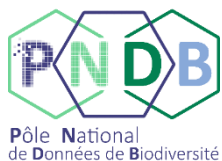
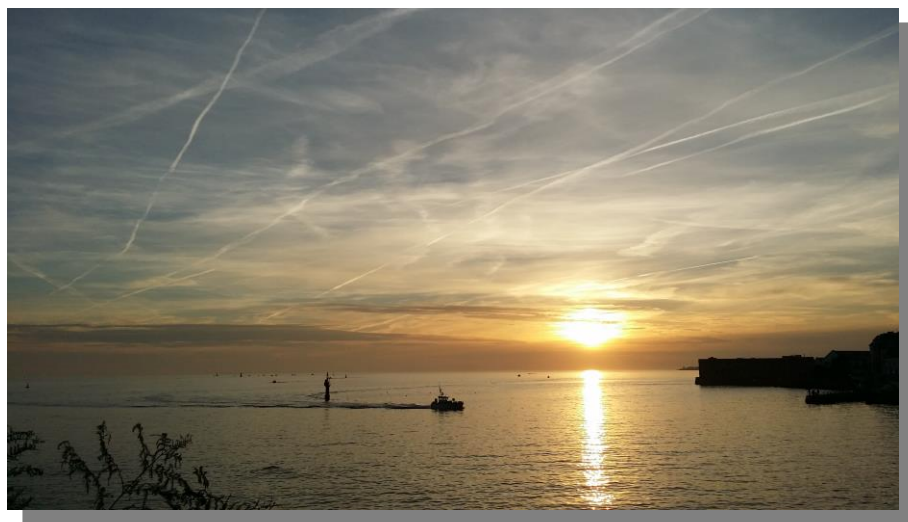
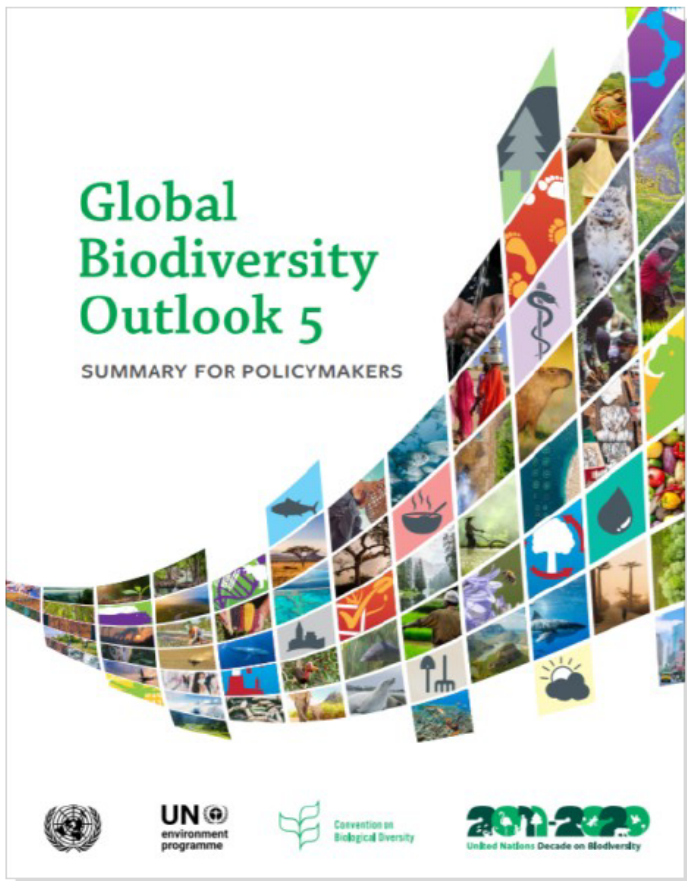


# Vocabularies for a national Biodiversity e-infrastructure



# Context

- **Biodiversity: we are losing the fight!**



September  
15<sup>th</sup>

« None of the Aichi Biodiversity  
Targets will be fully met »

Need to follow biodiversity dynamics and  
produce metrics to monitor its state and  
guide decision making






# Context

- **Biodiversity data access crisis**

- Research applications of primary biodiversity databases in the digital age, 2019 (**347 sources**)
- The Biodiversity Informatics Landscape: Elements, Connections and Opportunities, Bingham et al., 2017 (**71 sources**)
- Biodiversity and ecosystem services in environmental profit & loss accounts, Chaplin-Kramer et al., 2016 (**11 sources**)
- French Foundation for Biodiversity Research (**22 sources**)
- Healthy ecosystem metric framework: biodiversity impact (**7 sources**)

=> Compilation on <https://github.com/go-fair-ins/GO-FAIR-BiodiFAIRse/blob/master/Datasource/Listing.md>

## Data integration enables global biodiversity synthesis

 J. Mason Heberling,  Joseph T. Miller,  Daniel Noesgaard,  Scott B. Weingart, and  Dmitry Schigel

<https://doi.org/10.1073/pnas.2018093118>






# Context

- **Biodiversity data access crisis**

- Research applications of primary biodiversity databases in the digital age, 2019 (**347 sources**)    **90 Not Accessible in 2019**
- The Biodiversity Informatics Landscape: Elements, Connections and Opportunities, Bingham et al., 2017 (**71 sources**)
- Biodiversity and ecosystem services in environmental profit & loss accounts, Chaplin-Kramer et al., 2016 (**11 sources**)
- French Foundation for Biodiversity Research (**22 sources**)
- Healthy ecosystem metric framework: biodiversity impact (**7 sources**)

=> Compilation on <https://github.com/go-fair-ins/GO-FAIR-BiodiFAIRse/blob/master/Datasource/Listing.md>

## Data integration enables global biodiversity synthesis

 J. Mason Heberling,  Joseph T. Miller,  Daniel Noesgaard,  Scott B. Weingart, and  Dmitry Schigel

<https://doi.org/10.1073/pnas.2018093118>

# Context

- **Biodiversity data access crisis**

- Research applications of primary biodiversity databases in the digital age, (347 sources) **90 Not Accessible in 2019**
- The Biodiversity Informatics Landscape: Elements, Connections and Opportunities, Bingham et al., 2017 (**71 sources**)
- Biodiversity and ecosystem services in environmental profit & loss accounts, Chaplin-Kramer et al., 2016 (**11 sources**)
- French Foundation for Biodiversity Research (**22 sources**)
- Healthy ecosystem metric framework: biodiversity impact (**7 sources**)

=> Compilation on <https://github.com/go-fair-ins/GO-FAIR-BiodiFAIRse/blob/master/Datasource/Listing.md>

=> ~25% Not Accessible

No licences

Need to ask access

Where is the data ?

PDF, no dataset

Error 404 Not Found

Only derivated / degraded data

login AAAS required

Dataset corrupted



# Context

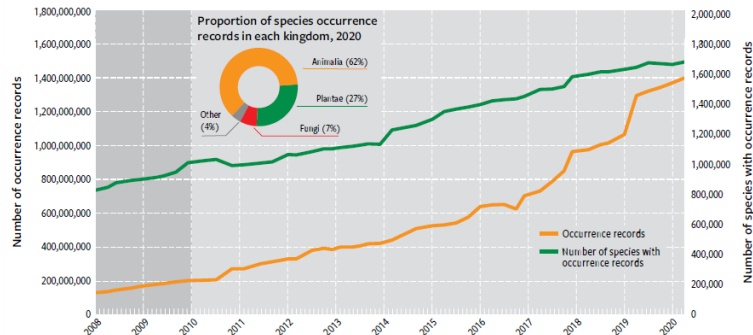
- But there is hope!!!  
=> Places to find data:



**Summary of target achievement**

Significant progress has been made since 2010 in the generation, sharing and assessment of knowledge and data on biodiversity, with big-data aggregation, advances in modelling and artificial intelligence opening up new opportunities for improved understanding of the biosphere. However, major imbalances remain in the location and taxonomic focus of studies and monitoring. Information gaps remain in the consequences of biodiversity loss for people, and the application of biodiversity knowledge in decision making is limited. **The target has been partially achieved (medium)**

Figure 19.2. Growth in GBIF-mediated species occurrence records<sup>12</sup>



Mobilization of open-access data through the Global Biodiversity Information Facility (GBIF). The lines show the number of species occurrence records over time, and the number of species having occurrence records.

## Summary of Holdings

A summary of all datasets in our catalog.

### Summary of holdings

825,972 datasets

The total number of publicly-available metadata records. Only the latest version of each metadata record is counted. A "dataset" here is defined by a single metadata record which may be packaged with one or more data files.

79 TB

of content

The volume of all publicly-available metadata and data files in this repository. Only the latest version of each file is included.

### Metadata Assessment

This graph shows the assessment score for all metadata in this repository, based on the **DataONE FAIR Suite**. The FAIR suite evaluates metadata based on these criteria: Findable, Accessible, Interoperable, Reusable. Each point represents the average score for that category across all the versions of datasets that were changed in that month.



<https://search.dataone.org/profile>

=> And one common AMAZING Language:

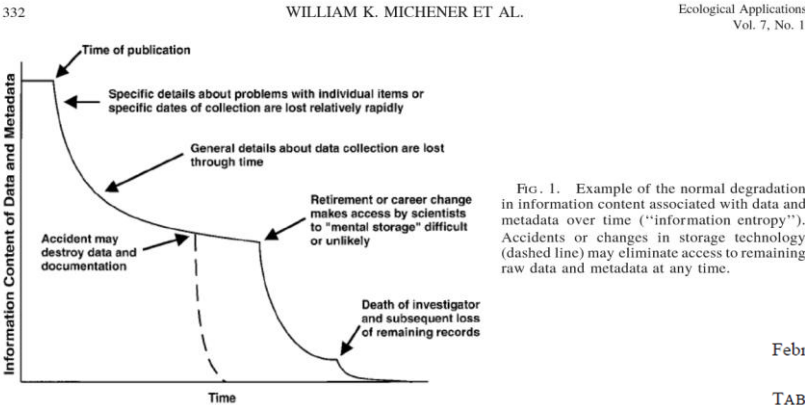


# Ecological Metadata Language (EML)

<https://eml.ecoinformatics.org/>

# Data and metadata focus

## PNDB position



### NONGEOSPATIAL METADATA FOR THE ECOLOGICAL SCIENCES, *Michener et al. 1997*

February 1997 ECOLOGICAL METADATA 339

TABLE 2. Content of metadata (refer to classes in Table 1) associated with three levels of secondary data utilization.

Metadata descriptor classes	Levels of secondary data utilization and associated metadata content		
	Level I: exchange with expert colleague	Level II: searchable and third party data reuse	Level III: publishable and auditable
I. Data set descriptors	X	X	X
II. Research origin descriptors		X	X
III. Data set status and accessibility		X	X
IV. Data structural descriptors	X	X	X
V. Supplemental descriptors			X

Description des variables

Provenance

# Data and metadata focus

## PNDB position

### 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:44108e76-405d-4d58-b1b3-fb4b55e3fff9>

Size: 2801033 byte

Text Format

Number of Header Lines
1
Record Delimiter
#x0A
Attribute Orientation
column
Simple Text
Field Delimeter
,

Number Of Records: 12142

2 derivations

**Positionnement  
PNDB**

### NONGEOSPATIAL METADATA FOR THE ECOLOGICAL SCIENCES, Michener *et al.* 1997

Level	Planned Use			
III	Publishable and auditable	Inadequate	Minimal	Good practice
II	Searchable and third party reuse	Minimal	Good practice	Excessive
I	Exchange with expert colleague	Good practice	Excessive	Excessive
		LOW (Free format, ASCII, narrative, or hard copy)	MEDIUM (Mixed format, partially parameterized)	HIGH (Fixed format, highly parameterized, executable, language-dependent)
Amount of structure (Formalization, level of effort)				

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



## (meta)data lanscape through *Ecological Metadata Language*

The diagram illustrates a workflow for data management. It starts with 'Data' and 'Metadata' (indicated by a plus sign) being processed by 'MetaSHARK' (with a Shiny logo and the URL <https://github.com/earnaud/MetaShARK-v2>) to create a 'Datapackage'. The 'Datapackage' is then processed by 'Metacat' (indicated by a downward arrow) to be stored in 'Systems' (including Infrastructures and organismes) and 'Dataverse'.

- External Information Systems
- Infrastructures
- - organismes



## Semantics enrichment

cedarr

CEDAR R package for API linking in an R interface.

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

PNDR

DONNÉES

RÉSUMÉ

A PROPOS

aller à:

GO

YVAN LE BRAS

[< Back to search](#)

[Home](#) / [Search](#) / [Metadata](#)

Institut de Recherche pour le Développement, UMR DIADE, France ., SouthGreen Development Platform, Agropolis Campus, Montpellier, France ., Africa Rice Center, Benin ., CEA, Institut de Biologie Français 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". urn:node:METACAT\_TEST. urn:uuid:b004039b-ca27-4719-9df9-f8e785bc2432.

55 Citations

0

Downloads

0

Views

0

Copy Citation

Analyser les données ▼

Edit

Publish with DOI ★

Files in this dataset: Package: resource\_map\_urn:uuid:b7f31123-22b7-4b07-abb8-212c7d5bc085

File Name	File type	Size	Download All
<div>  Metadata: African_rice_population_genomics_dataset_or.xml </div>	EML v2.1.1	26 KB	Download
<div>  passeport_data_F1.csv </div>	<a href="#">More info</a> text/csv	15 KB	Download
<div>  passeport_data_F2.csv </div>	<a href="#">More info</a> text/csv	17 KB	Download

Example of french data portal:

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

# Data and metadata focus

(meta)data lanscape through *Ecological Metadata Language*

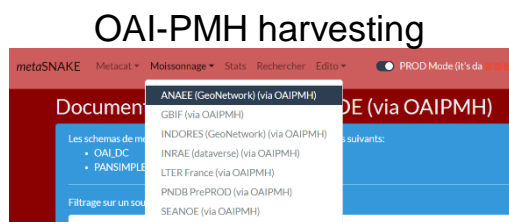
**Batch mode**

Data



**Enriched curated metadata**

Primary variables description Via EML  
Semantic enrichment



**MetaSNAKE**

**API MetaSHARK ?**

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

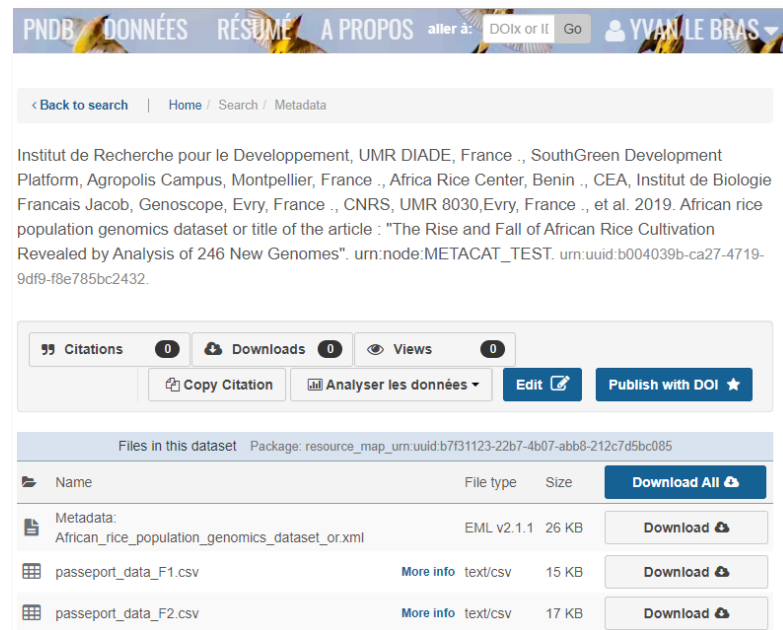
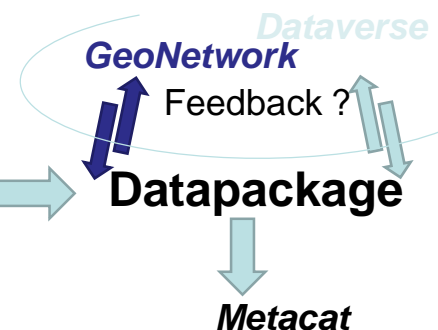
- External Information Systems
- Infrastructures
- organismes

Dataverse

**GeoNetwork**

**Dublin core, ISO19115,**

**Heterogeneous Initial Metadata**



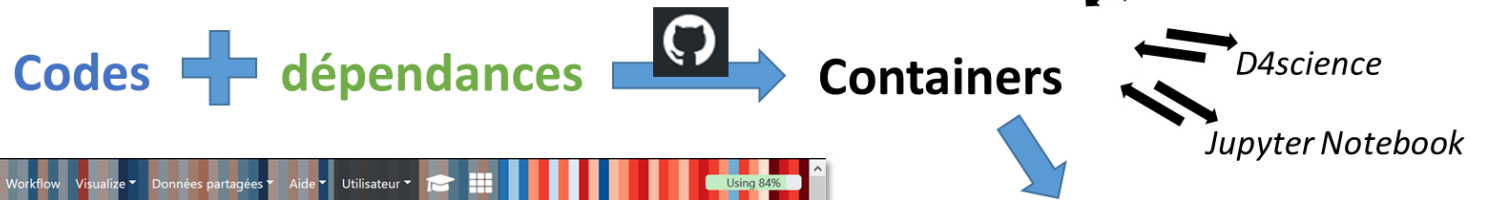
Example of french data portal:

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



# Analysis

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



**Galaxy / Ecology**

Analyse 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

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

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**

**Galaxy / Ecology**

Analyse 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, Substrat and Group

Convert Formats

FASTA/FASTQ

FASTQ Quality Control

Assembly

NCBI Blast

RAD-seq

Metagenomic Analysis

QIIME

Mothur

**tuto GDF data pre-processing**

13: data 12.GeneSON

12: Filter on data 10

11: Count on data 1

10: Filter on data 7

9: Summary Statistics

**Workflows**

**Histories**

# A preliminary integrated vision

Data -> Analysis

Pôle National de Données de Biodiversité Data Catalog
DATA SUMMARY ABOUT Jump to: DOIx or If 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.** urn:node:PNDB\_METACAT\_PRODUCTION. doi:10.48502/8bb5-pk85.

Citations 0
Downloads 0
Views 0
Copy Citation
Analyze
Edit

RStudio
Jupyter Notebook
Galaxy for Ecology
Download All

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
BDD_Kakila_v2_20210420_observateur.tsv	text/csv	11 KB
BDD_Kakila_v2_20210420_observation.tsv	text/csv	463 KB
BDD_Kakila_v2_20210420_sortie.tsv	text/csv	217 KB

[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 OMMAG (Observatory of Marine Mammals of the Guadeloupe Archipelago), the NPO BREACH, and whale-watching companies Cétacés Caraïbes, Guadeloupe Evasion Découverte et Aventures Marines.

# A preliminary integrated vision

Analysis <- Data

The screenshot shows the Galaxy / Ecology web interface. A 'Download' dialog box is open, displaying a search bar and a list of datasets. The datasets listed are:

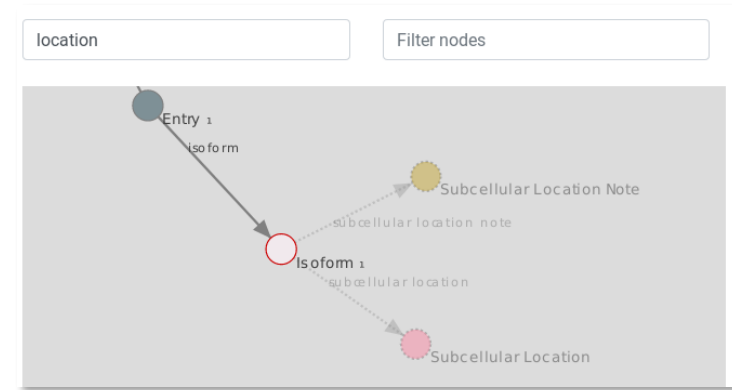
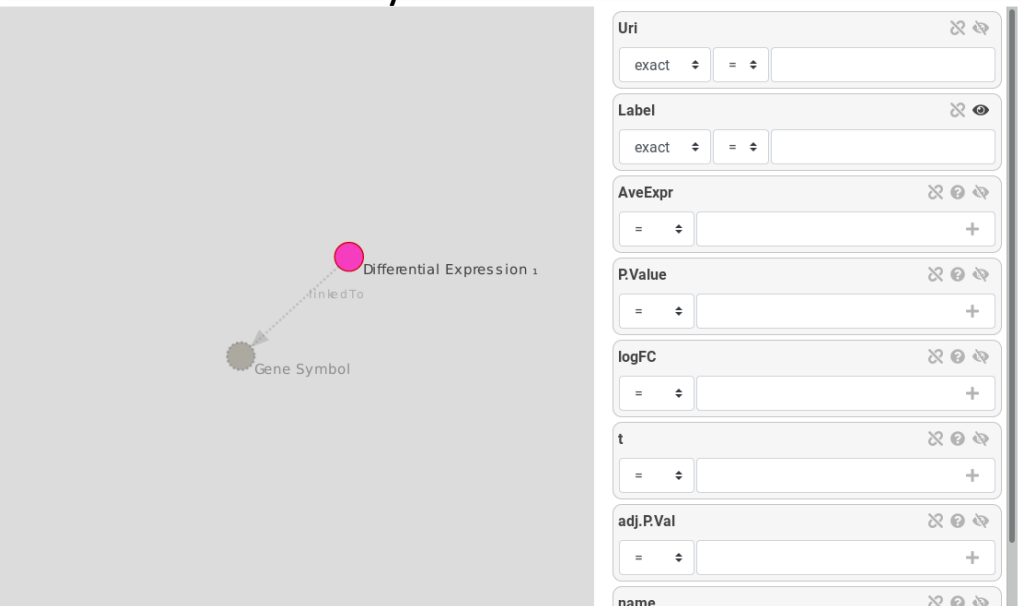
- Sentinel-3**: National Human Genome Research Institute (NHGRI)
- Sentinel-5P Level 2**: European Commission's Copernicus Earth Observation Programme. Sentinel-3 is a polar orbiting satellite that completes 14 orbits of the Earth a day.
- Coupled Model Intercomparison Project 6**: Observations from 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.
- CMIP6 GCMs downscaled using WRF**: The sixth phase of global coupled ocean-atmosphere general circulation model ensemble
- GBIF European region public datasets**: High-resolution historical and future climate simulations from 1980-2100
- NOAA Global Forecast System (GFS)**: 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 Unified Forecast System Subseasonal to Seasonal Prototype 5**: 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 dialog box has 'Cancel' and 'Ok' buttons at the bottom right.

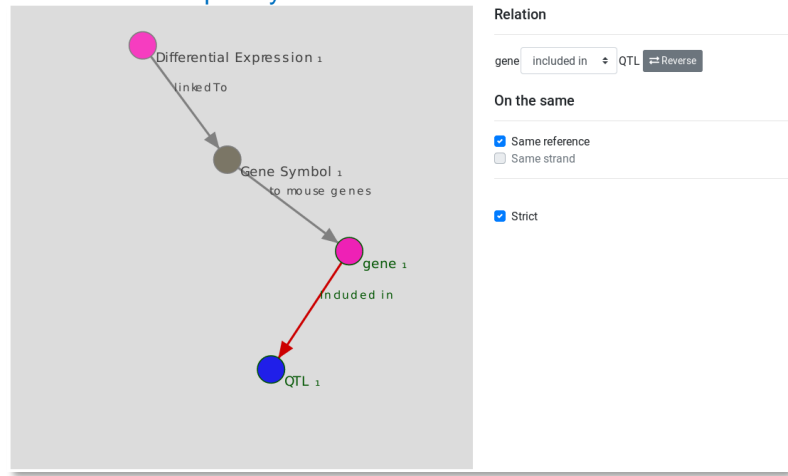


# A preliminary integrated vision

Analysis = Data



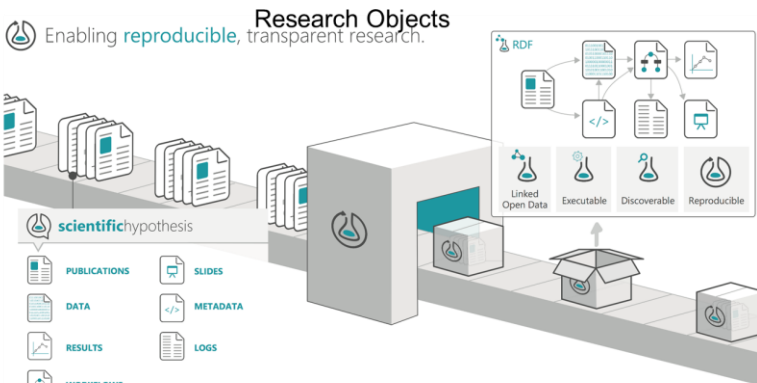
<https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/rna-seq-analysis-with-askomics-it/tutorial.html>





# Vocabularies

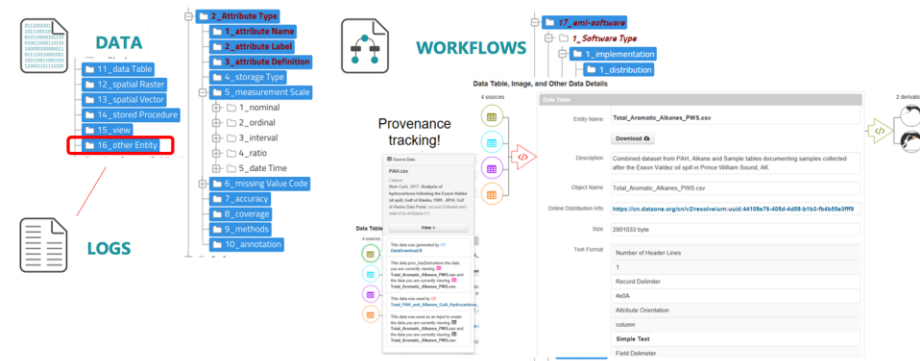
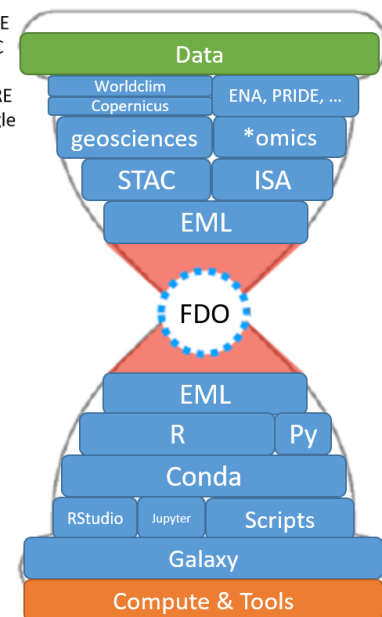
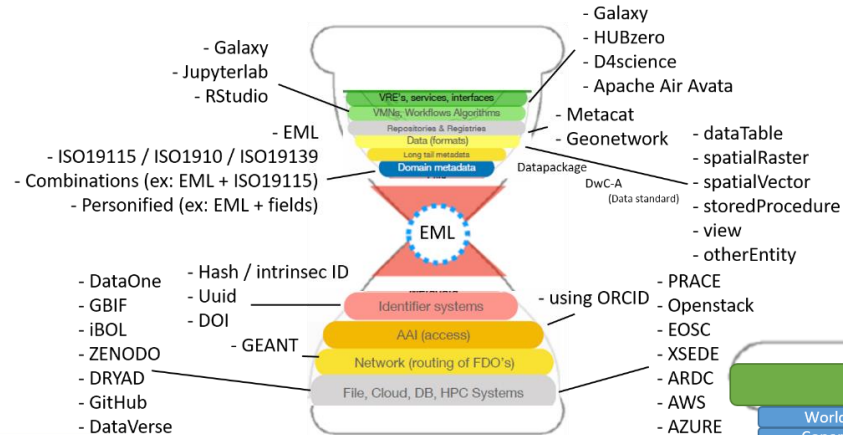
## • Metadata standard as THE preliminary « vocabulary » layer



<http://www.researchobject.org/>

Matthew B. Jones, Margaret O'Brien, Bryce Mecum, Carl Boettiger, Mark Schildhauer, Mitchell Maier, Timothy Whiteaker, Stevan Earl, Steven Chong. 2019. **Ecological Metadata Language version 2.2.0.**

**METADATA**



RO must be associated to metadata -> here we see the interest of the « completeness » of EML as this standard seems to be perfect to inform about each component of ROs

# Vocabularies

- Metadata standard as THE preliminary « vocabulary » layer
- Terminological resources as a second layer for any kind of Research Object / Digital Object

## *EML xml metadata file: keyword*

```
<taxonomicClassification>
  <taxonRankValue>Ziphius cavirostris</taxonRankValue>
  <taxonId provider="worms">137127</taxonId>
</taxonomicClassification>
</taxonomicCoverage>
</coverage>
<annotation id="kw3">
  <propertyURI label="is about">http://purl.obolibrary.org/obo/IAO_0000136</propertyURI>
  <valueURI label="biodiversity">http://aims.fao.org/aos/agrovoc/c_33949</valueURI>
</annotation>
<annotation id="kw4">
  <propertyURI label="is about">http://purl.obolibrary.org/obo/IAO_0000136</propertyURI>
  <valueURI label="marine mammals">http://aims.fao.org/aos/agrovoc/c_22acc50e</valueURI>
</annotation>
```

## *A CEDAR dedicated R package*

### cedarr

CEDAR R package for API linking in an R interface.

### Installation

You can install this package with the following command:

```
devtools::install_github("earnaud/cedarr", dependencies = TRUE)
```

The screenshot displays the 'Pôle National de Données de Biodiversité Data Catalog' interface. The main heading is 'Data catalog'. Below it, there's a navigation bar with links: '< Back to search', 'Home / Search / Metadata'. The dataset entry is for 'Lorraine Coché, Elie Arnaud, Bouveret Laurent, Romain David, Eric Foulquier, et al. 2021. Kakila database of marine mammal observation data in the AGOA sanctuary - French Antilles.xml'. The DOI is '10.48502/8bb5-pk85'. Below the entry, there are statistics: Downloads (0), Citations (0), and Views (0). A table lists the files in the dataset: 'Metadata: Kakila database of marine mammal observation data in the AGOA sanctuary - French Antilles.xml', 'BDD\_Kakila\_v2\_20210420\_observateur.tsv', 'BDD\_Kakila\_v2\_20210420\_observation.tsv', and 'BDD\_Kakila\_v2\_20210420\_sortie.tsv'. At the bottom, there's a 'General' section with annotations. One annotation is 'is about biodiversity', which is highlighted, showing a tooltip with the dataset name and a link to the dataset page.

# Vocabularies

- Metadata standard as THE preliminary « vocabulary » layer
- Terminological resources as a second layer for any kind of Research Object / Digital Object

## *EML xml metadata file*

```
<attribute id="att-experobs">
  <attributeName>expertise_observateur</attributeName>
  <attributeDefinition>Level of expertise of the observer (beginner, intermediate, expert). The level of expertise is used for the identification of cetaceans.</attributeDefinition>
  <storageType>string</storageType>
  <measurementScale>
    <nominal>
      <nonNumericDomain>
        <textDomain>
          <definition>Level of expertise of the observer (beginner, intermediate, expert). The level of expertise is used for the identification of cetaceans.</definition>
        </textDomain>
      </nonNumericDomain>
    </nominal>
  </measurementScale>
  <missingValueCode>
    <code>" "</code>
    <codeExplanation>" "</codeExplanation>
  </missingValueCode>
  <annotation>
    <propertyURI label="is similar to">https://schema.org/isSimilarTo</propertyURI>
    <valueURI label="identificationRemarks">http://rs.tdwg.org/dwc/terms/identificationRemarks</valueURI>
  </annotation>
</attribute>
```

## *Data catalog*

Pôle National de Données de Biodiversité Data Catalog

Data Table

Entity Name: BDD\_Kakila\_v2\_20210420\_observateur.tsv

Download

Description: Content of BDD\_Kakila\_v2\_20210420\_observateur.tsv

Object Name: BDD\_Kakila\_v2\_20210420\_observateur.tsv

Size: 11585 bytes

Authentication: 4a8b2111c4e93bc3b80575027da6c0fb Calculated By MD5

Text Format

Number of Header Lines

Record Delimiter

Attribute Orientation

Simple Text

Field Delimiter

Number Of Records: 498

Attribute Information

Variables

- code\_observateur
- code\_organisme
- expertise\_observateur

Name: expertise\_observateur

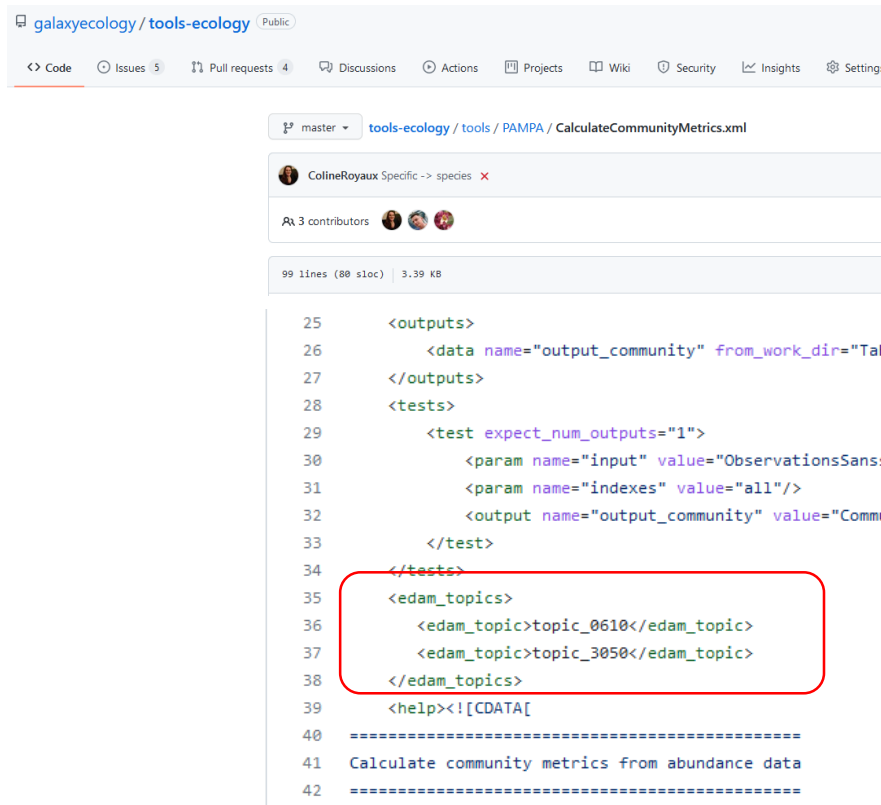
Annotations

is similar to identificationRemarks

Label

# Vocabularies

- Metadata standard as THE preliminary « vocabulary » layer
- Terminological resources as a second layer for any kind of Research Object / Digital Object



```
25     <outputs>
26         <data name="output_community" from_work_dir="TabCommunityIndexes.tabular" format="tabular"/>
27     </outputs>
28     <tests>
29         <test expect_num_outputs="1">
30             <param name="input" value="ObservationsSansszcl1_cropped.tabular"/>
31             <param name="indexes" value="all"/>
32             <output name="output_community" value="Community_metrics_cropped.tabular"/>
33         </test>
34     </tests>
35     <edam_topics>
36         <edam_topic>topic_0610</edam_topic>
37         <edam_topic>topic_3050</edam_topic>
38     </edam_topics>
39     <help><![CDATA[
40 =====
41 Calculate community metrics from abundance data
42 =====
```

## *Vocabularies for Processing*

# Difficulties / Issues / Challenges

- **Vocabularies are in the same time THE solution to connect people with others people + people to machine + machine to machine AND the less understandable / useable scientific component by domains researchers**
  - XML / JSON / RDF / ontologies / thesaurus / concept / provenance are not terms the users like ;)
  - We need GUI and UX to interface resources and humans!!!!
- **We need common approach/tools/framework/API BUT this is dangerous to focus on a lonely solution**
  - A unique solution is easier to use + communicate on (Google)
- **A strong movement of « semantics » guys to directly plug semantics to Research objects**
  - Not something optimal! The jargon & technical aspects behind semantics need to be hidden, and metadata standards can help here!

# Thank you !

## PNDB team

**Coline Royaux** – engineer R /  
Galaxy dev (workflows Galaxy pour  
calcul indicateurs espèces /  
communautés)

**Elie Arnaud** – engineer R Shiny /  
knowledge – metadata dev

**Julien Sananikone** – engineer  
DevOps / sys admin / web dev

**Yvan Le Bras** – Beta tester

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

PNDB « bricks »:

MetaShARK Metadata work: [Yvan.le-bras@mnhn.fr](mailto:Yvan.le-bras@mnhn.fr)

<https://youtu.be/OVVISMzRGtw>

Data metadata portal:

<https://youtu.be/STwsYDHEt2A>

Galaxy Europe demo:

- <https://youtu.be/HeIAHggX6D4>

- [Essential biodiversity variables on Galaxy: implementing the PAMPA application](#)

- [Producing biodiversity indicators from citizen science projects: update of birds and bats monitoring schemes on Galaxy-E](#)

