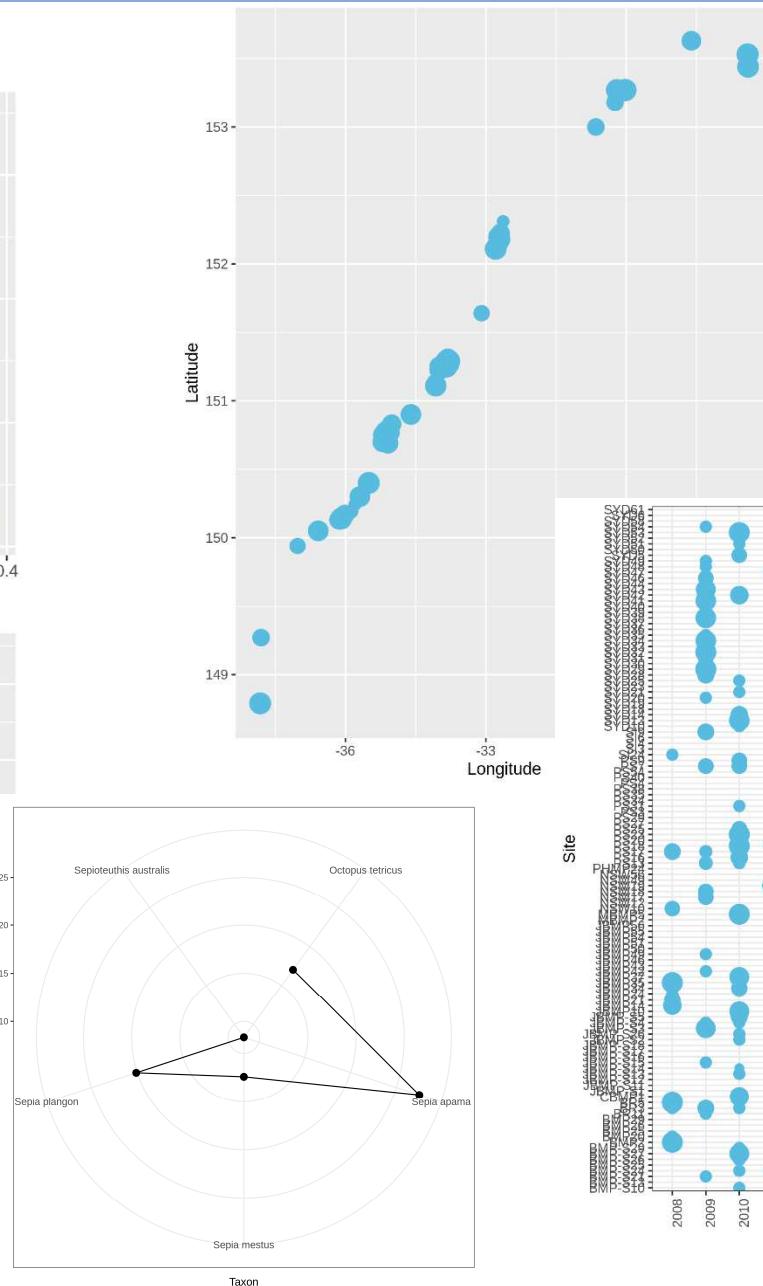
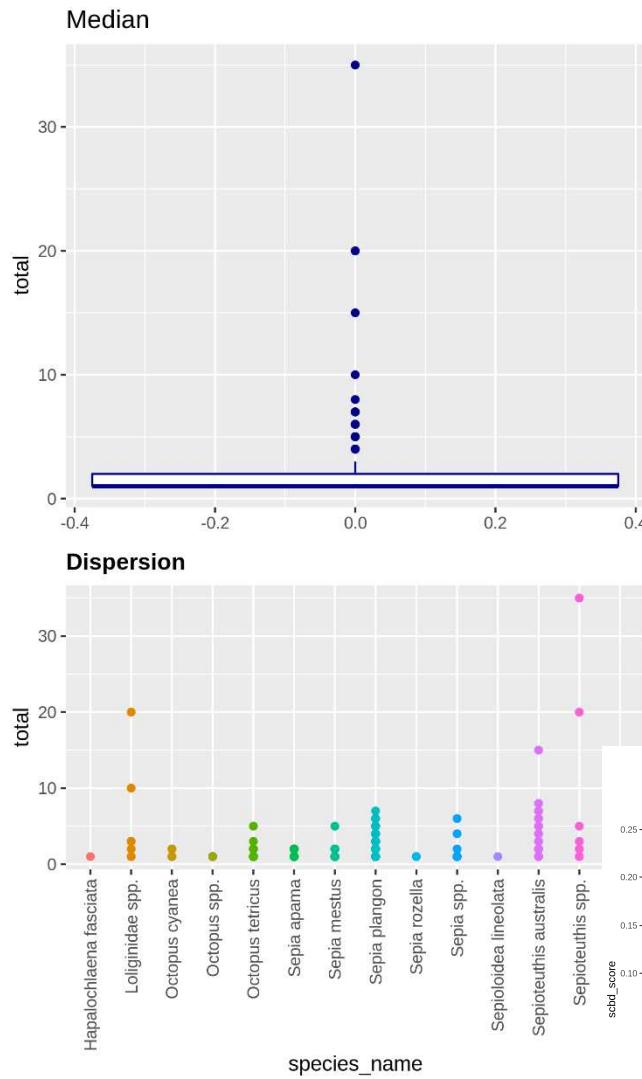
Galaxy-E Killer workflows

- 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

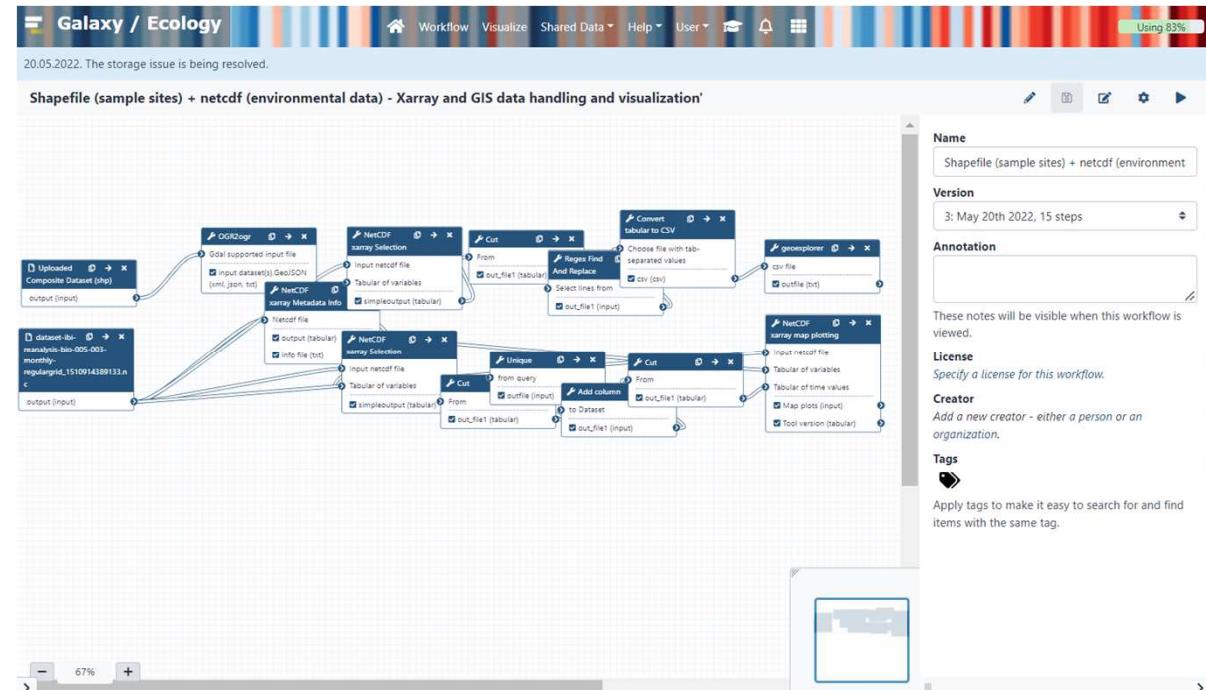
Fouilloux *et al.* EGU22 Pangeo for everyone with Galaxy

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

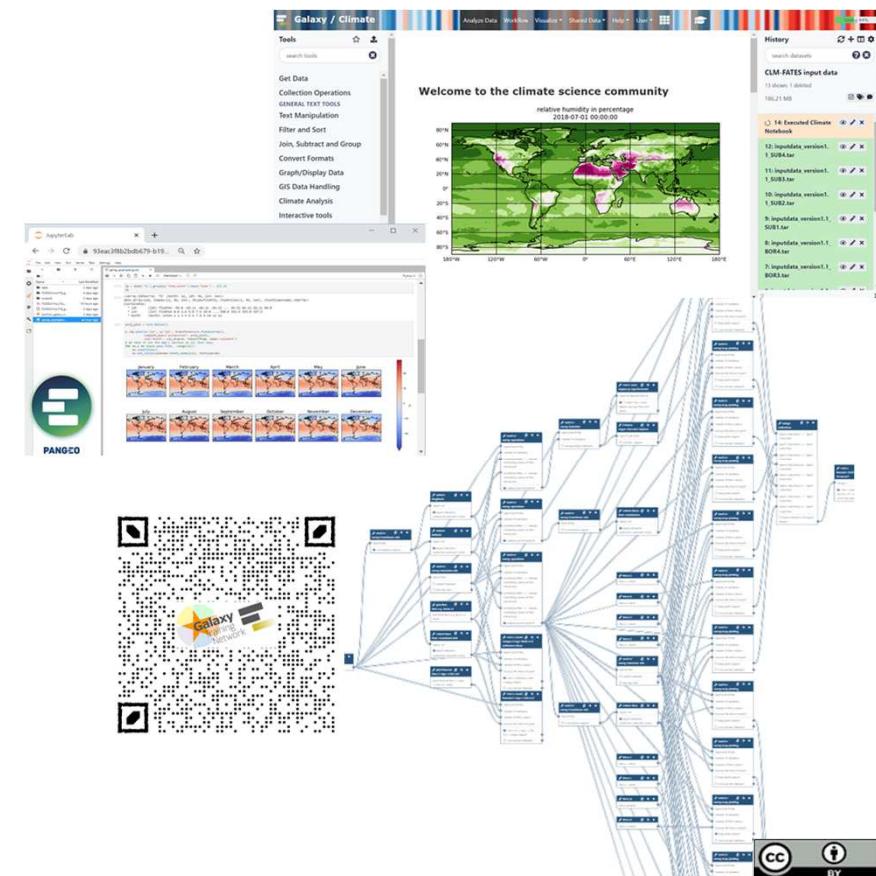
Now: a **Galaxy workflow** mixes scripts, GDAL & Xarray tools making it easily accessible and (re)-runnable.



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.



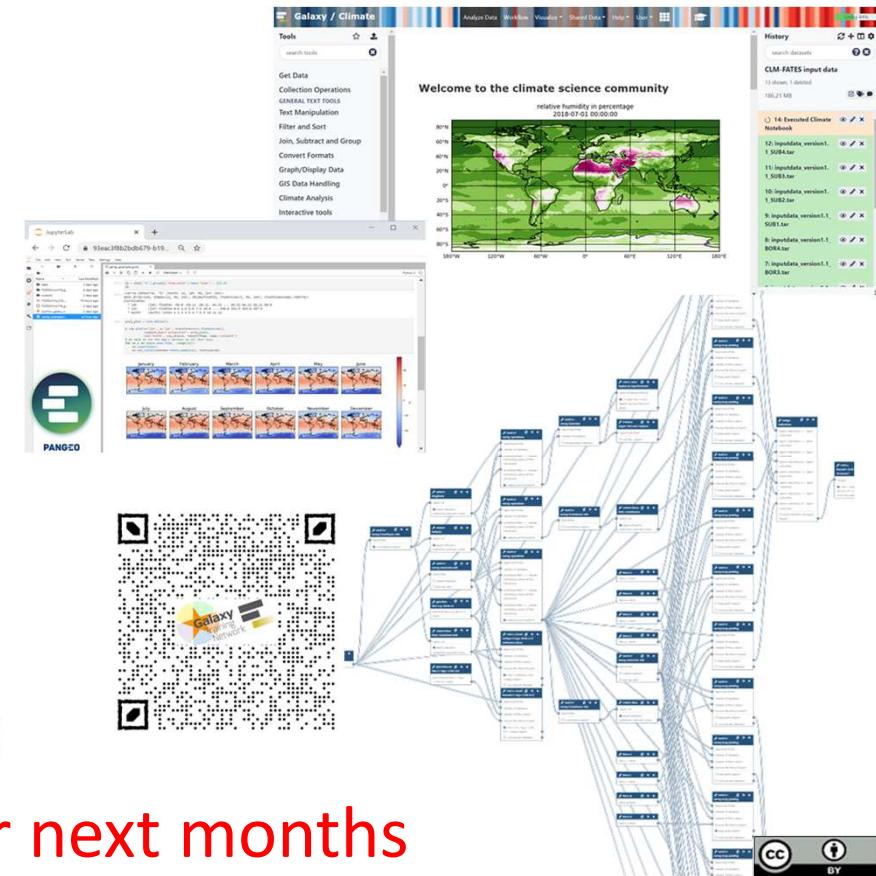
Slides from Fouilloux et al 2022 From EGU22 <https://meetingorganizer.copernicus.org/EGU22/EGU22-5709.html?reply>



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

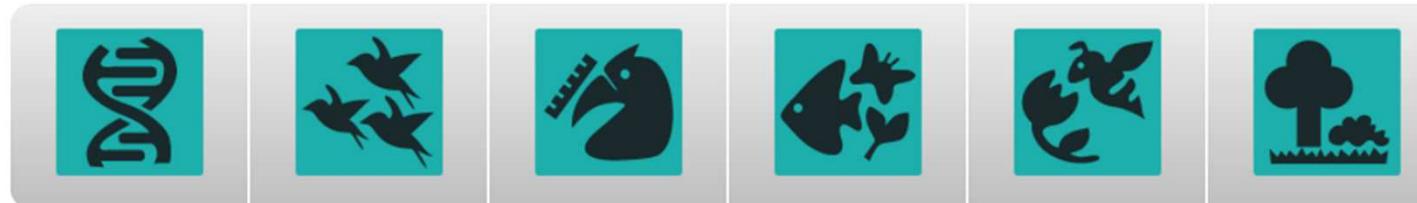
PNDB / Galaxy-E work for next months



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



Essential Biodiversity Variables workflows



Composition génétique

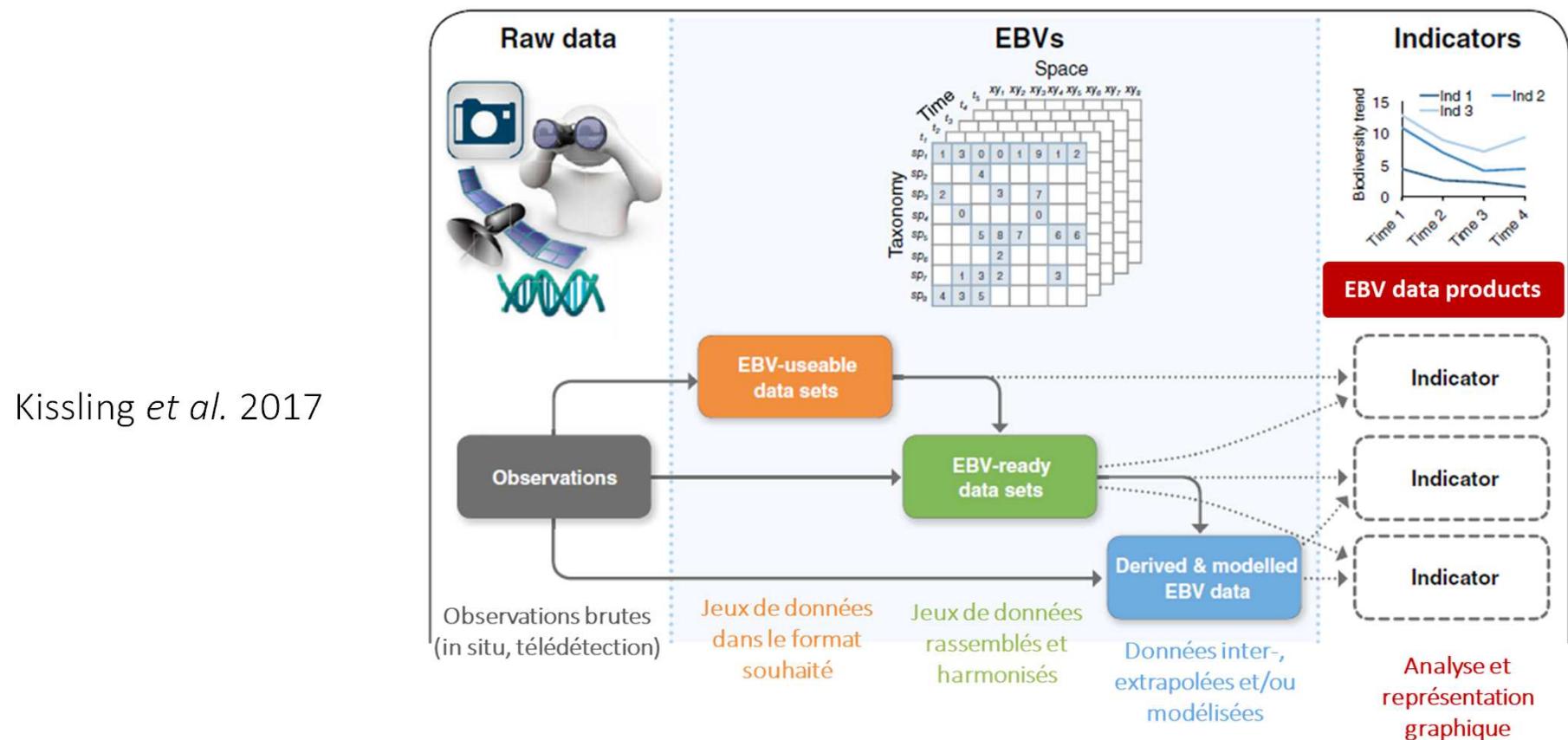
Espèces - Populations

Espèces - Traits

Composition de la communauté

Fonction de l'écosystème

Structure de l'écosystème



EBV workflows: STOC

<https://bit.ly/3AdFDK>



Biodiversity data

1 Preprocess data

2 Filter data



Community

Analyze community indexes

Temporal trend indicator using Galaxy

Spec filtered input
No tabular dataset available.

input Stoc count file, shaped and filtered
Coordinates of sites
No tabular dataset available.

input location file
No tabular dataset available.

Species file
No tabular dataset available.

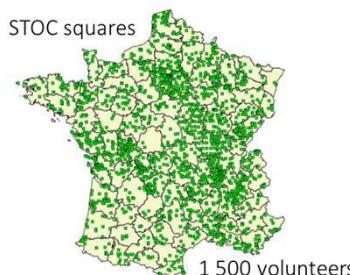
Input species tabular file, with 3 columns (species ID, species name, species scientific name, specialization status).

Indication rate file
No tabular dataset available.

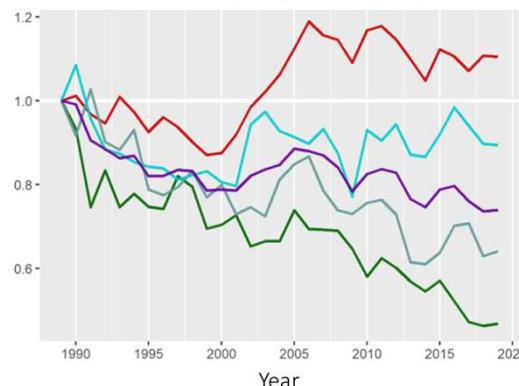
Choose the index you want to compute
Generalists

Available index
Specify advanced parameters
No, use program defaults.

STOC Temporal population trend indicator

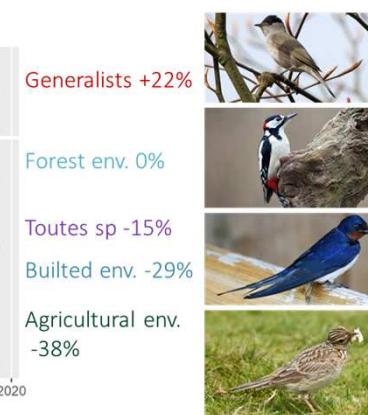


Estimate variations per specialization group



3

Analyze species abundance



Preprocess population data for evolution trend analyzes (Galaxy Version 0.0.1)

Input file

No tabular dataset available.

Population count file, with location, date, species and abundance.

Execute

STOC preprocess population data

Filter species with rare and low abundances (Galaxy Version 0.0.1)

Preprocessed Stoc input file

No tabular dataset available.

Output file from the 'Preprocess population data tool'

Execute

STOC Filter species with rare and low abundances

Favorite Options

Favorite Options

Favorite Options



Species - population

Estimate temporal population evolution by specialization group (Galaxy Version 0.0.1)

Yearly variation dataset

No tabular dataset available.

Output from the 'Estimate temporal population evolution by species' tool.

Global tendencies dataset

No tabular dataset available.

Output from the 'Estimate temporal population evolution by species' tool.

Species file

No tabular dataset available.

Input species tabular file, with 5 columns (species ID, species name, species scientific name, specialization status).

Specify advanced parameters

No, use program defaults.

Email notification

Send an email notification when the job completes.

Execute

STOC Estimate species population evolution

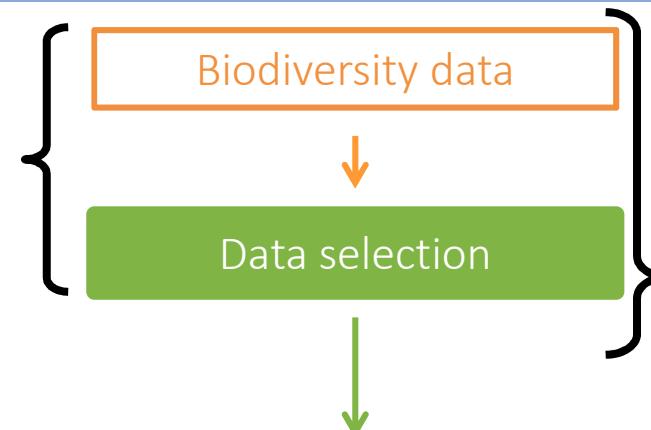


EBV workflows: PAMPA

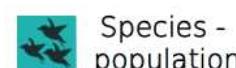
wwz.ifremer.fr/pampa

Existing accessible & reusable Galaxy tools

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



Community



Pre-processed data



Species -
population

PAMPA

Compute
community metrics

1

Compute
population metrics

Calculate presence absence table (calculate presence absence table from observation data (Galaxy Version 0.8))

Input file
No tabular dataset available

Observation data file: with location, year, species and abundance.

Email notification
Yes No

Send an email notification when the job completes.

✓ Execute

Calculate presence absence table from abundance data

GLM on community
metrics

2

Metric ~ site + year + habitat

GLM on population
metrics

Time-series plot
from GLM results

3

Compute GLM on community data (Compute GLM on community data (Galaxy Version 0.8))

Input matrix file
No tabular dataset available

Observation data file: with location, year, species and abundance.

Choose the community metrics you want to compute:
Select all

Execute

Calculate community metrics from abundance data

Compute GLM on community data with selected interest variables (Compute GLM on community data with selected interest variables (Galaxy Version 0.8))

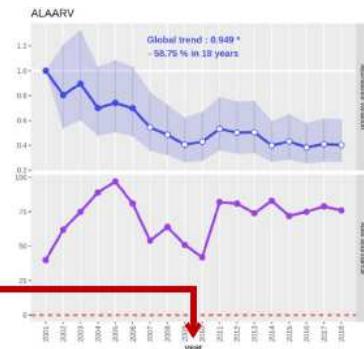
Input matrix file
No tabular dataset available

Observation data file: with location, year, species and abundance.

Choose the interest variables you want to include in your analysis:
Year effect
Habitat effect

Execute

Compute GLM on community data with selected interest variables



Compute GLM on population data (Compute GLM on population data (Galaxy Version 0.8))

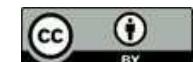
Input matrix file
No tabular dataset available

Observation data file: with location, year, species and abundance.

Choose the population metrics you want to compute:
Select all

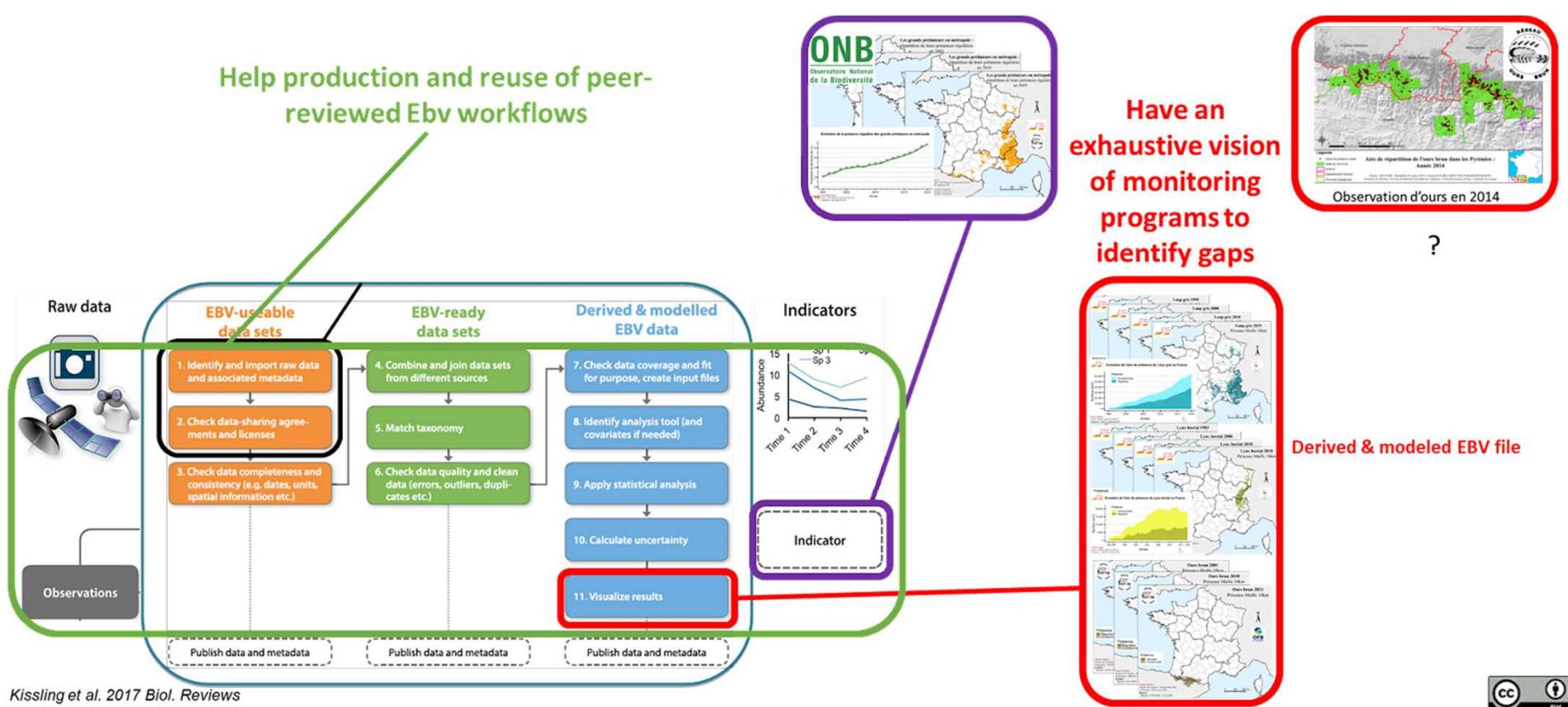
Execute

Compute GLM on population data with selected interest variables



Essential Biodiversity Variables workflows

Help BONs to identify gaps & reuse EBV workflows



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

Get species occurrences data from GBIF, ALA, iNAT and others (Galaxy Version 0.9.0)

Scientific name of the species

Genus species format, eg : Canis lupus

Data source to get data from
 Select/Unselect all

Any combination of gbif, bison, inat, ebird, antweb, ala, idigbio, obis, ecoengine and/or vertnet

Number of records to return
500

This is passed across all sources

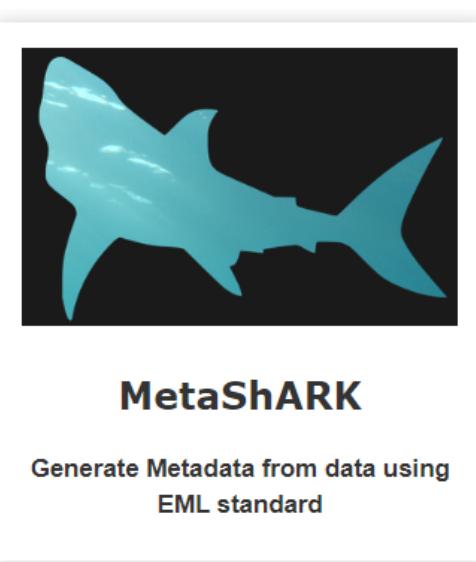
Email notification
 No
Send an email notification when the job completes.

Execute

Get species occurrences data

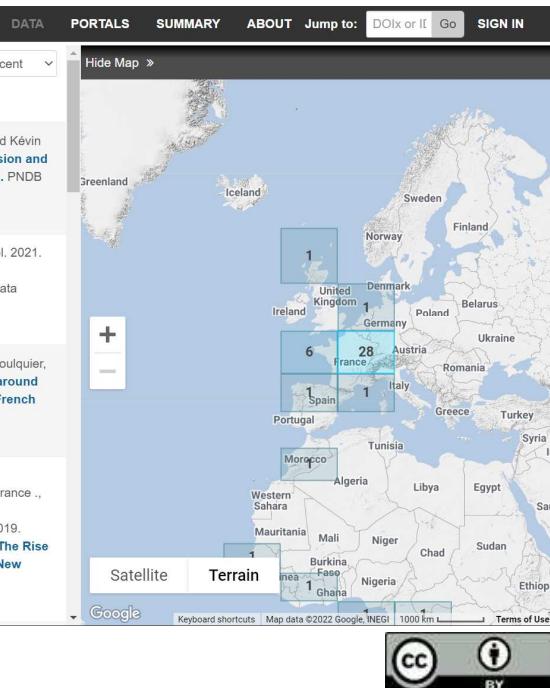
What it does
Search species occurrences across a single or many data sources.

Import Biodiversity
occurrences data



MetaShARK

Generate Metadata from data using EML standard



Convert metadata

xmlstarlet convert a metadata XML file in one standard to another (Galaxy Version 1.6.1)

input xml file to convert
 18: xmlstarlet on data 2 and data 17: Converted xml

A xml file corresponding to a xsd schema you want to convert in another.

input xsd conversion file
 2: iso2eml_all_in_one.xsl

A xsd file describing the mapping between a first xsd specification to another.

Email notification
 No
Send an email notification when the job completes.

Execute

What it does

This tool converts a xml file to another using a xsd conversion file to specify the translation to be done, from a xsd schema to another.

Inputs

A xml metadata file using a standard (for exemple EML, ISO19115,...) and a xsd file describing the mapping between the standard terms from input searched output standard.

Outputs

A xml metadata file using a new standard (for example ISO19115, EML, ...).

Upload data an metadata

Pôle National de Données de Biodiversité Data Catalog

DATA PORTALS SUMMARY ABOUT Jump to: DOIx or If Go SIGN IN

Search Search phrase

Datasets 1 to 25 of 68 Sort by Most recent

1 2 3 Next

Filter by:

- Data attribute
- Data files
- Creator
- Year
- Identifier
- Taxon
- Location
- Access

Camille Leroux, Christian Kerbiriou, Isabelle Le Viol, Nicolas Valet, and Kévin Barré. 2022. **Data from: Distance to hedgerows drives local repulsion and attraction of wind turbines on bats: implications for spatial siting.** PNDB Data Repository. um:uuid:4267c75d-1707-41f0-8fe6-5e1348902d4e.

Constance Blary, Kévin Barré, Christian Kerbiriou, and Isabelle Le Viol. 2021. **Assessing the importance of field margins for bat species and communities in intensive agricultural landscapes - Data.** PNDB Data Repository. um:uuid:cb192b3b-dd23-4f6c-abd6-d0e3964c4b79.

Lorraine Coché, Elie Arnaud, Bouvet Laurent, Romain David, Eric Foulquier, et al. 2021. **Kakilia database of marine mammal observation data around the French archipelago of Guadeloupe in the AGOA sanctuary - French Antilles.** PNDB Data Repository. doi:10.48502/8bb5-pk85.

Institut de Recherche pour le Développement, UMR DIADE, France., SouthGreen Development Platform, Agropolis Campus, Montpellier, France., Africa Rice Center, Benin., CEA, Institut de Biologie François Jacob, Genoscope, Evry, France., CNRS, UMR 8030, Evry, France., et al. 2019. **African rice population genomics dataset or title of the article : "The Rise and Fall of African Rice Cultivation Revealed by Analysis of 246 New Genomes".** PNDB Data Repository. doi:10.48502/xcah-3w89.

Satellite Terrain Google Keyboard shortcuts Map data ©2022 Google, INEGI Terms of Use

Ecological research data management

Get species occurrences data from GBIF, ALA, iNAT and others (Galaxy Version 0.9.0)

Scientific name of the species

Genus species format, eg : Canis lupus

Data source to get data from

Select/Unselect all

Any combination of gbif, bison, inat, ebird, antweb, ala, idigbio, obis, ecoengine and/or vertnet

Number of records to return

500

This is passed across all sources

Email notification

No
Send an email notification when the job completes.

Execute

Get species occurrences data

What it does

Search species occurrences across a single or many data sources.

Import Biodiversity
occurrences data

Amazing basis for

PNDB / Galaxy-E work for next months

Convert metadata

xmlstarlet convert a metadata XML file in one standard to another (Galaxy Version 1.0.1)

input xml file to convert

18: xmlstarlet on data 2 and data 17: Converted xmi

A xml file corresponding to a xsd schema you want to convert in another.

input xs conversion file

2: iso2eml_all_in_one.xsl

A xs file describing the mapping between a first xsd specification to another.

Email notification

No
Send an email notification when the job completes.

Execute

What it does

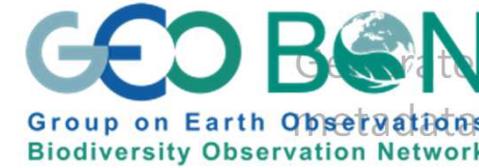
This tool converts a xml file to another using a xsl coversion file to specify the translation to be done, from a wsdl schema to another.

Inputs

A xml metadata file using a standard (for exemple EML, ISO19115...) and a xsd file describing the mapping between the standard terms from input searched output standard.

Outputs

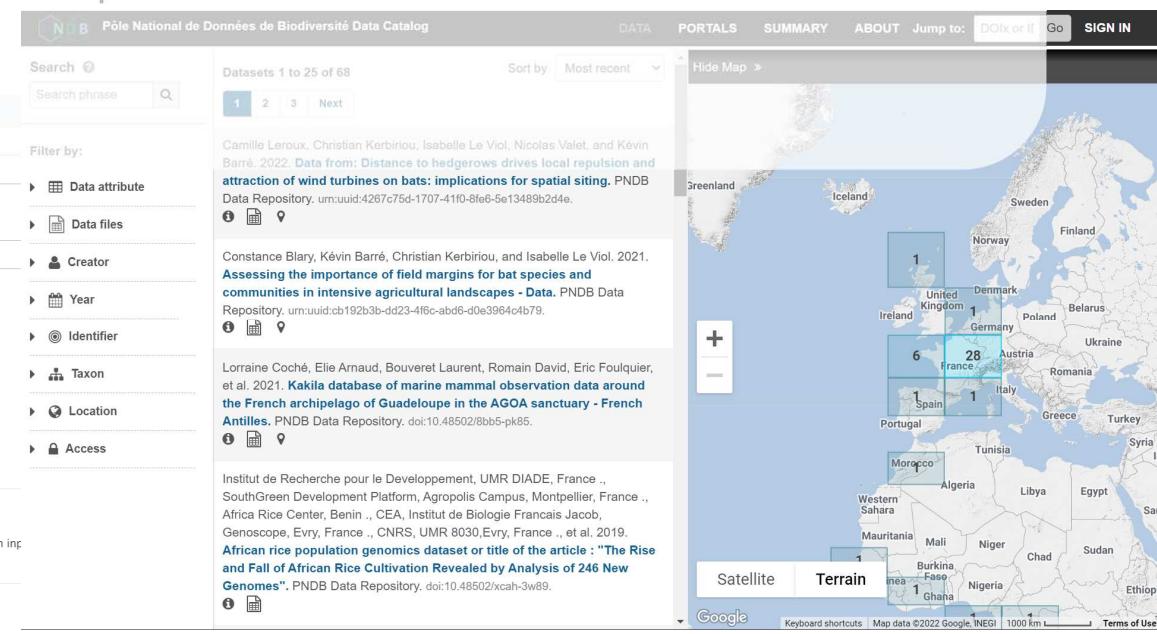
A xml metadata file using a new standard (for example ISO19115, EML, ...).



Biodiversity Observation Network

Group on Earth Observations

Biodiversity Observation Network

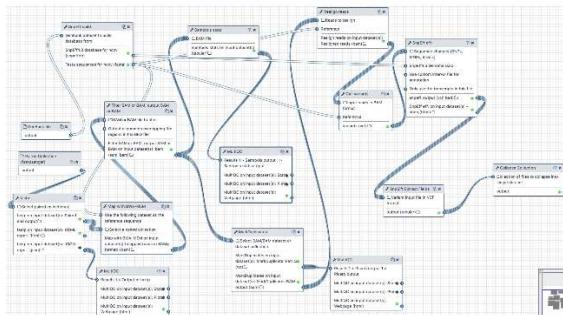


Not only asynchronously

=> Thanks to Galaxy interactive tools GxIT

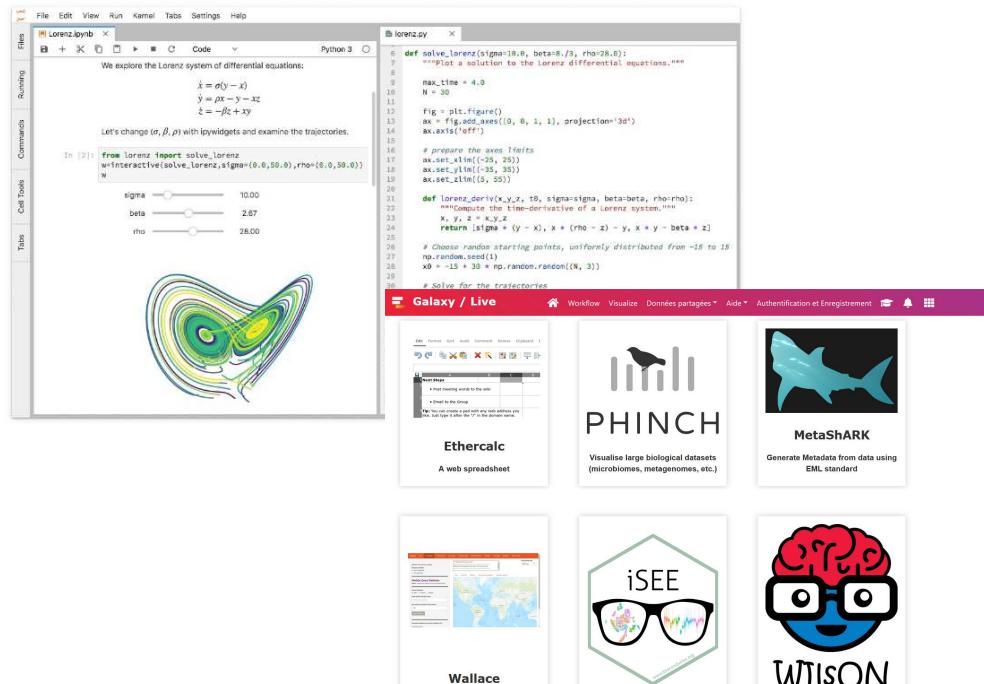
One specific workflow goal

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 async ones)



Interactive tools

Interactive Jupyter Notebook

[GPU enabled Interactive Jupyter Notebook for Machine Learning](#)

Interactive Climate Notebook

Interactive Pangeo Notebook

RStudio

Pylon Interactive Jupyter Notebook

[HiGlass](#) an interactive Hi-C data visualizer.

[OpenRefine](#) Working with messy data

Ubuntu XFCE Desktop

[Panoply](#) interactive 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

[radianit](#) Data analytics using R Shiny app

[EtherCalc](#)

[VRM Editor](#) interactive 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

A screenshot of the Refine web application. The top navigation bar includes 'Create Project', 'Open Project' (which is highlighted in blue), 'Import Project', and 'Language Settings'. Below the navigation is a table with columns 'Last modified' (today 5:29 PM) and 'Name' (input.tabular). The main area shows a 'Dataset' dropdown set to 'Galaxy' and two dropdowns for 'x Var' (key) and 'y Var' (coordinateUncertaintyInMeters). A red banner at the bottom of the interface reads 'Amazing for R Shiny apps!'

A screenshot of the sPOCC web application. It features a sidebar with 'Obtain Occurrence Data' and 'Modules Available' (Query Database, User-specified, Galaxy History User). The main area shows a map of North America with red dots representing Chrysomys picta occurrences. A search bar for 'Enter species scientific name' contains 'Chrysomys picta'. Below the map are buttons for 'Set maximum number of occurrences' (1000) and 'Query Database'. At the bottom, there's a download section for 'Download database occurrence localities (.csv)' and links for 'Get in Galaxy' and 'Download'.

Amazing for R Shiny apps!

Cofest « R Shiny » subject?

workflows (the `async` ones)

A screenshot of the geoExploreR R Shiny app. The left panel shows a scatter plot with axes 'x' and 'y' ranging from 0 to 300,000. The right panel shows a choropleth map of Europe where countries are colored based on a numerical value, with a legend ranging from -1,400,000,000 to -2,400,000,000. The title 'workflows (the `async` ones)' is displayed above the app.

The application you are currently using is a front-end tool for any user who wants to contribute to a DataOne node repository, also known as [metaData catalogue](#) or [MetaCat](#). To contribute to a MetaCat, you need to login to the metacat in which you want to upload your data package. Then, it will be possible for MetaSHARK to gather the needed informations, as you will see while using the app. The upload feature is accessible from the Upload tab. It will require to have a written EML file available in this instance of MetaSHARK.

MetaSHARK

Welcome in MetaSHARK

MetaSHARK (Metadata Shiny Automated Resources and Knowledge) is a tool designed for ecology data description tasks. The tool relies on ecology metadata standards, and mainly the Ecological Metadata Language. Its vocation is to allow any ecologist to fill in metadata for its dataset to permit the understanding, reusability and reproducibility of his work. But as metadata is becoming more and more complex, this tool is trying to get as user-friendly as possible.

MetaSHARK is written and maintained by the French National Biodiversity Data Hub (PNDB). You can interact with the development team on their [git repository](#).

About EML Assembly Line

Authorship

The [EML Assembly Line](#) package used in this app and its children is the intellectual property of the Environment Data Initiative (EDI). You can find further details on their [git repository](#)

Usage

EMLassemblyline is a metadata builder for scientists and data managers who need to easily create high quality EML metadata for data publication. It emphasizes auto-extraction of metadata, appends value added content, and accepts user supplied inputs through template files thereby minimizing user effort while maximizing the potential of future data discovery and reuse. EMLassemblyline requires no familiarity with EML, is great for managing 10-100s of data packages, accepts all data formats, and supports complex and fully reproducible science workflows. Furthermore, it incorporates EML best practices, is based on a simple file organization scheme, and is not tied to a specific data repository.

(preface by Colin Smith, EDI)

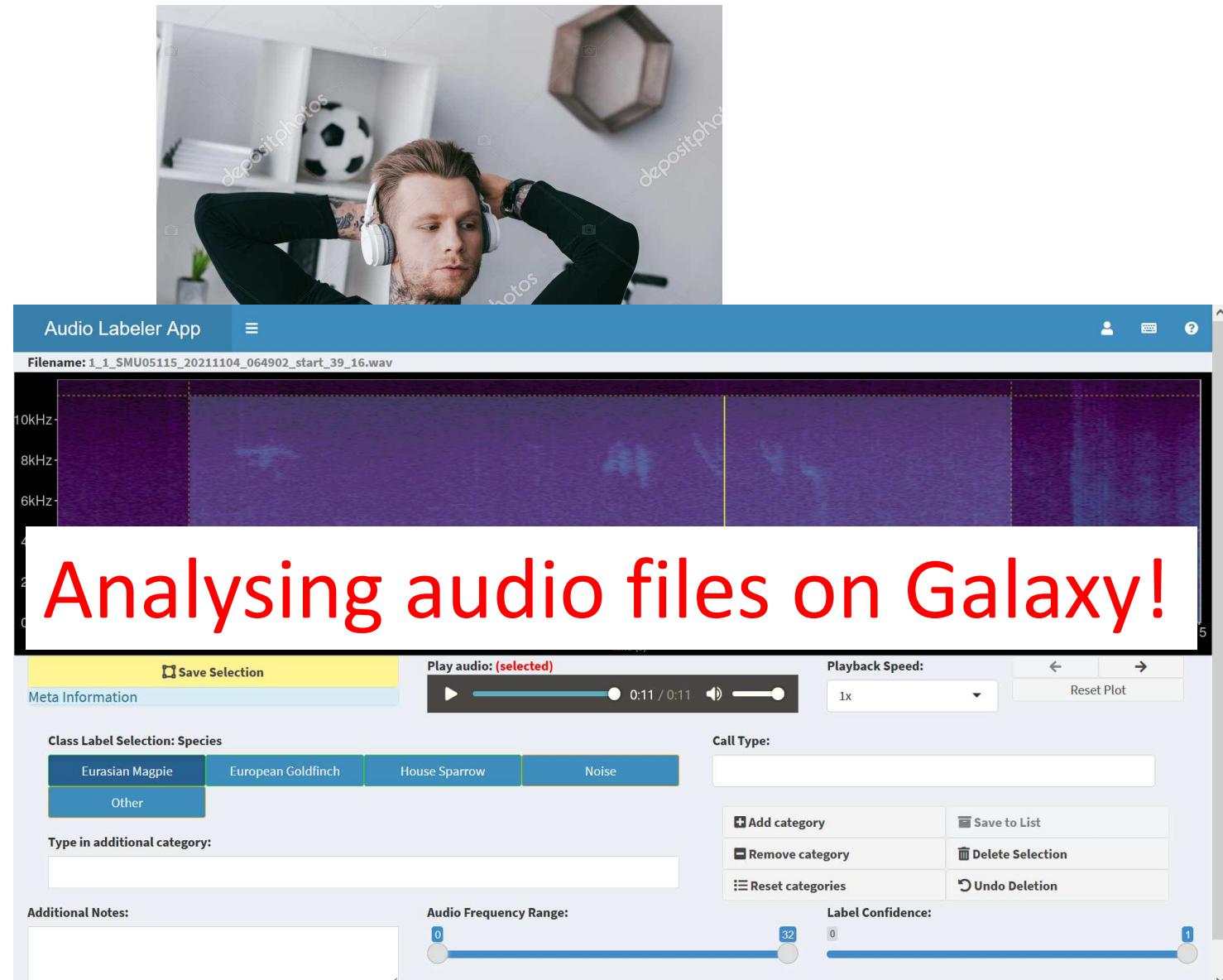
One specific workflow goal

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



One specific workflow goal

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



Not only for scientists

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

Galaxy-Bricks



Galaxy for pupils !
bricks.vigienature-ecole.fr

The screenshot displays the Galaxy Bricks interface, which is a user-friendly version of the Galaxy platform designed for school pupils. The interface includes a sidebar with various tools and a main workspace for managing data and running analyses.

Tools Sidebar:

- Question de recherche
- Importer des données
- Manipuler des données
 - Convertir des dates
 - Opérations sur des lignes (2)
 - Résumer des données
 - Sélectionner des lignes
 - Rechercher/REMPLACER
 - Operations sur des lignes (WIP-repeat)
 - Fusionner le contenu de deux colonnes
- Visualiser
 - Plot w ggplot2
 - Représenter
- Tests statistiques
 - Régression linéaire multiple
 - Regression linéaire simple

Main Workspace:

Bienvenue dans Galaxy Bricks !

Commencer une question de recherche !

Scenario:

Titre: Effet de l'environnement sur les oiseaux

Description: Quel est l'effet de l'environnement sur la diversité des oiseaux ?

Step 7: Représenter (sur Résumer des données on data 3)

A générée le fichier "Représenter on data 4" (type

Warning message:

NA introduced lors de la conversion automatique
Erreur : Must use either variable name or expression when facetting
Exécution arrêtée

Step 6: Représenter (sur Résumer des données on data 3)

Bottom Right:

Galaxy-Bricks



Galaxy for pupils !

bricks.vigienature-ecole.fr

The screenshot shows the Galaxy-Bricks web application interface. On the left, a vertical sidebar lists various data processing and visualization tools:

- Question de recherche (highlighted in red)
- Importer des données
- Manipuler des données
 - Convertir des dates
 - Opérations sur des lignes (2)
 - Résumer des données
 - Sélectionner des lignes
 - Rechercher/REMPLACER
 - Operations sur des lignes (WIP-repeat)
 - Fusionner le contenu de deux colonnes
- Visualiser
 - Plot w ggplot2
 - Représenter
- Tests statistiques
 - Régression linéaire multiple
 - Regression linéaire simple

The central area features a large "Galaxy BRICKS" logo and a button labeled "Commencer une question de recherche !". To the right, a scenario titled "Effet de l'environnement sur les oiseaux" is displayed, showing a list of steps and a warning message.

Scenarios:

- Step 7: Représenter (sur Résumer des données on data 3) - A generated file "Représenter on data 4" (type).
- Step 6: Représenter (sur Résumer des données on data 3) - A warning message: "Warning message: NA introduced during automatic conversion. Error: Must use either variable name or expression when faceting. Execution stopped."

Central Panel:

Amazing basis for Cofest « vue.js » GUI?

Commencer une question de recherche !

Bienvenue dans Galaxy Bricks!

Effet de l'environnement sur les oiseaux

Quel est l'effet de l'environnement sur la diversité des oiseaux ?

7 Représenter (sur Résumer des données on data 3) A généré le fichier "Représenter on data 4" (type)

6 Représenter (sur Résumer des données on data 3)

Warning message:
NA introduced during automatic conversion
Error: Must use either variable name or expression when faceting
Execution stopped

Crowd sourcing through Galaxy



GAPARS project

MOODA concept (Massively Open Online Data Analysis)

Crowdsourcing with hoverflies (syphres) images from SPIPOLL project

A screenshot of the Galaxy platform interface. The left sidebar shows various tools and workflows, including "Collection Operations" and "GENERAL TEXT TOOLS". The main area displays a "Citizen Science Project" titled "Help us identify the sex of these marmalade hoverflies! Look at the image, and then select a label below that you think fits best. These pictures are of *Episyrphus balteatus* which you can read more about on Wikipedia!". It features a large image of a hoverfly on a purple flower. Below the image are four smaller images labeled "Male", "Female", "Likely male" (which has a green border), and "Likely female". A "Submit" button is at the bottom. The right sidebar shows a history of analysis steps, such as "39: GlimmTMB - Temp trends plot on data 1 0, data 11, and others" and "ff 20 shown, 21 deleted".

Galaxy / Ecology

Analyse de données Workflow Visualize Données partagées Aide Utilisateur

Tools

Get Data

Collection Operations

GENERAL TEXT TOOLS

Text Manipulation

Filter and Sort

Join, Subtract and Group

Convert Formats

FASTA/FASTQ

FASTQ Quality Control

Assembly

NCBI Blast

RAD-seq

Metagenomic Analysis

Qiime

Mothur

DNA Metabarcoding

Statistics

Machine Learning

Graph/Display Data

GIS Data Handling

Animal Detection on Acoustic Recordings

Climate Analysis

Species abundance

Interactive tools

Preprocess population data for evolution trend analyses

WORKFLOWS

All workflows

Citizen Science Experiment!

Help us identify the sex of these marmalade hoverflies! Look at the image, and then select a label below that you think fits best. These pictures are of *Episyrphus balteatus* which you can read more about on Wikipedia!

Citizen Science Project

Male Female

Likely male

Cannot See Likely female

Submit

History

Rechercher des données

ff 20 shown, 21 deleted 401,42 KB

39: GlimmTMB - Temp trends plot on data 1 0, data 11, and others

32: GlimmTMB - Temp trends per year - on data 10, data 11, and others

31: GlimmTMB - Temp trends per year - on data 10, data 11, and others

30: GlimmTMB - Temp trends plot on data 1 0, data 11, and others

29: GlimmTMB - Temp trends per year - on data 10, data 11, and others

28: GlimmTMB - Temp trends per year - on data 10, data 11, and others

27: GlimmTMB - Temp trends plot on data 1 0, data 11, and others

0 bytes format: png, génome de référence: T Method: glimmTMB

Estimation de la variation annuelle glimmTMB(ct~ factor(year)+(~id,plot)

Estimation de la tendance glimmTMB(ct~ year+ (~id,plot) Saving 7 x 7 in image

Image in png format

16: GAM - Temp tren

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 repo

Dashboard Batches Tasks Project settings Data upload Import Data exports Production / SpiPoll Fly Logout

Overview
All time

Batches: 6 Active tasks: 4.76k Training tasks: 400 Average run: 2.16 Players: 2.00 Classifications: 73.59k

Classifications
Last 30 days

Date	Classifications
2022-09-04	~60
2022-09-06	~60
2022-09-08	~120
2022-09-10	~85
2022-09-12	~135
2022-09-14	~90
2022-09-16	~160
2022-09-18	~10
2022-09-20	~85
2022-09-22	~130
2022-09-24	~10
2022-09-26	~180
2022-09-28	~160
2022-09-30	~85
2022-10-03	~70

Players
Last 30 days

Date	Players
2022-09-04	1
2022-09-06	1
2022-09-08	1
2022-09-10	1
2022-09-12	1
2022-09-14	1
2022-09-16	1
2022-09-18	1
2022-09-20	1
2022-09-22	1
2022-09-24	1
2022-09-26	1
2022-09-28	1
2022-09-30	1
2022-10-03	1

Male Likely male Cannot See Likely female Female

Submit

```
THE FRAMEWORK PROGRAMME FOR RESEARCH AND INNOVATION  
HORIZON 2020  
Using 3%  
History  
ff  
20 shown, 21 deleted  
401.42 kB  
39: GlimmTMB - Temp trends plot on data 1 0, data 11, and others  
32: GlimmTMB - Temp trends per year - on data 10, data 11, and others  
31: GlimmTMB - Temp trends per year - on data 10, data 11, and others  
30: GlimmTMB - Temp trends plot on data 1 0, data 11, and others  
29: GlimmTMB - Temp trends per year - on data 10, data 11, and others  
28: GlimmTMB - Temp trends per year - on data 10, data 11, and others  
27: GlimmTMB - Temp trends plot on data 1 0, data 11, and others  
0 bytes  
format: png, génome de référence: ?  
Method: glimmTMB  
Estimation de la variation annuelle  
glimmTMB(ct~ factor(year)+ (~id_plot)  
Estimation de la tendance  
glimmTMB(ct~ year+ (~id_plot)  
Saving 7 x 7 in image  
Image in png format  
16: GAM - Temp tren
```

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 repo

Dashboard Batches Tasks Project settings Data upload Import Data exports Production / SpiPoll Fly Logout

Overview All time

Batches 6 Active tasks 4.76k Training tasks 400 Average run 2.16 Players 2.00 Classifications 73.59k

Classification Last 30 days

Players Last 30 days

Amazing basis for Cofest « citizen science » webhook?

2022-08-04 2022-08-06 2022-08-08 2022-08-10 2022-08-12 2022-08-14 2022-08-16 2022-08-18 2022-08-20 2022-08-22 2022-08-24 2022-08-26 2022-08-28 2022-08-30 2022-09-01

Male Likely male Cannot See Likely female Female Submit

History Rechercher des données ff 20 shown, 21 deleted 401.42 <5 39: GlimmTMB - Temp trends per year - on data 10, data 11, and others 31: GlimmTMB - Temp trends per year - on data 10, data 11, and others 30: GlimmTMB - Temp trends plot on data 1 0, data 11, and others 29: GlimmTMB - Temp trends per year - on data 10, data 11, and others 28: GlimmTMB - Temp trends per year - on data 10, data 11, and others 27: GlimmTMB - Temp trends plot on data 1 0, data 11, and others 0 bytes format: png, génome de référence: ? Method: glimmTMB Estimation de la variation annuelle glimmTMB(ct~ factor(year)+ (~id_plot) Estimation de la tendance glimmTMB(ct~ year+ (~id_plot) Saving 7 x 7 in image Image in png format 16: GAM - Temp tren

The screenshot shows the Galaxy web interface with various metrics for a project: 6 batches, 4.76k active tasks, 400 training tasks, an average run of 2.16, 2.00 players, and 73.59k classifications. Below these are two bar charts: one for 'Classifications' over the last 30 days and another for 'Players'. At the bottom, there's a user interface for classifying images of hoverflies (syphres), showing three categories: Male, Female, and Cannot See, with a 'Likely male' option highlighted by a green box. To the right, a sidebar displays a history of command-line operations related to GlimmTMB analysis. A large red watermark 'Amazing basis for Cofest « citizen science » webhook?' is overlaid across the center of the interface.

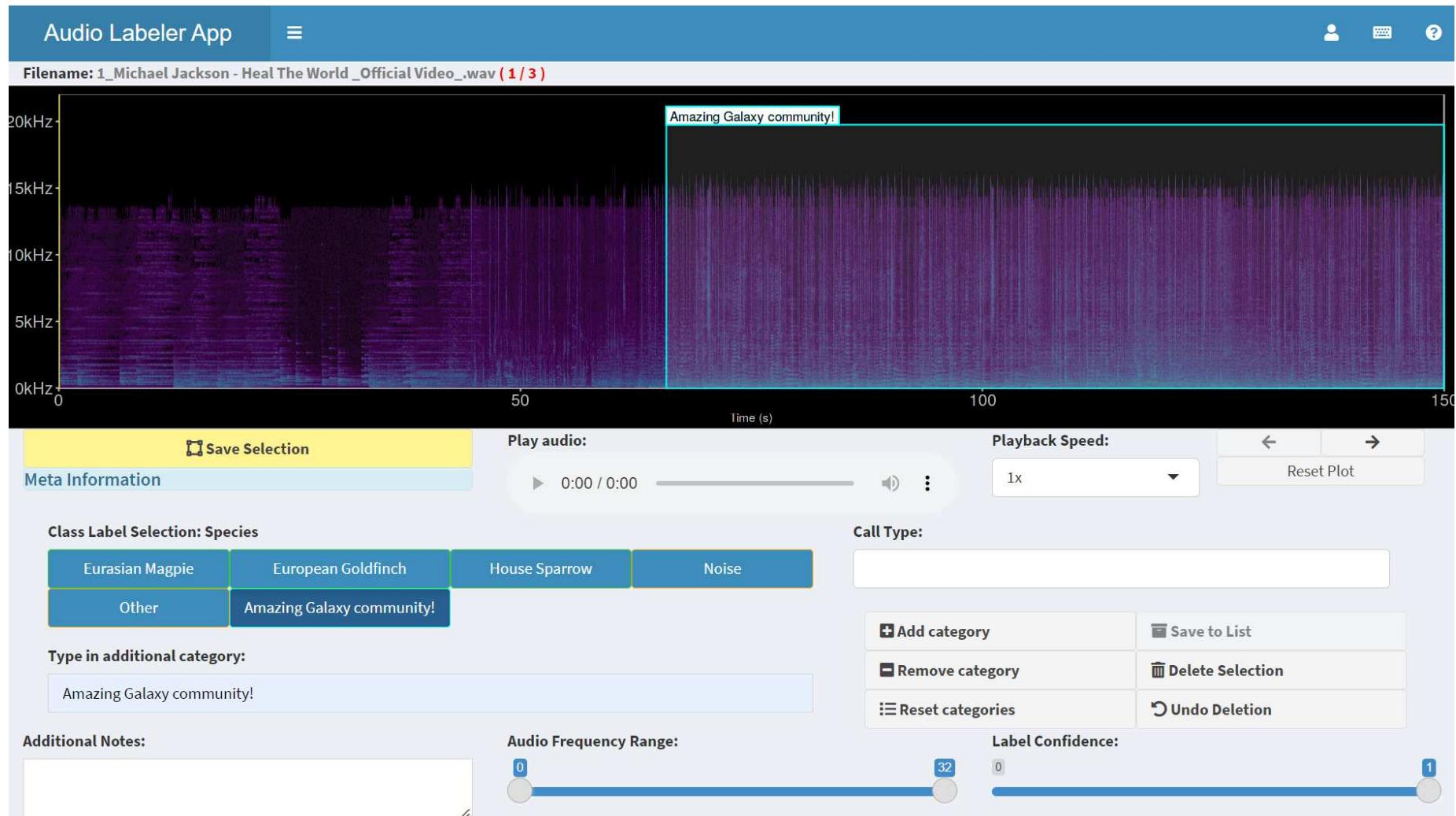
63 320 classifications in 2,5 years !



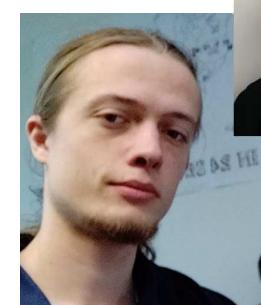
Crowd sourcing through Galaxy

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



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

