

Infrastructure de recherche PNDB

(Pôle National de Données de Biodiversité)

Reflexion données/métadonnées

Atelier ODATIS 26 05 2021



Descriptif

Connaissances relatives à la biodiversité (état, dynamiques, interactions, retro-actions...)
Communautés scientifiques très hétérogènes, peu structurées => données dispersées, hétérogènes

Partenaires : 9 organismes, 10 Universités, FRB
 Portage par le MNHN

Objectifs

Développer, offrir et mettre en synergie :

- outils et services
- visant description, mise à disposition, validation, analyse et réutilisation des données de biodiversité,

Fonder un cadre scientifique intégratif : temps long, tous milieux, de la molécule aux anthroposystèmes ; pressions anthropiques)

Compléter le dispositif Syst. Terre-Environnement, en lien étroit avec le SIB national porté par OFB/MTE



Science ouverte

Stratégie

S'appuyer sur l'existant et favoriser :

- pérennité, interopérabilité, ouverture, réutilisation
- croisement de données biodiv / autres (climat, env...)
 (sans se substituer aux porteurs de B de données)

Développer la description des jeux de données
 (métadonnées, « quality data »)

Faire monter les communautés en compétence

S'inscrire dans le paysage complexe des infras
 au niveau national, européen, international



FAIR

Services

Accès aux jeux de métadonnées-données

Services et outils associés

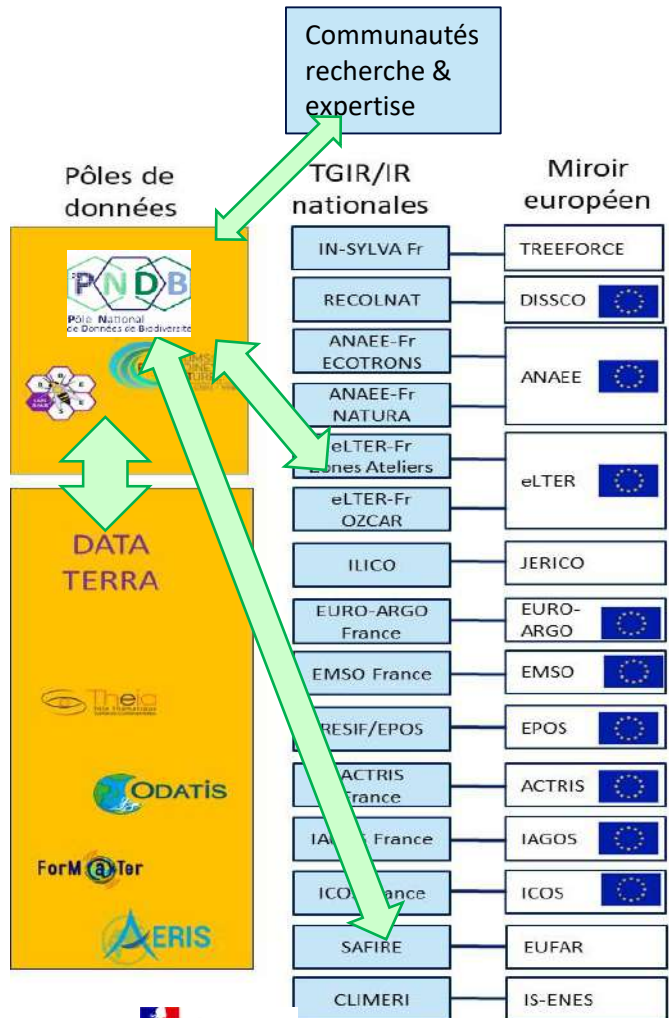
Aide à la bancarisation

Animation des communautés autour de Sci. ouverte

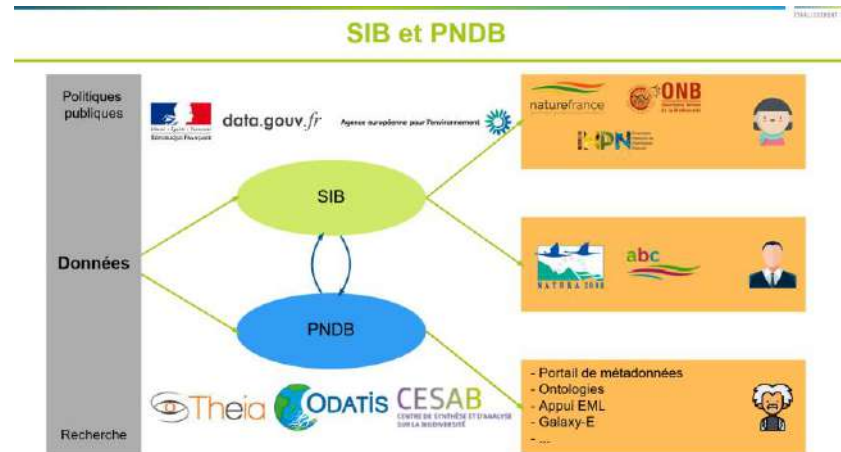
Stimulation des interactions producteurs/utilisateurs

Positionnement à l'échelle nationale

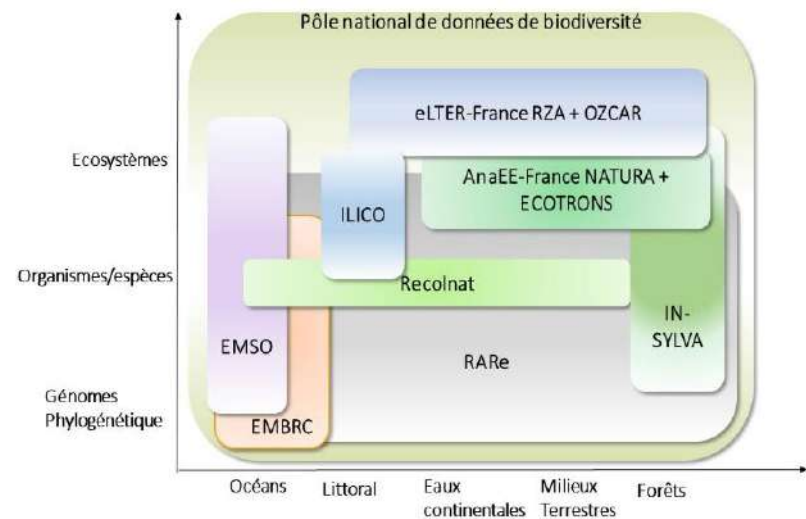
Par rapport aux producteurs de données



Par rapport aux utilisateurs de données



Par rapport aux milieu et aux niveaux biologiques



Services et données

- **Type(s)**
 - Structuration/standardisation de la donnée
 - Avec inférence de métadonnées
 - Métadonnées très fines/détaillées
 - Accès direct à des données ouvertes brutes
 - Analyse/couplage de données hétérogènes sur infrastructure cloud / HPC / distribuée via plateforme ET outils open sources et à fort degré « FAIR »
- **Gestion des données**
 - Conseils sur entrepôts de données les plus pertinents
 - Approche « tout est métadonnée »
 - maDMP (projet Equipex+)
 - Data paper (projet FNSO 2019)
- **Éléments d'ouverture**
 - TOUTES métadonnées et données ouvertes (uniquement licence ouverte Etalab v2 compatible CC-BY 4.0)
 - TOUS les scripts/outils/plateformes sous licence ouverte (GNU GPL/MIT/CECILL, ...)
 - 160 codes ouverts pour les outils « [Galaxy Ecology](#) » (PNDB)
 - Plus de 8000 accessibles via l' « [app store](#) » Galaxy



Données et métadonnées du PNDB

Positionnement PNDB

332

WILLIAM K. MICHENER ET AL.

Ecological Applications
Vol. 7, No. 1

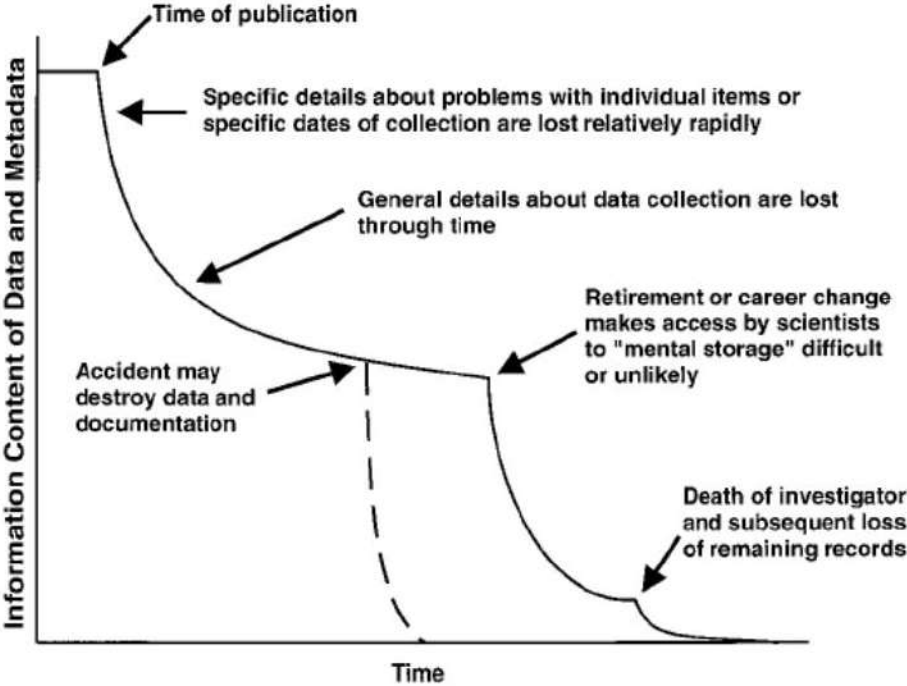


FIG. 1. Example of the normal degradation in information content associated with data and metadata over time ("information entropy"). Accidents or changes in storage technology (dashed line) may eliminate access to remaining raw data and metadata at any time.

Données et métadonnées du PNDB

Positionnement PNDB

Data Table, Image, and Other Data Details

4 sources

Data Table

Entity Name: **Total_Aromatic_Alkanes_PWS.csv**

[Download](#)

Description: Combined dataset from PAH, Alkane and Sample tables documenting samples collected after the Exxon Valdez oil spill in Prince William Sound, AK

Object Name: Total_Aromatic_Alkanes_PWS.csv

Online Distribution Info: <https://cn.dataone.org/cn/v2/resolve/urn:uuid:44108e76-406d-4d68-b1b3-fb4b05e3fff9>

Size: 2601033 byte

Text Format:

- Number of Header Lines: 1
- Record Delimiter: #x0A
- Attribute Orientation: column
- Simple Text
- Field Delimiter: .

Number Of Records: 12142

2 derivations

Positionnement
PNDB

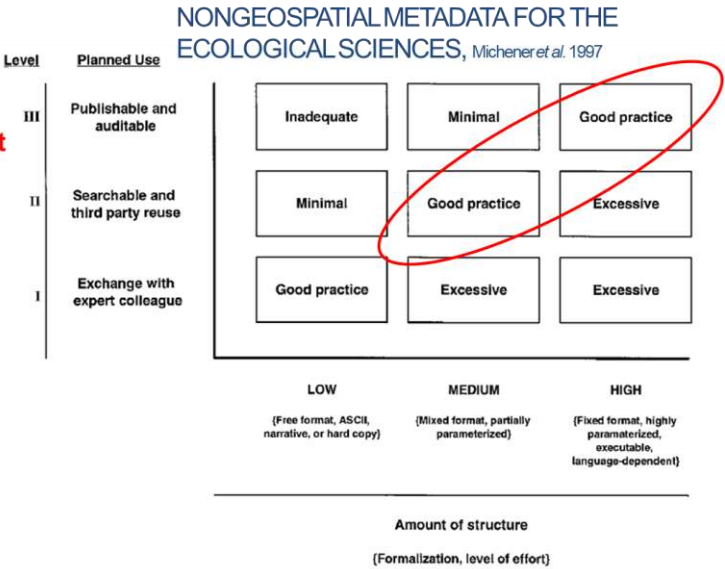


FIG. 3. Degree of metadata format/structure sufficient for three levels of projected secondary data utilization.

Données et métadonnées du PNDB

Le paysage (méta)données via l' Ecological Metadata Language

**Mode
single**

Données



Métadonnées

MetaSHARK

<https://github.com/earnaud/MetaShARK-v2>



Datapackage

Metacat

GeoNetwork

SI externes

Infrastructures

- organismes
Dataverse

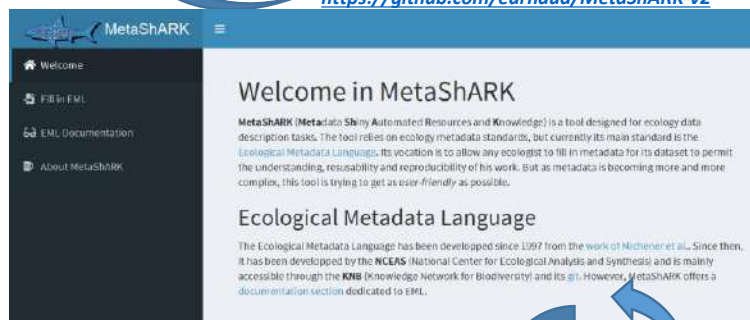
dataset1

dataset2

script1

article1

protocol1



EML Assembly Line

<https://ediorg.github.io/EMLassembleline/index.html>

EML



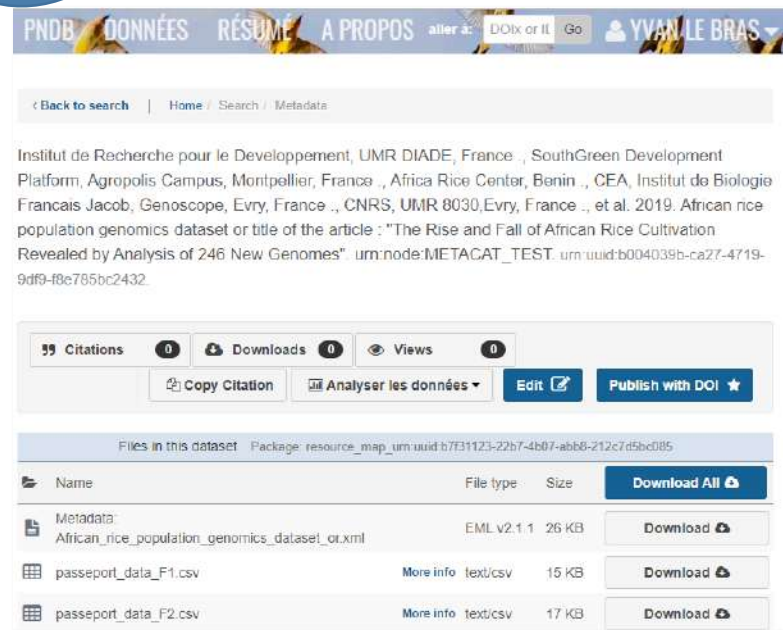
Morpho->Outdated!

**Semantics
enrichment**

cedarr

CEDAR R package for API linking in an R interface.

<https://github.com/earnaud/cedarr>



Portail de données métadonnées :

<https://openstack-192-168-100-101.genouest.org/metacatui>

<http://data.test.pndb.fr/data> <https://data.pndb.fr/data>



Données et métadonnées du PNDB

Le paysage (méta)données via l' Ecological Metadata Language

Mode batch

Données + Métadonnées enrichies unifiées

- dataset1
- dataset2
- script1
- article1
- protocol1



MetaSNAKE

SI externes
- Infrastructures
- organismes

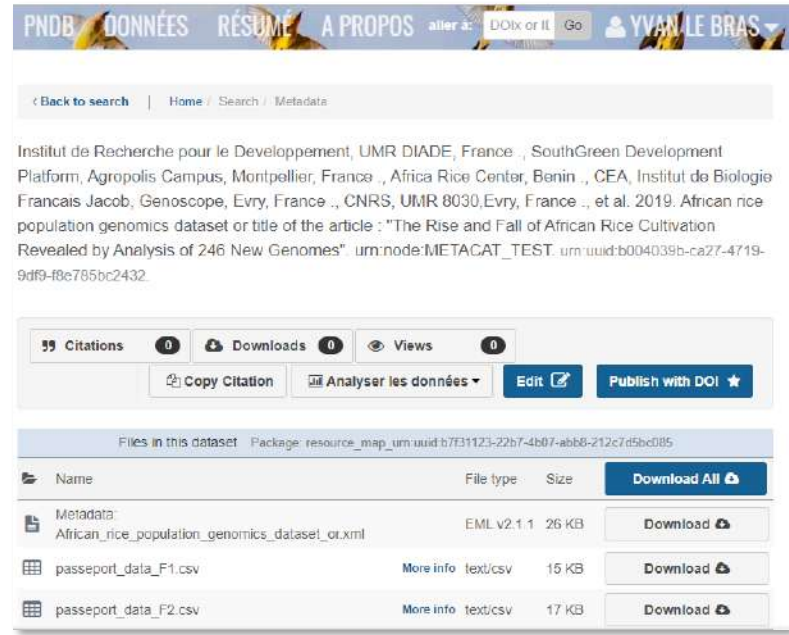
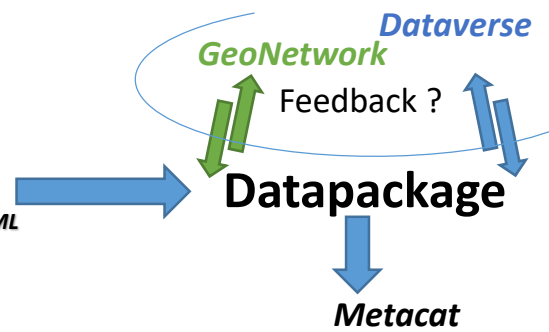
Dataverse
GeoNetwork
Solution à façon

Dublin core,
ISO19115,
...

Métadonnées initiales hétérogènes

Primary variables description
Semantic enrichment

API MetaSHARK ?
<https://github.com/earnaud/MetaShARK-v2>



Portail de données métadonnées :
<https://openstack-192-168-100-101.genouest.org/metacatui>
<http://data.test.pndb.fr/data> <https://data.pndb.fr/data>

Standardisation par la métadonnée vs par un standard de données

- **Pour faciliter l'accès aux données brutes par tous types d'acteurs**
 - Les participants aux atlas de la biodiversité communale n'ont pas accès à la donnée, ou que une partie filtrée après passage par SI d'association puis SI politique publique
- **Pour faciliter le calcul d'indicateurs de Biodiversité**
 - On ne peut pas se contenter de données d'occurrence ou de présence seule pour les calculs d'indicateurs à l'échelle des communautés ou écosystèmes
- **Problème du fait que un standard de données peut obliger à contraindre / modifier le sens d'une variable primaire**
 - Pour certaines données de recherche notamment, il faut rentrer dans le cadre même si la correspondance est pas parfaite

Volet analyse

Le paysage **analyse** via *Github*, *Conda*, *Containers*, *Cloud* et *Galaxy*



Galaxy / Ecology

Analyses de données Workflow Visualize Données partagées Aide Utilisateur

Using 84%

Tools

search tools

Compute GLM on community data
Compute a GLM of your choice on community data

Compute GLM on population data
Compute a GLM of your choice on population data

Calculate presence absence table
calculate presence absence table from observation data

Calculate community metrics
calculate community metrics from abundance data

Create a plot from GLM data as temporal trend

Temporal trend indicator using GlimmTMB or GAM models

Estimate temporal population evolution by specialization group

Estimate temporal population evolution by species

Filter species with rare and low abundances

Model temporal trend with a simple linear regression

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

Favorite Versions Options

Yearly variation dataset

16: GLM - Results from your population analysis on data 1...

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

Global tendencies dataset

16: GLM - Results from your population analysis on data 1...

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

Species file

16: GLM - Results from your population analysis on data 1...

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

Yes No

Send an email notification when the job completes.

Execute

History

Rechercher des données

imported: PAMPA NS-IBTS G. morhua

20 shown, 1 hidden

36.52 MB

20: Report

19: Create a plot from GLM data on data 15, data 9, and data 16

a list with 1 item

18: Your analysis rating file on data 15 and data 9

2 tabs

17: Simple statistics on chosen variables on data 15 and data 9

16: GLM - Results from your population analysis on data 15 and data 9

15: Regex Find And Replace on data 14

STOC Estimate species population evolution

Machines virtuelles (local ou cloud)
Galaxy

Galaxy / Ecology

Analyses de données Workflow Visualize Données partagées Aide Utilisateur

Using 84%

Tools

search tools

Get Data

Collection Operations

GENERAL TEXT TOOLS

Text Manipulation

Filter and Sort

Join, Substrait and Group

Convert Formats

FASTA/FASTQ

FASTQ Quality Control

Assembly

NCBI Blast

RAD-seq

Metagenomic Analysis

QIIME

Workflow

Complete EBV workflow example from GIBF data

History

Rechercher des données

tuto GIBF data pre-processing

12 shown, 10 hidden

237 MB

11: data 12 Dec 2018

1 tab

12: Filter on data 10

1 tab

13: Count on data 1

1 tab

14: Filter on data 1

1 tab

15: Summary Statistics

1 tab

Workflows

Histories

Details

Insert Add an input dataset

Email notification

Yes No

Send an email notification when the job has completed.

Output cleanup

Yes No

Upon completion of this step, delete any shared outputs from completed workflow steps if they are no longer required as inputs.

Ogrinfo

Whisker

The

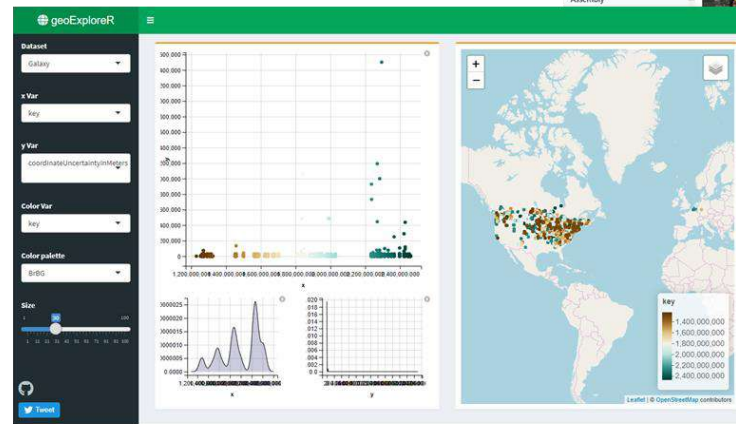
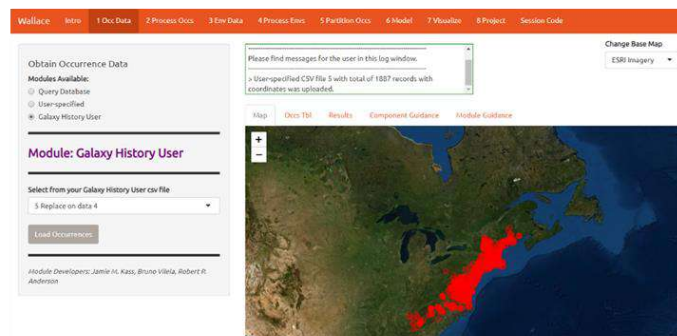
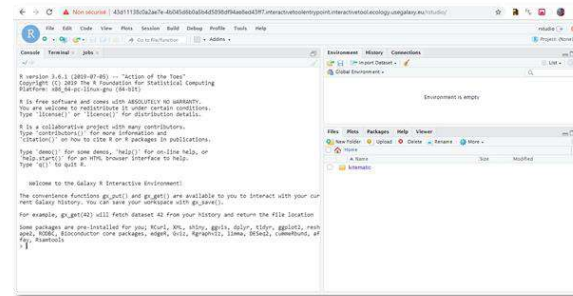
CC BY

```

graph TD
    RStudio --> Autres[Autres plateformes]
    RStudio <--> D4science
    D4science <--> Jupyter[Jupyter Notebook]
    Autres --> Jupyter
  
```

du RStudio

du Jupyter Notebook



Volet analyse

Testez la plateforme d'analyse / couplage de données du PNDB <https://ecology.usegalaxy.eu/>

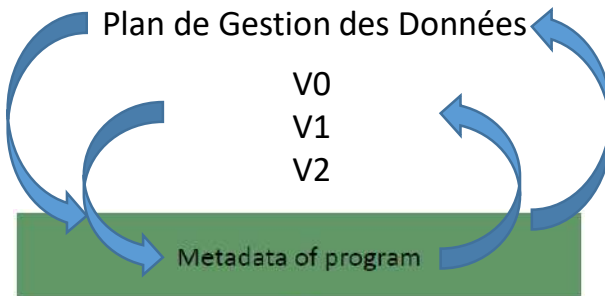
Tutoriels : <https://training.galaxyproject.org/>

Codes sources : <https://github.com/65MO/Galaxy-E>
<https://github.com/galaxyecology/tools-ecology>

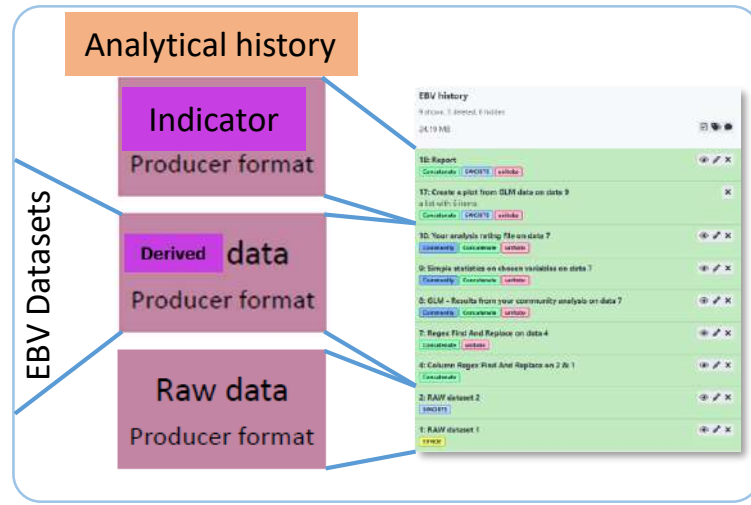
Standardisation des codes sources par conda/container/Galaxy

- **Pour faciliter la définition de dépendances via un seul système**
 - Conda est utilisable pour tous les langages de programmation
- **Pour faciliter le déploiement sur tous types d'architectures matérielles et logicielles**
 - Les technologies de containerisation facilite le déploiement sur infrastructures de calcul. Le déploiement sous Windows reste une difficulté tout de même!
- **Pour un haut degré d'accessibilité, transparence, et reproductibilité**
 - Une plateforme web orientée workflow et collaboration

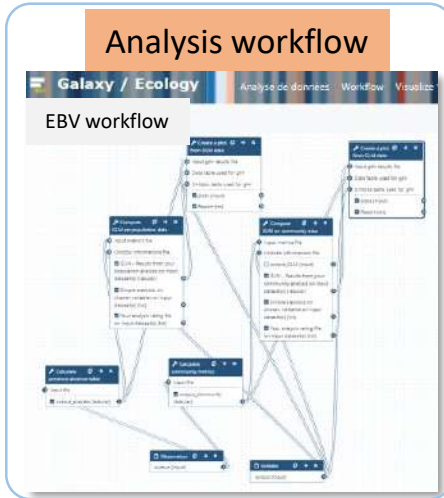
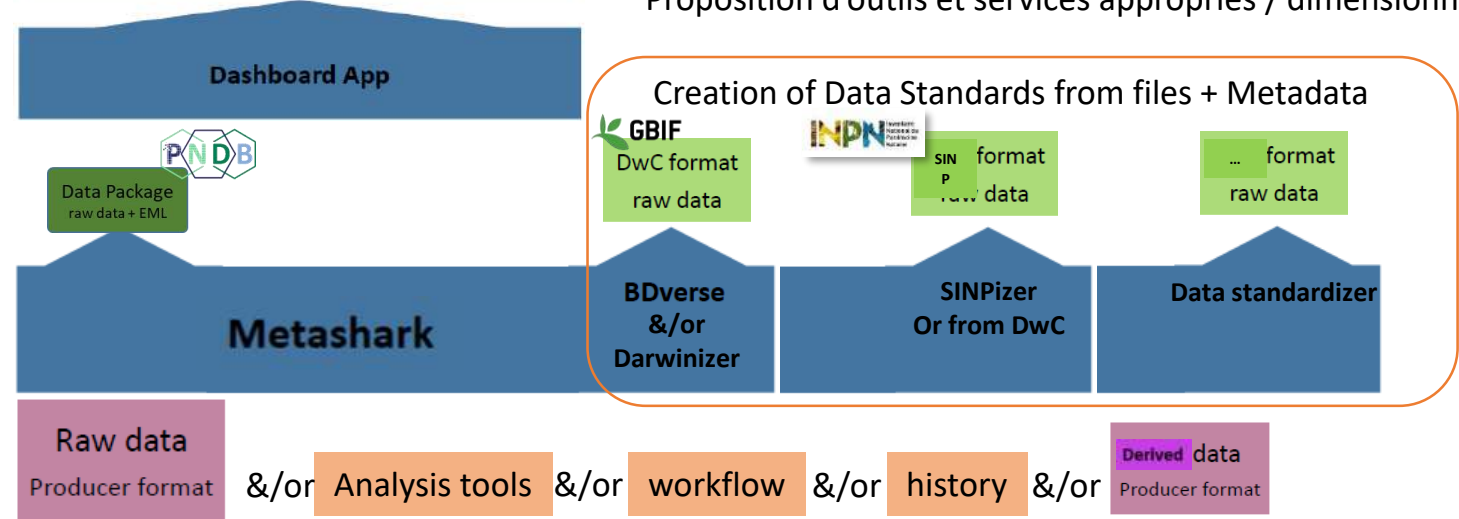
Vision intégrée (actuelle ;))



- Infrastructure
- Stockage
 - Calcul
 - Standards
 - ...



Proposition d'outils et services appropriés / dimensionnés



Vision intégrée préliminaire

Data -> Analysis

Pôle National de Données de Biodiversité Data Catalog

DATA SUMMARY ABOUT Jump to: DOI or ID Go YVAN LE BRAS

< Back to search Home / Search / Metadata

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.** um:node:PND_B_METACAT_PRODUCTION. doi:10.48502/8bb5-pk85

Citations 0 Downloads 0 Views 0

Copy Citation Analyze Edit

Files in this dataset Package: um:node:be50a625-e8a7-4cd5-b3ba-9ab591d41aa2

Name	File type	Size	
Metadata: Kakila database of marine mammal observation data in the AGOA sanctuary - French Antilles.xml	EML v2.2.0	140 KB	Download
BDD_Kakila_v2_20210420_observateur.tsv	text/csv	11 KB	Download
BDD_Kakila_v2_20210420_observation.tsv	text/csv	463 KB	Download
BDD_Kakila_v2_20210420_sortie.tsv	text/csv	217 KB	Download

Show 5 more items in this data set

General

Annotations

is about biodiversity is about marine mammals is about geographical distribution is about Species distribution is about Sea regions is about Ecological stocktaking is about Natural area

is about Landscape is about Ecosystem is about Animal ecology

Identifier doi:10.48502/8bb5-pk85

Abstract Database collected as part of the Lorraine Coché master's 2 internship "Inventory and structuring of marine mammal observation data around Guadeloupe" in 2020 (Master Tropical marine ecosystems at the University of the Antilles). This database centralizes and harmonizes the data collected by the team of the Agoa Sanctuary (Aire Marine Protégée), the OIMAG (Observatory of Marine Mammals of the Guadeloupe Archipelago), the NFO BREACH, and whale-watching companies Cétacés Caraïbes, Guadeloupe Evasion Découverte et Aventures Marines.

Vision intégrée préliminaire

Analysis <- Data

The screenshot shows the Galaxy / Ecology web interface. A 'Download' dialog box is open, displaying search results for datasets. The dialog has a 'Type to Search' input field and a list of datasets with their descriptions.

Type to Search	Search Results
	National Human Genome Research Institute (NHGRI)
Regular	
■ Sentinel-3	European Commission's Copernicus Earth Observation Programme. Sentinel-3 is a polar orbiting satellite that completes 14 orbits of the Earth a day.
■ Sentinel-5P Level 2	Observations from the Sentinel-5 Precursor satellite of the Copernicus Earth Observation Programme. It contains a polar orbiting satellite that completes 14 orbits of the Earth a day.
■ Coupled Model Intercomparison Project 6	The sixth phase of global coupled ocean-atmosphere general circulation model ensemble
■ CMIP6 GCMs downscaled using WRF	High-resolution historical and future climate simulations from 1980-2100
■ GBIF European region public datasets	The Global Biodiversity Information Facility is an international network and data infrastructure aimed at providing anyone, anywhere, open access to data about all types of life on Earth.
■ NOAA Global Forecast System (GFS)	The Global Forecast System (GFS) is a weather forecast model produced by the National Centers for Environmental Prediction (NCEP).
■ NOAA Unified Forecast System Subseasonal to Seasonal Prototype 5	The Unified Forecast System Subseasonal to Seasonal prototype 5 (UFS S2Sp5) dataset is reforecast data from the UFS atmosphere-ocean.

The background interface shows a sidebar with 'Tools' and 'Get Data' sections, a main workspace with a 'COVID-19 Re' dataset, and a 'History' panel on the right listing various datasets like 'STOC-EPS data from GBIF fin' and 'GimmTMB - Te mp trends per year'.

Vision intégrée préliminaire

Analysis = Data

Uri

exact =

Label

exact =

AveExpr

= +

PValue

= +

logFC

= +

t

= +

adj.P.Val

= +

name

<https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/rna-seq-analysis-with-askomics-it/tutorial.html>

Relation

gene included in QTL ☒ Reverse

On the same

☒ Same reference

☐ Same strand

☒ Strict