

Infrastructure PNDB, ouverture et FAIRisation des données et codes sources utilisés dans les domaines de la Biodiversité

Yvan Le Bras, Elie Arnaud, Coline Royaux,
Julien Sananikone, Marie Jossé, Mélanie
Madelin, Olivier Norvez



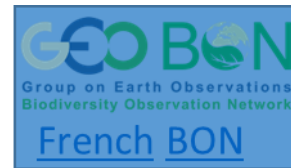
Ouvrir
la science !
RECHERCHE DATA GOUV



CENTRES DE RÉFÉRENCE THÉMATIQUES
Experts disciplinaires de la donnée



Le Fonds national
pour la science
ouverte
OpenMetaPaper



DataONE: More than « just » a network

Through NCEAS, EDI, US LTER & Amazing EML!



PNDB Pôle National de Données de Biodiversité Data Catalog DATA PORTALS SUMMARY ABOUT Jump to: DOIx or II Go SIGN IN

Search

Filter by:

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

Datasets 1 to 25 of 68 Sort by: Most recent

1 2 3 Next

Hide Map >>

Greenland Iceland Sweden Finland Norway Denmark Poland Belarus Ireland United Kingdom Germany Austria Romania Ukraine Portugal Spain Italy Greece Turkey Tunisia Syria Morocco Algeria Libya Egypt Western Sahara Mauritania Mali Niger Chad Sudan Ethiopia Burkina Faso Ghana Nigeria

1 6 28 1 1 1

Satellite Terrain

Google Keyboard shortcuts Map data ©2022 Google, INEGI 1000 km Terms of Use

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-1707-41f0-8fe6-5e13489b2d4e.

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, Bouveret Laurent, Romain David, Eric Foulquier, et al. 2021. **Kakila 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 Française 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.

DataONE: More than « just » a network

Through NCEAS, EDI, US LTER & Amazing EML!



PNDB Pôle National de Données de Biodiversité Data Catalog

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

Search

Filter by:

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

Datasets 1 to 25 of 68 Sort by: Most recent

1 2 3 Next

Hide Map >

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-1707-41f0-8fe6-5e13489b2d4e.

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, Bouveret Laurent, Romain David, Eric Foulquier, et al. 2021. **Kakila 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 Française 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

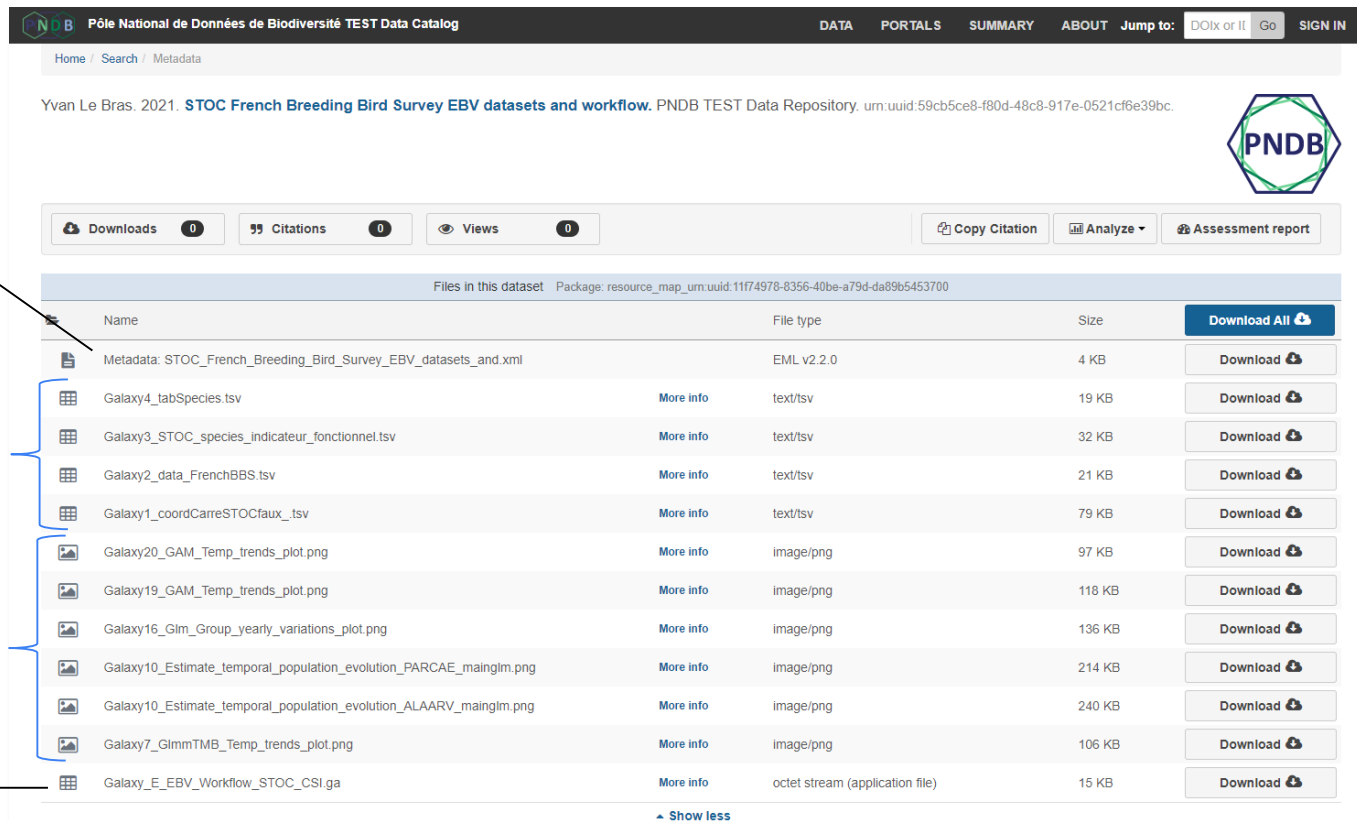
Google Keyboard shortcuts Map data ©2022 Google, INEGI 1000 km Terms of Use

Powered by Metacat (MetacatUI v2.20.0)



DataONE: More than « just » a network

Through NCEAS, EDI, US LTER & Amazing EML!



PNDB Pôle National de Données de Biodiversité TEST Data Catalog

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

Home / Search / Metadata

Yvan Le Bras. 2021. [STOC French Breeding Bird Survey EBV datasets and workflow](#). PNDB TEST Data Repository. urn:uuid:59cb5ce8-f80d-48c8-917e-0521cf6e39bc.

PNDB

Downloads 0 Citations 0 Views 0

Copy Citation Analyze Assessment report

Files in this dataset Package: resource_map_urn:uuid:11f74978-8356-40be-a79d-da89b5453700

Name	File type	Size	Download All
Metadata: STOC_French_Breeding_Bird_Survey_EBV_datasets_and.xml	EML v2.2.0	4 KB	Download
Galaxy4_tabSpecies.tsv	More info text/tsv	19 KB	Download
Galaxy3_STOC_species_indicateur_fonctionnel.tsv	More info text/tsv	32 KB	Download
Galaxy2_data_FrenchBBS.tsv	More info text/tsv	21 KB	Download
Galaxy1_coordCarreSTOCfaux_.tsv	More info text/tsv	79 KB	Download
Galaxy20_GAM_Temp_trends_plot.png	More info image/png	97 KB	Download
Galaxy19_GAM_Temp_trends_plot.png	More info image/png	118 KB	Download
Galaxy16_Glm_Group_yearly_variations_plot.png	More info image/png	136 KB	Download
Galaxy10_Estimate_temporal_population_evolution_PARCAE_mainglm.png	More info image/png	214 KB	Download
Galaxy10_Estimate_temporal_population_evolution_ALAARV_mainglm.png	More info image/png	240 KB	Download
Galaxy7_GlmmTMB_Temp_trends_plot.png	More info image/png	106 KB	Download
Galaxy_E_EBV_Workflow_STOC_CSI.ga	More info octet stream (application file)	15 KB	Download

Show less

EML metadata file

EBV-useable file

EBV-ready file

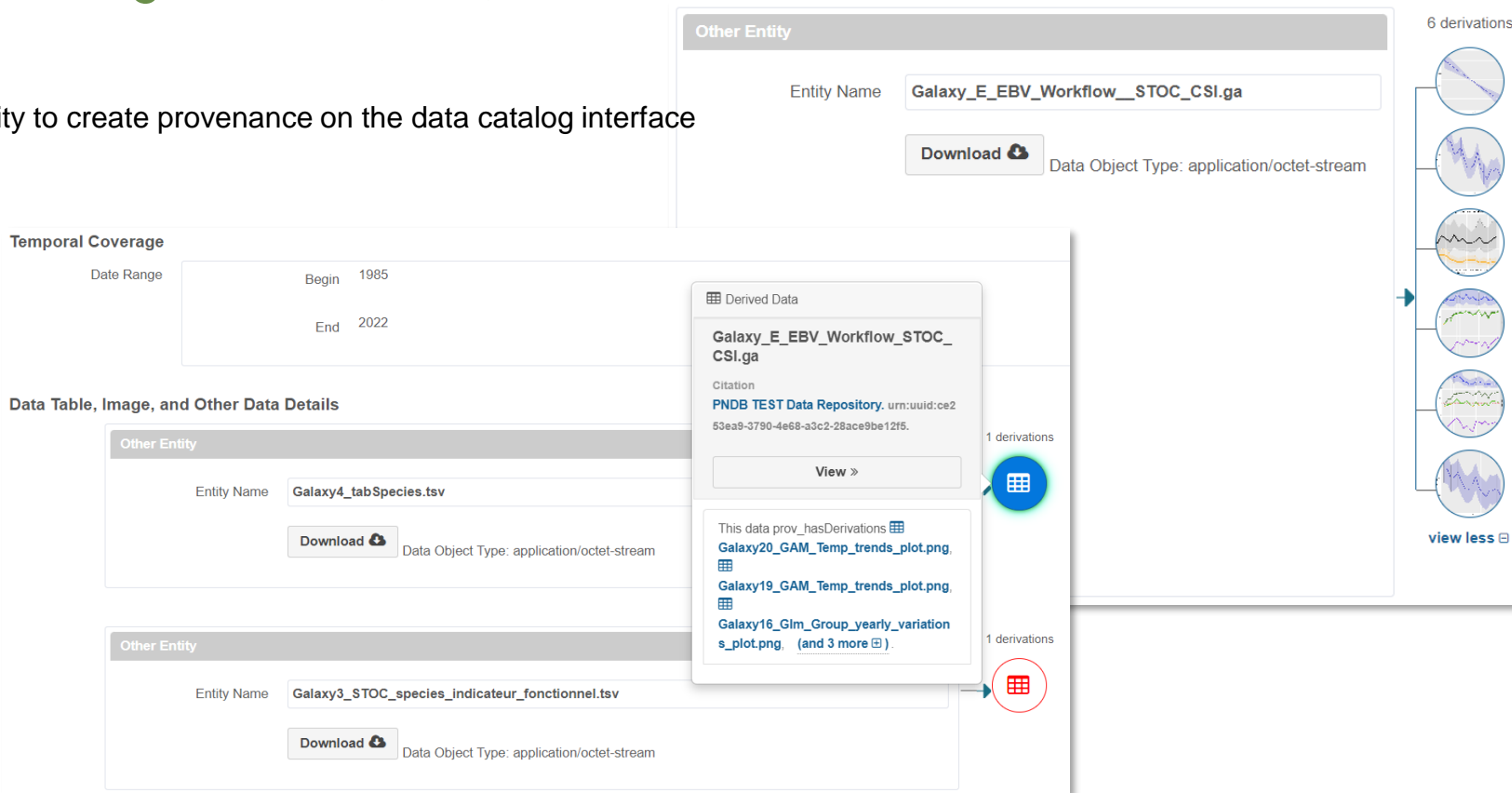
Derived & modeled
EBV file

EBV workflow file

DataONE: More than « just » a network

Through NCEAS, EDI, US LTER & Amazing EML!

Possibility to create provenance on the data catalog interface



The screenshot displays the DataONE interface with several data entities and their associated derivations. The main interface shows three data entities, each with a download button and a data object type of application/octet-stream.

- Entity 1:** Entity Name: Galaxy_E_EBV_Workflow__STOC_CSI.ga. It has 6 derivations, shown as a vertical stack of six circular thumbnails representing different data visualizations.
- Entity 2:** Entity Name: Galaxy4_tabSpecies.tsv. It has 1 derivation, indicated by a blue grid icon.
- Entity 3:** Entity Name: Galaxy3_STOC_species_indicateur_fonctionnel.tsv. It has 1 derivation, indicated by a red grid icon.

A modal window titled "Derived Data" is open, showing the details for the derivation of Galaxy_E_EBV_Workflow__STOC_CSI.ga:

- Derived Data:** Galaxy_E_EBV_Workflow__STOC_CSI.ga
- Citation:** PNDB TEST Data Repository. urn:uuid:ce253ea9-3790-4e68-a3c2-28ace9be12f5.
- View »** button
- This data prov_hasDerivations:** Galaxy20_GAM_Temp_trends_plot.png, Galaxy19_GAM_Temp_trends_plot.png, Galaxy16_Gim_Group_yearly_variation_s_plot.png, (and 3 more ☰).

Additional interface elements include:

- Temporal Coverage:** Date Range: Begin 1985, End 2022.
- Data Table, Image, and Other Data Details:** Section header.
- view less ☰** button at the bottom right of the derivations list.

DataONE: More than « just » a network

Through NCEAS, EDI, US LTER & Amazing EML!



PNDB Pôle National de Données de Biodiversité Data Catalog


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

PNDB Pôle National de Données de Biodiversité Data Catalog

DATA PORTALS SUMMARY ABOUT Jump to: DOIx or If Go YVAN LE BRAS ▾

Home Search Metadata

Clémence Epinoux, Marine Barbarin, Justine Castrec, Carine Churlaud, Mathilde Dabrowski, et al. 2022. **Water Interdisciplinary Biology and Ecology database "WIBE": Towards FAIR, open and interdisciplinary data on biomarkers to monitor the ecological status of coastal waters.** PNDB Data Repository. urn:uuid:99abf52c-b271-4b66-ae50-c504e492bc4c.



Downloads 0 Citations 0 Views 0

Copy Citation Analyze Assessment report Edit Publish with DOI

Files in this dataset Package: resource_map_urn:uuid:7bf52280-5ed6-48e5-b857-a4f9adecb90f

Name	File type	Size	Download All
Metadata: CoastWIBE_database_Coastal_Water_Interdisciplinarity_Biomarkers.xml	EML v2.2.0	183 KB	Download
Field_organisms_samples.csv	text/csv	94 KB	Download
Fieldwork.csv	text/csv	13 KB	Download
Project.csv	text/csv	785 B	Download

Show 5 more items in this data set

General

Annotations

- is about biomarker
- is about marine
- is about aquatic
- is about pollution
- is about corsica
- is about ecotoxicology
- is about interdisciplinary
- is about ecology
- is about biology
- is about FAIR data
- is about open science
- is about database

Identifier Water Interdisciplinary Biology and Ecology database "WIBE"

Abstract The Water Interdisciplinary Biology and Ecology database "WIBE" database presents data from different scientific projects proposing the monitoring of the ecological status of waters by a multi-biomarker study. This database gathers biological data of selected bioindicator species and environmental contextual data. In this work, data on physico-chemical parameters and concentrations of trace elements and organic pollutants were collected in the waters as well as biomarkers of effect and exposure to pollutants in marine organisms. The collected data, after various analyses, were cleaned and reworked to meet FAIR and open data principles. We focused on developing a data dictionary linked to existing ontologies and compliant with standards in order to make it as reusable as possible for the ecotoxicology research community. All datasets are available on the public repository of the National Biodiversity Data Centre. The current dataset can be used by port and coastal water managers but also by marine ecotoxicology researchers who will benefit from the first, to our knowledge, completely open database on marine biomarkers allowing the monitoring of coastal water contamination and thus the proposal of remediation measures if necessary.

https://data.pndb.fr

DataONE: More than « just » a network

Through NCEAS, EDI, US LTER & Amazing EML!



0 sources

0 derivations

Data Table

Entity Name: **Water_samples_measurements.csv**

Description: Water_samples_measurements.csv

Object Name: Water_samples_measurements.csv

Size: 55049 bytes

Authentication: eeafa7538d922594288eb8c210e3ed52 Calculated By MD5

Text Format

Number of Header Lines	1
Record Delimiter	\r\n
Attribute Orientation	column

Simple Text

Field Delimiter	;
-----------------	---

Number Of Records: 133

Attribute Information

Variables

- Water_sample_ID
- Fieldwork_ID
- Water_subsurface_temperature_celsius**
- Water_subsurface_salinity_psu
- Water_subsurface_dissolved_oxygen_mg
- Water_subsurface_pH
- Water_subsurface_turbidity_fmu

Name: Water_subsurface_temperature_celsius

Annotations: is similar to Celsius

Label: Celsius

Attribute: Water_subsurface_temperature_celsius in dataTable

Definition: the sampling station measured in celsius degrees.

Storage Type: similar to Celsius

DataONE: More than « just » a network

Through NCEAS, EDI, US LTER & Amazing EML!



PNDB Pôle National de Données de Biodiversité Data Catalog

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

PNDB Pôle National de Données de Biodiversité Data Catalog

DATA PORTALS SUMMARY ABOUT Jump to: DOIx or IJ Go YVAN LE BRAS

PNDB Pôle National de Données de Biodiversité Data Catalog

DATA PORTALS SUMMARY ABOUT Jump to: DOIx or IJ Go YVAN LE BRAS

0 sources

Data Table

0 derivations

PNDB Pôle National de Données de Biodiversité Data Catalog

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

Metadata Assessment Report

Lorraine Coché, Elie Arnaud, Bouveret Laurent, Romain David, Eric Foulquier, et al. 2021. **Kakila 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.

After running your metadata against our standard set of metadata, data, and congruency checks, we have found the following potential issues. Please assist us in improving the discoverability and reusability of your research data by addressing the issues below.

Assessment suite: FAIR Suite v0.3.1

Findable: 87% complete

Accessible: 57% complete

Interoperable: 90% complete

Reusable: 73% complete

51 checks

Number Of R

Attribute Infor

Variables

- Water_sample_ID
- Fieldwork_ID
- Water_subsurfa
- Water_subsurfa
- Water_subsurfa
- Water_subsurface_pH
- Water_subsurface_turbidity_fnu

Storage Type similar to Celsius

Galaxy for Ecology, an ecosystem!

Galaxy platform



Web platform to share and process research data

Permit easy access to cloud analysis and High Performance Computing through interfacing of any programming language

Four key principles

Accessibility

Reproducibility

Transparency

Peer review

Galaxy for Ecology, an ecosystem!

Guidelines to help create shareable “analytical bricks”

Currently, in ecology ...

One R script for one input datafile

```
direction="wide")
tab[is.na(tab)] <- 0
chemin <- "touverunnon"
chemin <- paste(rep(filename, sep="/")
write.table(tab, chemin)
colNames(tab) <- sub("nombre.", "", colNames(tab))
return(tab)
}
```

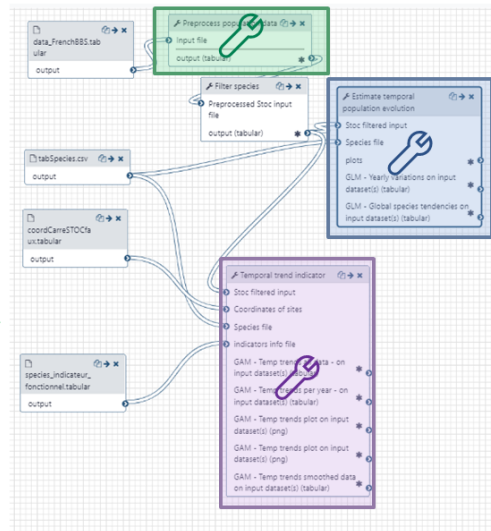
```
## sous jeux de donnees si choix d espece d annee ou d un pourcentage de carres
makeSousTab <- function(tab, vecsp=NULL, echantillon=1,
methodeechantillon="carre", vecannees=NULL) {
cat(" -- Fabrication du sous jeu de donnees --\n")
flush.console()
## reduction de la table à certaines espèces
if(!is.null(vecsp)) {
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 carres sur l'annee
if(echantillon != 1) {
if(echantillon < 1 & echantillon > 0) {
nbnit <- nrow(tab)
if(methodeechantillon == "global") {
nb <- round(nrow(tab)*echantillon)
cat(" echantillonnage", echantillon*100,
"% des donnees par la methode", methodeechantillon, "\n")
cat(" -> conservation de", nb, "lignes sur", nbnit, "\n")
flush.console()
tab <- tab[sample(1:nrow(tab))[1:nb],]
} else {
if(methodeechantillon == "carre") {
cat(" echantillonnage", 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, "\n")
tab=subset(tab, subset = carre %in% chat)
cat(" (" , nrow(tab), " lignes sur ", nbnit, ")", "\n", sep="")
} else {
stop("Methode d echantillonnage non reconnue")
}
```



With Galaxy...

Several atomized R scripts for several input datafiles



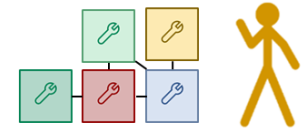
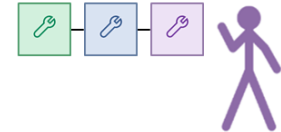
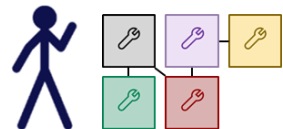
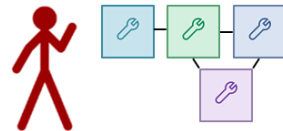
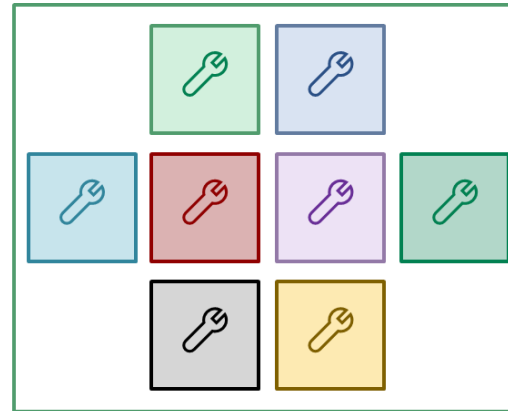
Galaxy for Ecology, an ecosystem!

Guidelines to help create shareable “analytical bricks”

	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

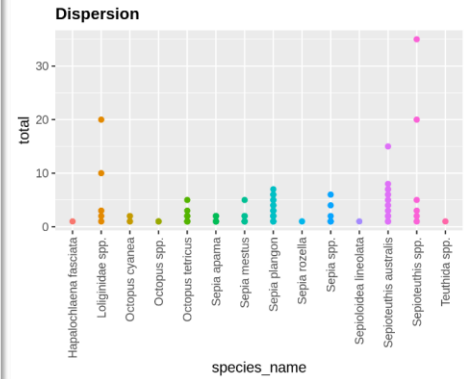
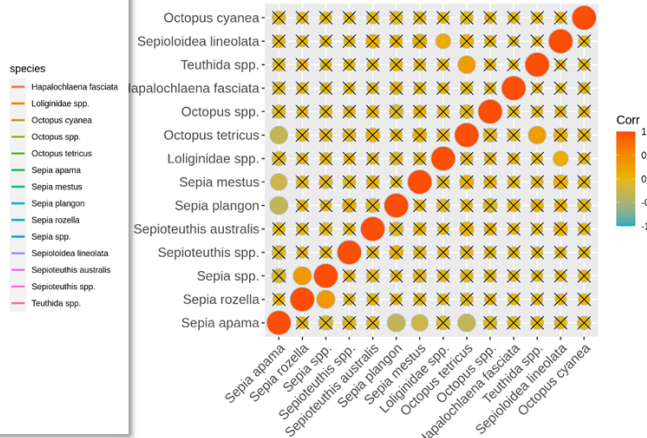
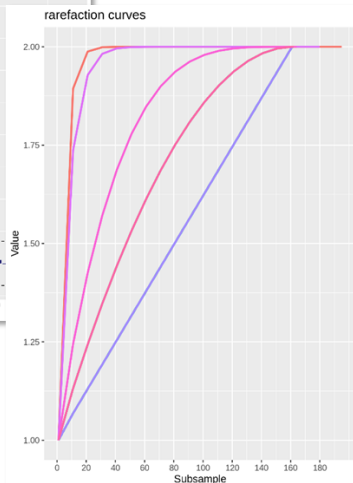
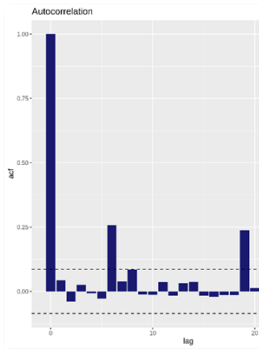
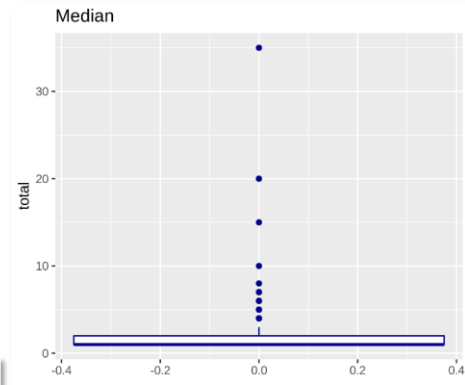
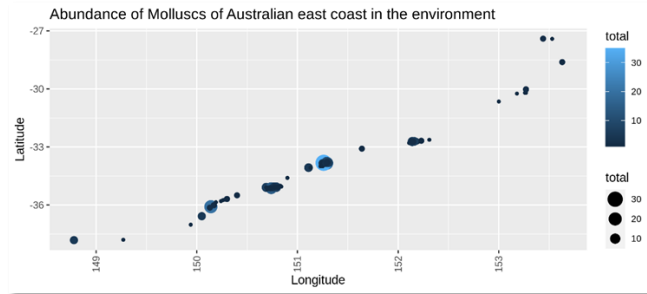
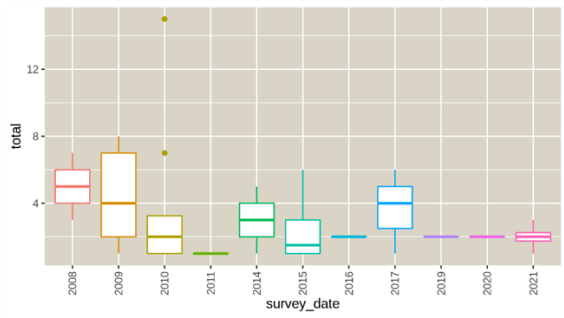


	Unitas	rotation	codeSp	sexe	taille	classe_taille	poids	nb_ind
1	AS140155	3	Hemifasc	-999	-999	P	-999	1
2	AS140159	1	Nasosp.	-999	-999	P	-999	3
3	AS140159	3	Gompvari	-999	-999	P	-999	1
4	AS140160	3	Gompvari	-999	-999	P	-999	1
5	AS140099	2	Parumult	-999	-999	P	-999	1
6	AS140088	1	Varilout	-999	-999	P	-999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

	Survey	Year	Quarter	Area	AphiaID	Species	InstClass	CSI#_number_per_hour
1	BIT5	1991	1	22	126281	Anguilla anguilla	0	0.000000
2	BIT5	1991	1	22	126281	Anguilla anguilla	720	0.009160
3	BIT5	1991	1	22	126417	Clupea harengus	0	0.000000
4	BIT5	1991	1	22	126417	Clupea harengus	80	0.075785
5	BIT5	1991	1	22	126417	Clupea harengus	85	0.088277
6	BIT5	1991	1	22	126417	Clupea harengus	95	0.037892
7	BIT5	1991	1	22	126417	Clupea harengus	100	0.063293
8	BIT5	1991	1	22	126417	Clupea harengus	105	0.032482
9	BIT5	1991	1	22	126417	Clupea harengus	110	0.028257

Galaxy for Ecology, an ecosystem!

Killer workflows! Biodiversity data exploration tools



Galaxy for Ecology, an ecosystem!

Killer workflows! *Dealing with GIS and netcdf files*

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

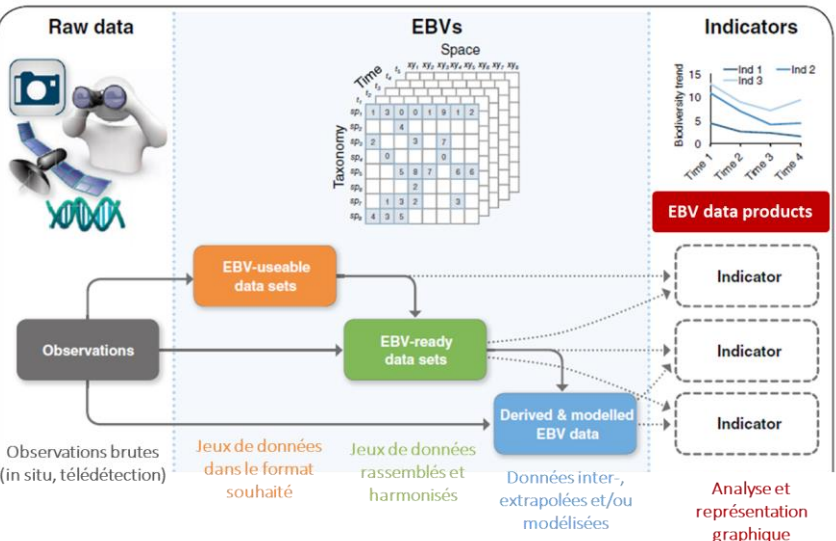
The screenshot shows a Galaxy workflow interface. The title bar reads "Galaxy / Ecology" and "20.05.2022. The storage issue is being resolved." The workflow title is "Shapefile (sample sites) + netcdf (environmental data) - Xarray and GIS data handling and visualization". The workflow consists of several steps: "Upload Dataset (big)", "Gdal supported input file", "NetCDF xarray Selection", "Cut", "Convert tabular to CSV", "geoplotter", "dataset-lib", "netcdf-lib", "NetCDF xarray Selection", "Unique", "Cut", "Add columns", "Cut", "NetCDF xarray map plotting". The right sidebar shows the workflow's name, version (3: May 20th 2022, 15 steps), and a license section.

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.

Galaxy for Ecology, an ecosystem!

Killer workflows! EBV workflows



STOC workflow

"carre"	"annee"	"espece"	"abond"
"440072"	2014	"ALAARV"	4
"440072"	2014	"PARCAE"	5
"440168"	2014	"ALAARV"	1
"440168"	2014	"PARCAE"	10
"440168"	2017	"ALAARV"	2
"440168"	2017	"PARCAE"	12
"440168"	2018	"ALAARV"	2
"440168"	2018	"PARCAE"	20
"440211"	2014	"ALAARV"	10
"440211"	2014	"PARCAE"	5
"440211"	2015	"ALAARV"	11

STOC data

- Site
- Year
- Species
- Occurrence



PAMPA workflow

STOC data

"carre"	"annee"	"espece"	"abond"
"440072"	2014	"ALAARV"	4
"440072"	2014	"PARCAE"	5
"440168"	2014	"ALAARV"	1
"440168"	2014	"PARCAE"	10

ode	Site	SiteLat	SiteLong	SurveyDate	Depth	Taxon	Block	Total	Dn
22	Ujung Tunku Nth	5.88	95.25	2009-02-28T13:00:00Z	5	Echinostrephus aciculatus	1	100	RS
22	Ujung Tunku Nth	5.88	95.25	2009-02-28T13:00:00Z	5	Echinostrephus aciculatus	2	100	RS
22	Ujung Tunku Nth	5.88	95.25	2009-02-28T13:00:00Z	5	Echinostrephus aciculatus	1	5	RS
22	Ujung Tunku Nth	5.88	95.25	2009-02-28T13:00:00Z	5	Echinostrephus aciculatus	2	2	RS
22	Ujung Tunku Nth	5.88	95.25	2009-02-28T13:00:00Z	5	Echinostrephus aciculatus	1	1	RS
22	Ujung Tunku Nth	5.88	95.25	2009-02-28T13:00:00Z	5	Echinostrephus aciculatus	2	1	RS

Reef Life Survey data

Survey	Year	Quarter	Area	AphiaID	Species	LngtClass	CPUE_number_per_hour
EVH0E	1997	4	Cn	126436	Gadus morhua	0	0.000000
EVH0E	1997	4	Cn	126436	Gadus morhua	300	0.333500
EVH0E	1997	4	Cn	126436	Gadus morhua	350	0.333500
EVH0E	1997	4	Cn	126436	Gadus morhua	460	0.333500
EVH0E	1997	4	Cn	126436	Gadus morhua	820	0.333500
EVH0E	1997	4	Cn	126436	Gadus morhua	870	0.333500
EVH0E	1997	4	Cn	126436	Gadus morhua	910	0.333500

DATRAS ICES data

Kissling et al. 2017

Galaxy for Ecology, an ecosystem!

Killer workflows! *EBV* workflows



1

Biodiversity data

Preprocess data

2

Filter data



Community

Analyze community indexes

3

Analyze species abundance



Species - population

Preprocess population data for evolution trend analysis (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

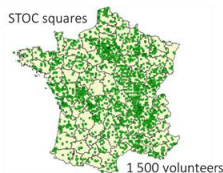
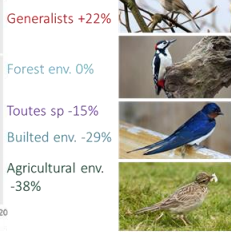
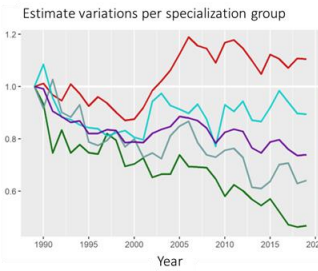
No tabular dataset available.

Output file from the "Preprocess population data tool"

Execute

STOC Filter species with rare and low abundances

STOC Temporal population trend indicator



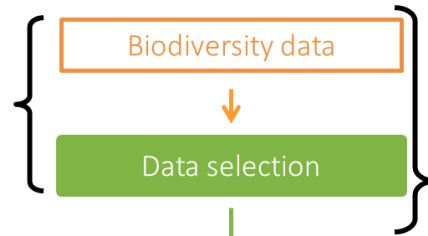
STOC Estimate species population evolution



Galaxy for Ecology, an ecosystem!

Killer workflows! *EBV* workflows

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



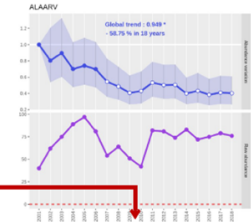
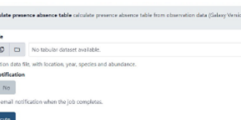
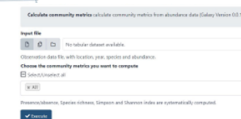
1
Compute community metrics

Compute population metrics

2
GLM on community metrics

GLM on population metrics



3
Time-series plot from GLM results



Galaxy for Ecology, an ecosystem!

Training material following Carpentries practices

Compute and analyze Essential Biodiversity Variables with PAMPA toolsuite

By:  Coline Royaux,  Yvan Le Bras

Overview

Questions

- How to evaluate properly species populations and communities biological state with abundance data?
- How does trawl exploited populations of Baltic sea, Southern Atlantic and Scotland are doing over time?
- How to compute and analyze Essential Biodiversity Variables (EBV) on abundance data?

Objectives

- Upload data from DATRAS portal of ICES
- Pre-process population data with Galaxy
- Learning how to use an Essential Biodiversity Variables (EBV) scientific workflow from raw data to graphical representation
- Learning how to construct a Generalized Linear (Mixed) Model from a usual ecological question
- Learning how to interpret a Generalized Linear (Mixed) Model

Requirements

- [Introduction to Galaxy Analyses](#)

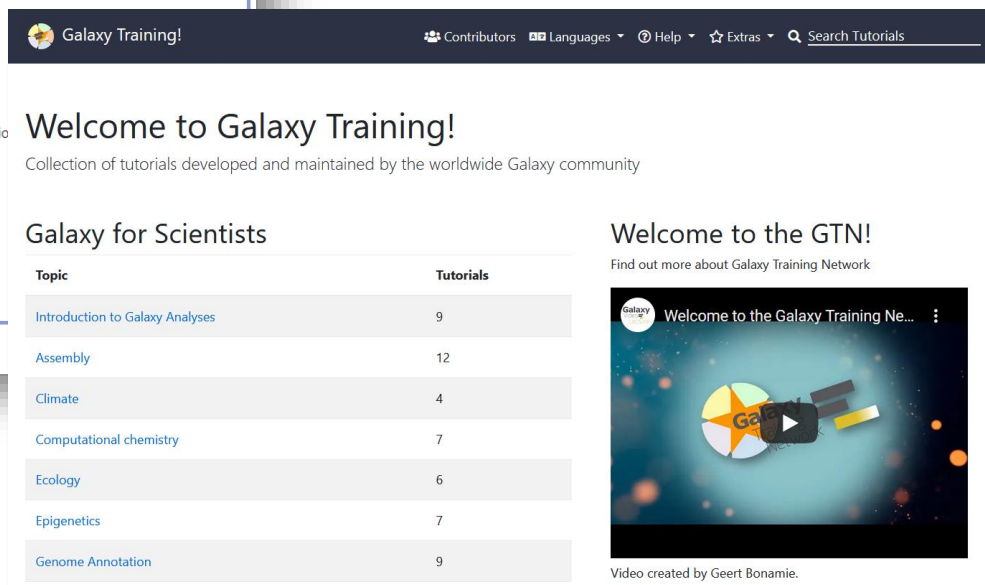
Time estimation: 2 hours

Supporting Materials

[Datasets](#) [Workflows](#) [Available on these Galaxies](#)

Last modification: Dec 1, 2020

<https://bit.ly/2VjDTaQ>



Galaxy Training! Contributors Languages Help Extras Search Tutorials

Welcome to Galaxy Training!


Collection of tutorials developed and maintained by the worldwide Galaxy community

Galaxy for Scientists

Topic	Tutorials
Introduction to Galaxy Analyses	9
Assembly	12
Climate	4
Computational chemistry	7
Ecology	6
Epigenetics	7
Genome Annotation	9

Welcome to the GTN!

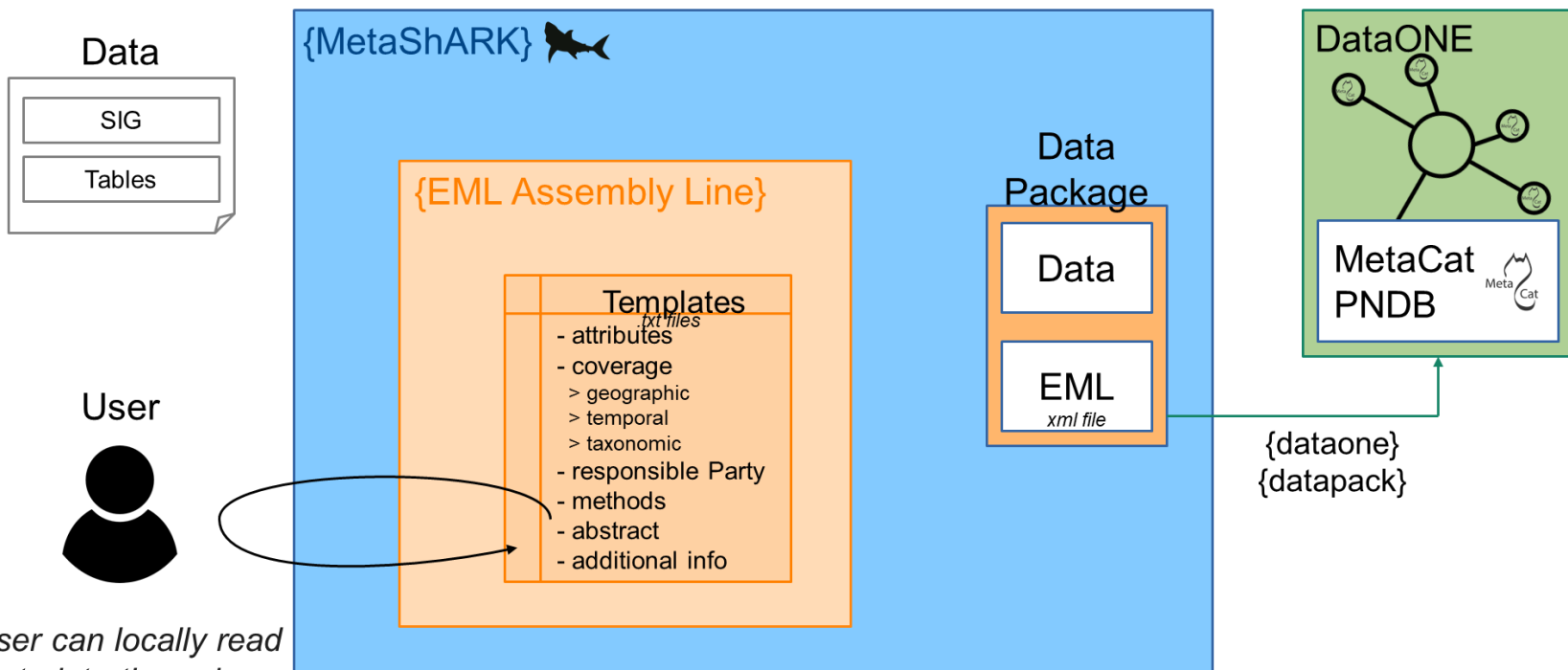
Find out more about Galaxy Training Network



Video created by Geert Bonamie.

MetaShARK application

The role of MetaShARK is to get rid of coding in treatment



User can locally read metadata through metadata table files by downloading them

MetaShARK application

The MetaShARK process is based on EML Assembly Line's one

Screenshots

{EML Assembly Line}

template_core_metadata()

template_table_attributes()

template_categorical_variables()

template_geographic_coverage()

template_taxonomic_coverage()

No specific templating operation

make_eml()

{MetaShARK} 

Attributes step

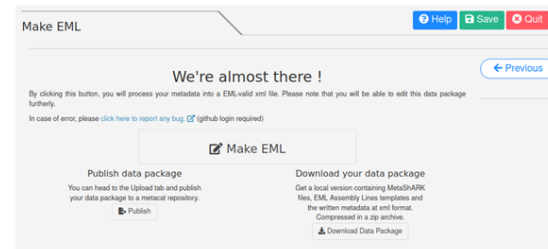
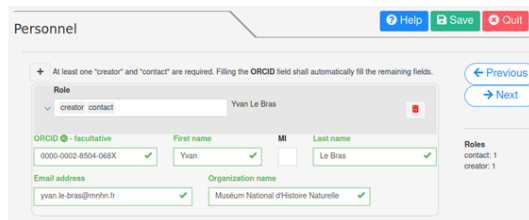
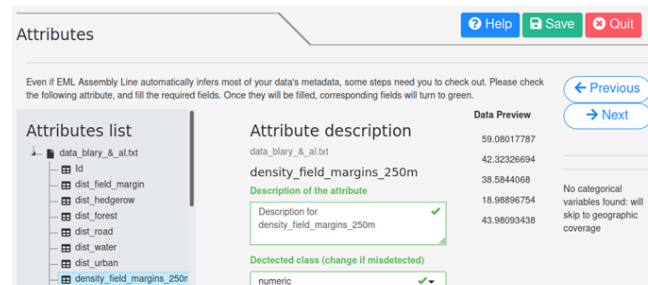
Custom Units step

Geographic Coverage step

Taxonomic Coverage step

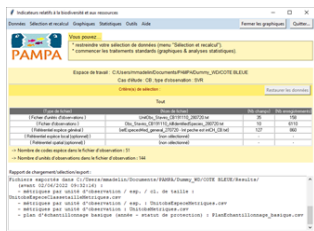
Personnel step
Miscellaneous step

Make EML step

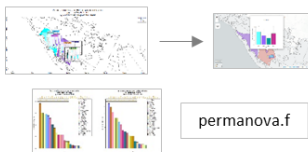
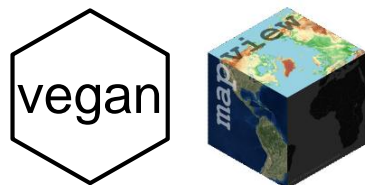


PAMPA application

From a Tcl/tk « old » app to a R package & R Shiny app



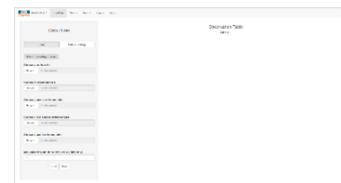
PAMPA
Interface



Interface update



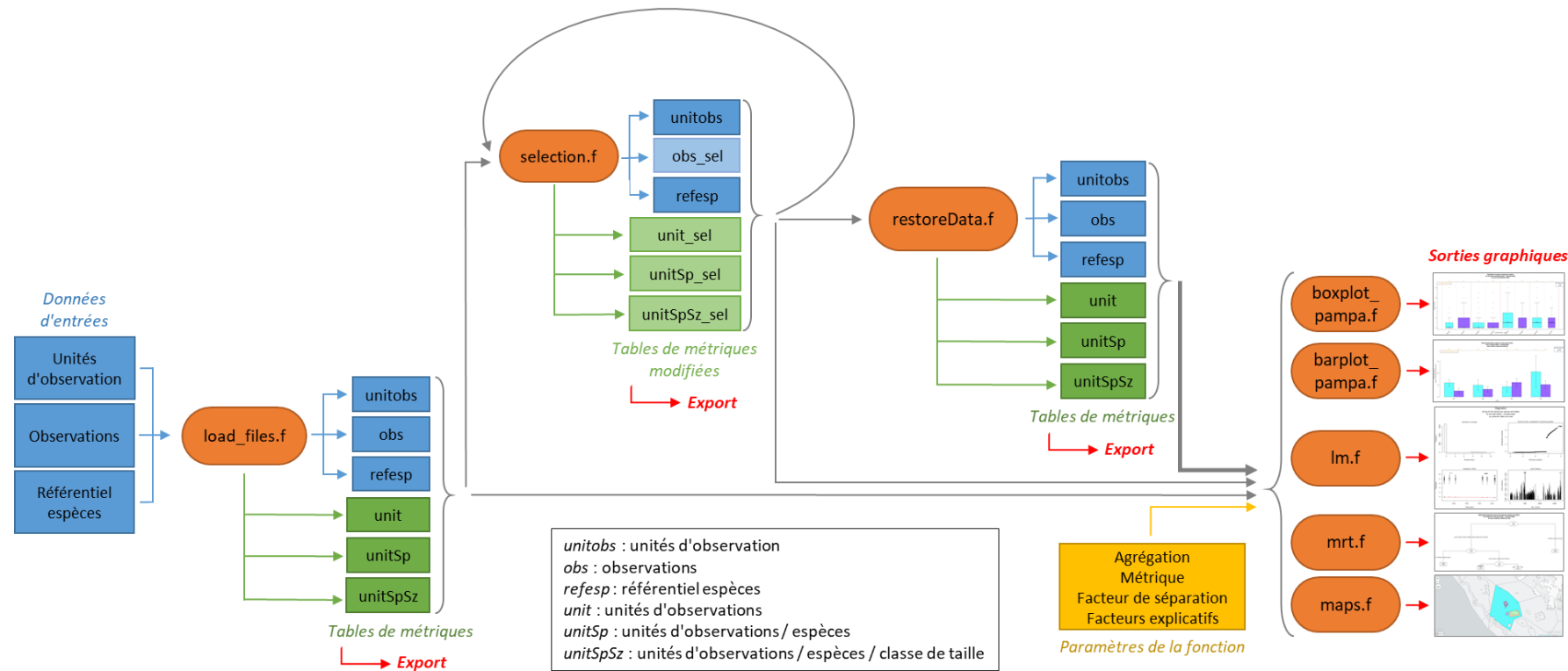
R package
creation



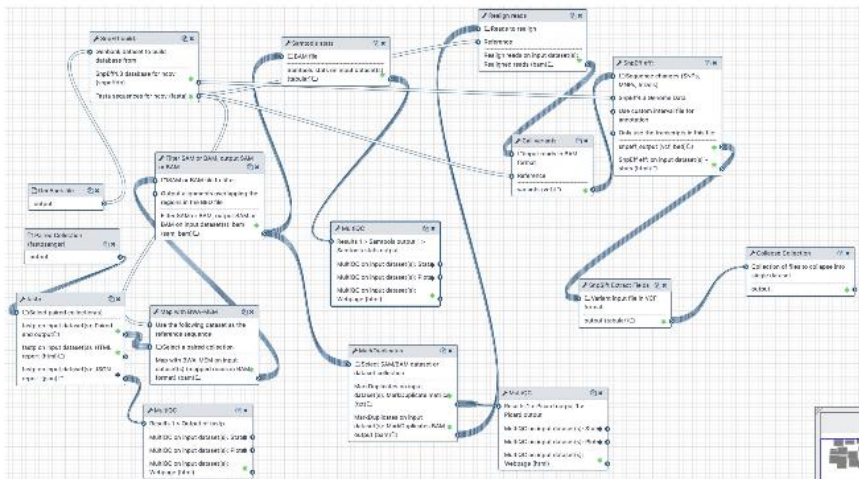
R Shiny®
Application

PAMPA application

PAMPA workflow



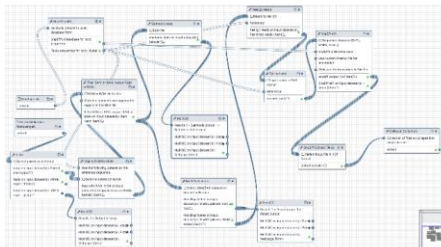
Galaxy-E: A gateway to synchronous as asynchronous workflows (the async ones)



```
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
```


Galaxy-E: A gateway to synchronous as asynchronous workflows

workflows (the async ones)



workflows (the async ones)

The screenshot shows a JupyterLab interface with a Python script in the 'Code' editor and its output in the 'Running' tab. The script defines a function to solve the Lorenz system and plots the trajectories. The output shows the Lorenz attractor plot and the parameter sliders.

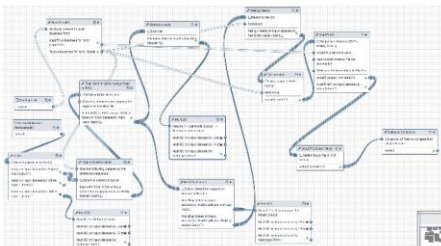
```
File Edit View Run Kernel Tabs Settings Help
Lorenz.ipynb Python 3
We explore the Lorenz system of differential equations:

$$\begin{aligned} \dot{x} &= \sigma(y - x) \\ \dot{y} &= \rho x - y - xz \\ \dot{z} &= -\beta z + xy \end{aligned}$$

Let's change  $(\sigma, \beta, \rho)$  with ipynwidgets and examine the trajectories.
In [2]: from lorenz import solve_lorenz
w=interactive(solve_lorenz, sigma=(0.0,50.0), rho=(0.0,50.0))
w
sigma 10.00
beta 2.67
rho 28.00
def solve_lorenz(sigma=10.0, beta=8./3, rho=28.0):
    """Plot a solution to the Lorenz differential equations."""
    max_time = 4.0
    N = 30
    fig = plt.figure()
    ax = fig.add_axes([0, 0, 1, 1], projection='3d')
    ax.axis('off')
    # prepare the axes limits
    ax.set_xlim((-25, 25))
    ax.set_ylim((-35, 35))
    ax.set_zlim((5, 55))
    def lorenz_deriv(x,y,z, t0, sigma=sigma, beta=beta, rho=rho):
        """Compute the time-derivative of a Lorenz system."""
        x, y, z = x,y,z
        return [sigma * (y - x), x * (rho - z) - y, x * y - beta * z]
    # Choose random starting points, uniformly distributed from -15 to 15
    np.random.seed(1)
    x0 = -15 + 30 * np.random.random((N, 3))
    # Solve for the trajectories
    t = np.linspace(0, max_time, int(250*max_time))
    x,t = np.asarray([integrate.odeint(lorenz_deriv, x0[i], t)
        for x0i in x0])
    # choose a different color for each trajectory
    colors = plt.cm.viridis(np.linspace(0, 1, N))
    for i in range(N):
        x, y, z = x,t[1:,i].T
        lines = ax.plot(x, y, z, '-', c=colors[i])
        plt.setp(lines, linewidth=2)
    angle = 104
    ax.view_init(30, angle)
```

Galaxy-E: A gateway to synchronous as asynchronous 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:
input: the_input
```

workflows (the async ones)

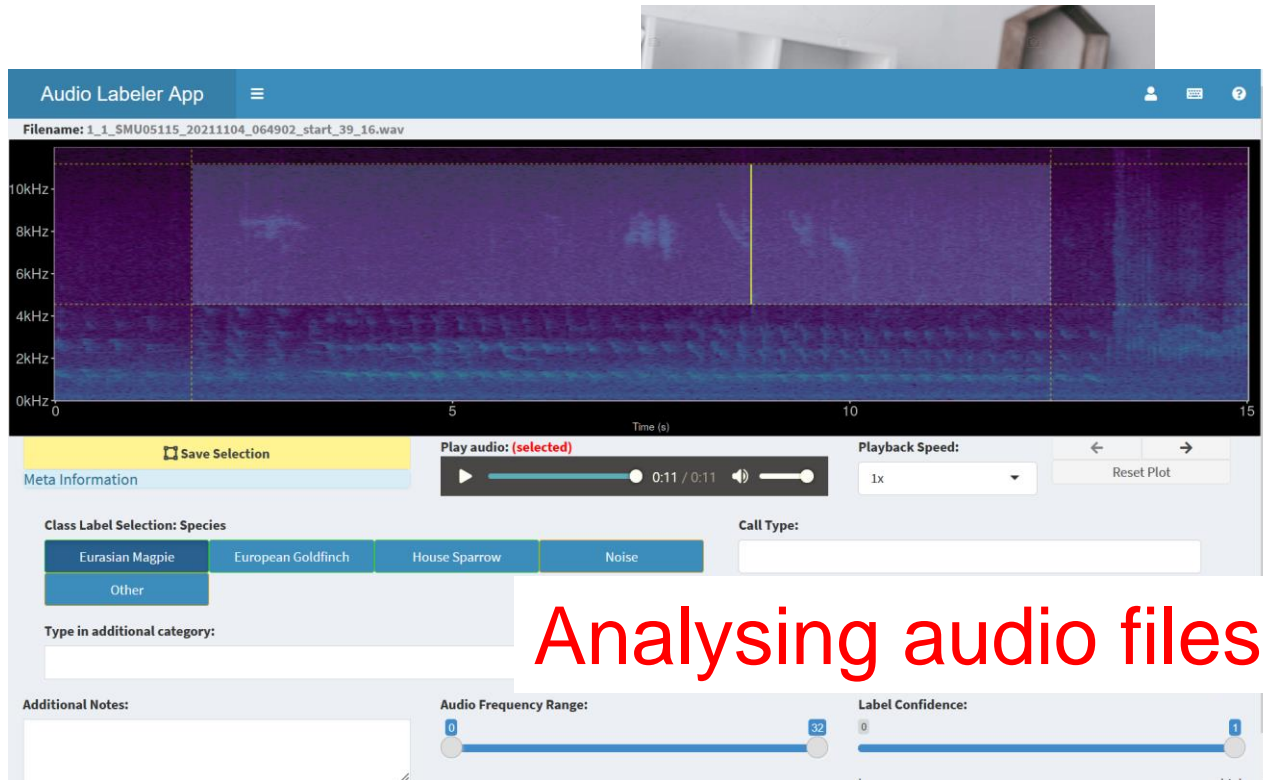
Galaxy-E: A gateway to synchronous as asynchronous workflows

What this guy is doing?



Galaxy-E: A gateway to synchronous as asynchronous workflows

What this guy is doing?



Analysing audio files on Galaxy!

Galaxy-E: A gateway to synchronous as asynchronous workflows

Not only for scientists!

MOODA concept (Massively Open Online Data Analysis)

Crowdsourcing with hoverflies (syphtres) images from SPIPOLL project



GAPARS project

SPIPOLL

The screenshot shows the Galaxy-E web interface. The main content area displays a 'Citizen Science Project' for identifying the sex of marmalade hoverflies. A large image of a hoverfly on a purple flower is shown. Below the image are four classification options: 'Male', 'Likely male' (highlighted with a green box), 'Cannot See', and 'Likely female'. A 'Submit' button is at the bottom. The right sidebar shows a 'History' panel with a list of previous tasks, including 'GlimmTMB - Temp trends per year' and 'GlimmTMB - Temp trends plot on data 1'. The top navigation bar includes 'Analyse de données', 'Workflow', 'Visualisation', 'Données partagées', 'Aide', and 'Utilisateur'.

63 320 classifications in 2,5 years !

Galaxy-E: A gateway to synchronous as asynchronous workflows

Not only for scientists!

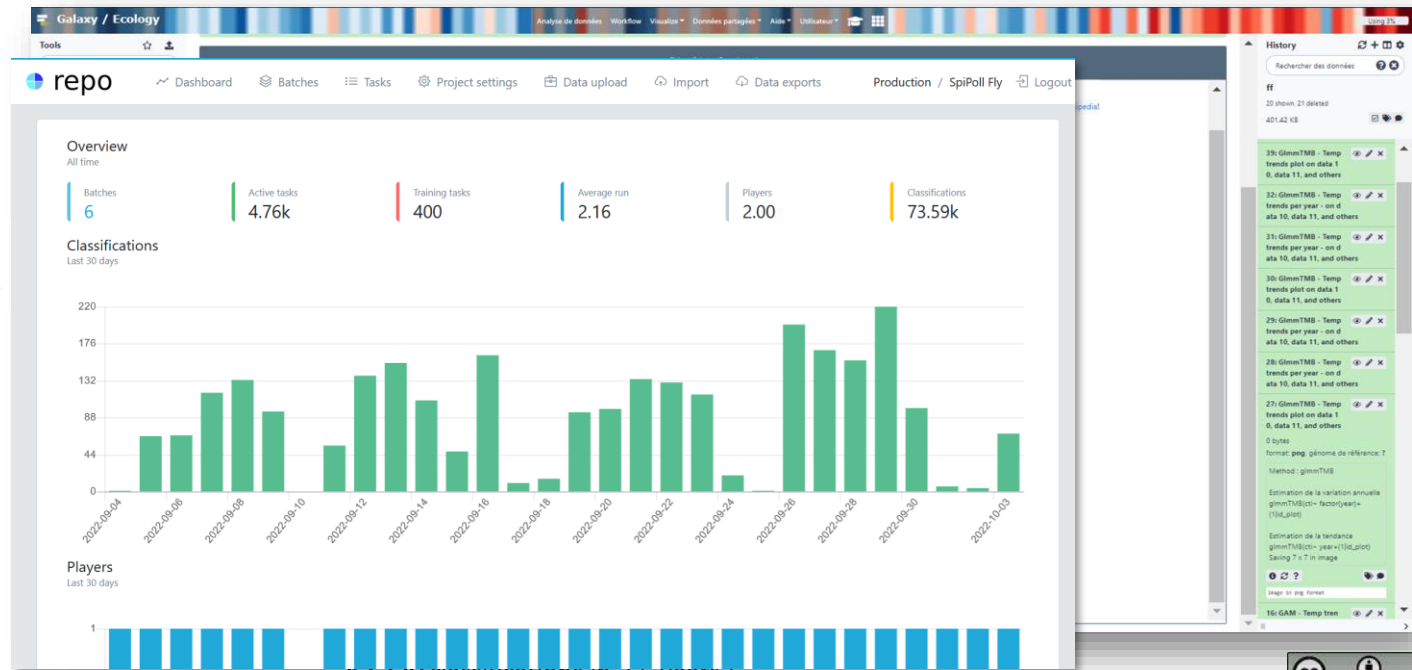
MOODA concept (Massively Open Online Data Analysis)

Crowdsourcing with hoverflies (syphtres) images from SPIPOLL project



GAPARS project

SPIPOLL



33 320 classifications in 2,5 years !

Galaxy-E: A gateway to synchronous as asynchronous workflows

Not only for scientists!

Galaxy for pupils !

bricks.vigienature-ecole.fr

VIGIENATURE
École

Galaxy BRICKS

Le projet Analyser les données Visualiser les données Exemples de scenarios Forum

Bienvenue dans Galaxy Bricks !

Commencer une question de recherche !

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 B)

Aperçu des premières lignes du fichier:

Warning message:
NAs introduits lors de la conversion automatique
Erreur : Must use either a variable name or expression when facetting
Exécution arrêtée

6 Représenter (sur Résumer des données on data 3)

Acknowledgements

- ◆ Royaux Coline – PhD student
- ◆ Arnaud Elie – R Shiny / knowledge
- ◆ Jossé Marie– R / Galaxy dev
- ◆ Madelin Mélanie - R Shiny
- ◆ Sananikone Julien – DevOps / sys admin / web dev
- ◆ Norvez Olivier – animation coordinator
- ◆ Le Bras Yvan – Beta tester

<https://www.pndb.fr/>



Pôle National de Données de Biodiversité

