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 :

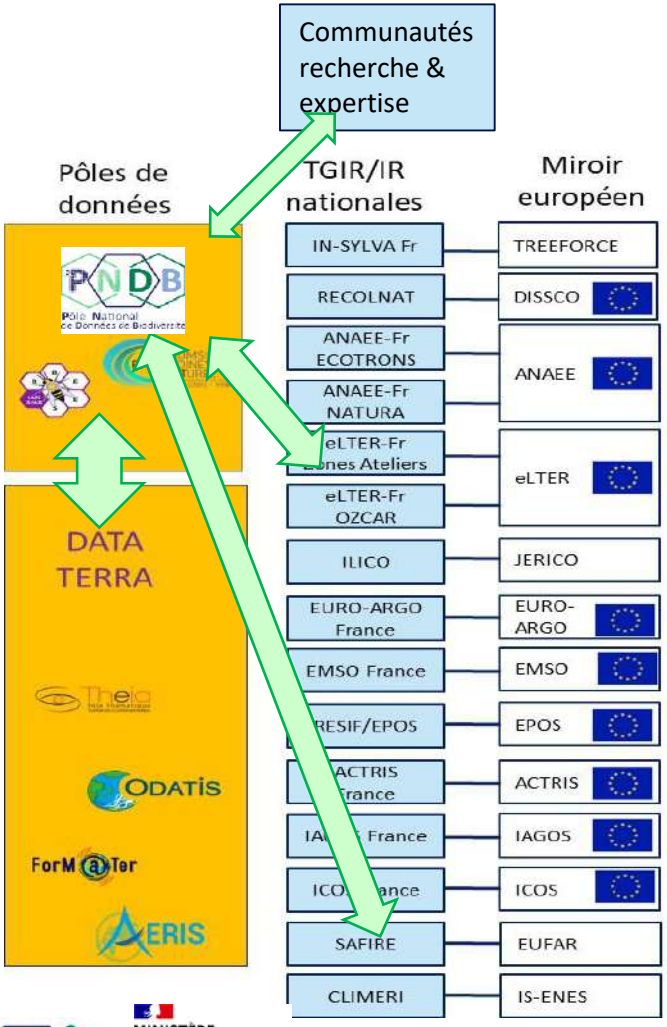
Accès aux jeux de métadonnées-données
Services et outils associés
Aide à la bancarisation
Animation des communautés autour de Sci. ouverte
Stimulation des interactions producteurs/utilisateurs

Pôle National de Données de Biodiversité

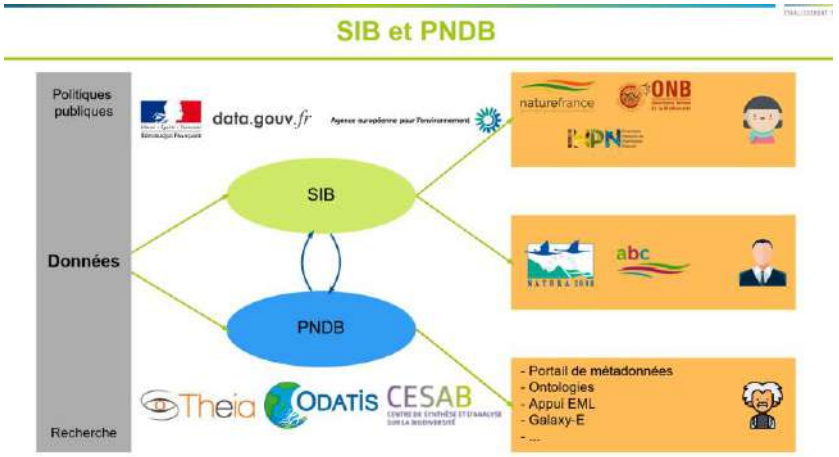
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'INNOVATION
Liberté
Égalité
Fraternité

Positionnement à l'échelle nationale

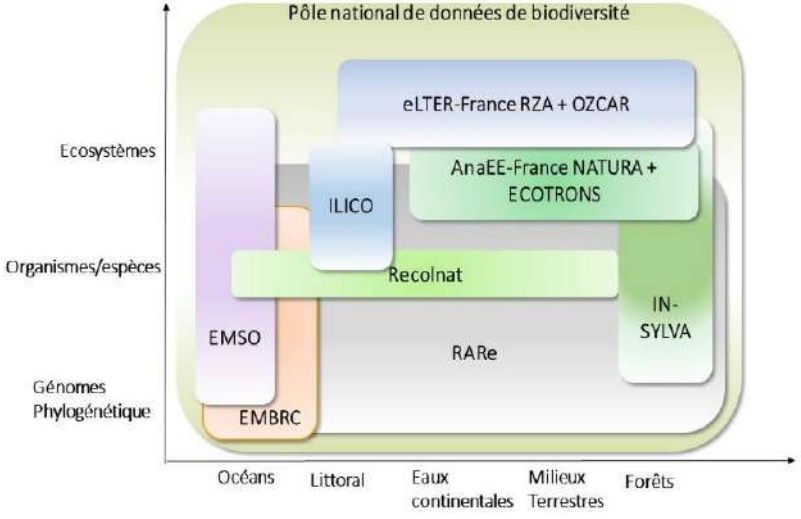
Par rapport aux producteurs de données



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)

- **Eléments 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

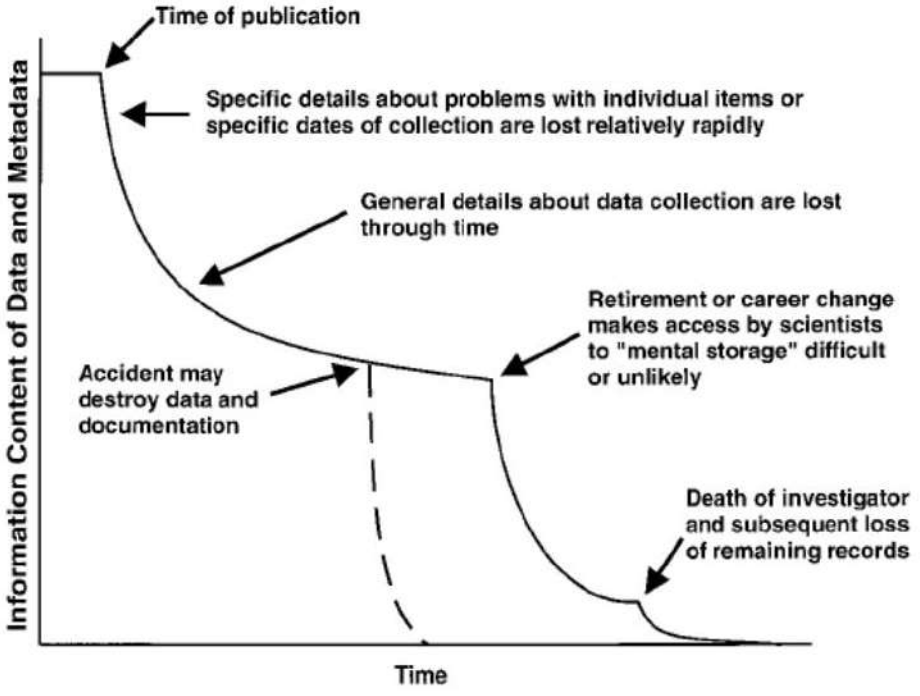


FIG. 1. Example of the normal degradation in information content associated with data and metadata over time ("information entropy"). Accidents or changes in storage technology (dashed line) may eliminate access to remaining raw data and metadata at any time.

Données et métadonnées du PNDB

Positionnement PNDB

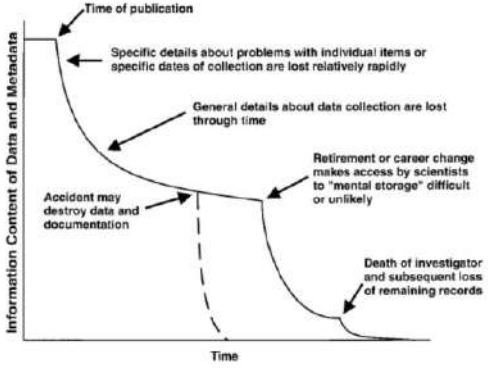


FIG. 1. Example of the normal degradation in information content associated with data and metadata over time ("information entropy"). Accidents or changes in storage technology (dashed line) may eliminate access to remaining raw data and metadata at any time.

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

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

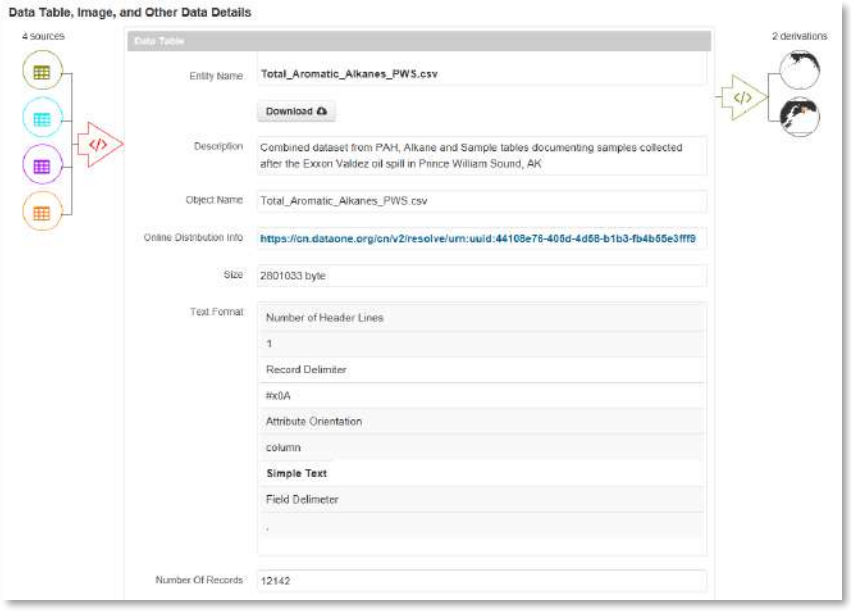
Metadata descriptor classes	Levels of secondary data utilization and associated metadata content		
	Level I: exchange with expert colleague	Level II: searchable and third party data reuse	Level III: publishable and auditable
I. Data set descriptors	X	X	X
II. Research origin descriptors		X	X
III. Data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

Description des variables

Provenance

Données et métadonnées du PNDB

Positionnement PNDB



Positionnement PNDB

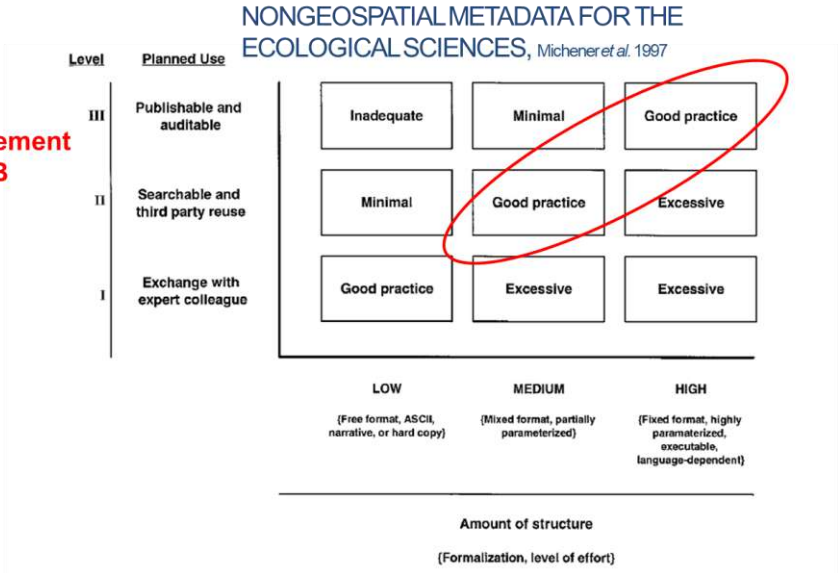


FIG. 3. Degree of metadata format/structure sufficient for three levels of projected secondary data utilization.

Données et métadonnées du PNDB

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

Mode single

Données + **Métadonnées** → **Datapackage**

GeoNetwork

SI externes
-
Infrastructures
- organismes

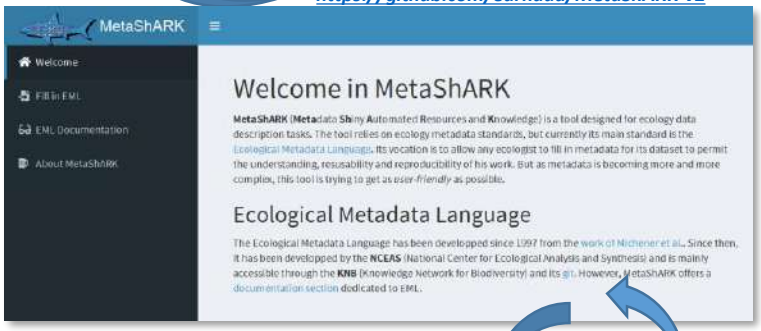
Dataverse

MetaSHARK

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

Metacat

- dataset1
- dataset2
- script1
- article1
- protocol1



EML Assembly Line

https://ediorg.github.io/EMLassemblyline/index.html

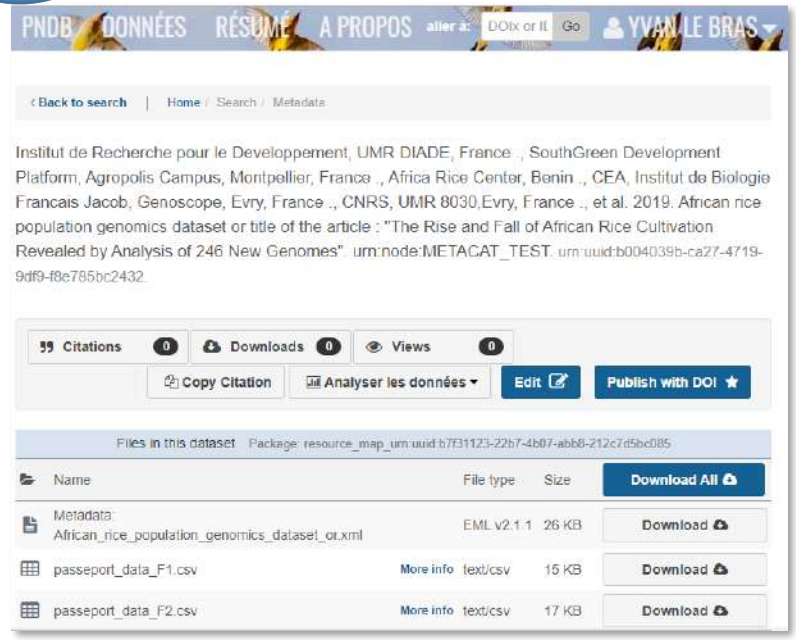
EML

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

Pôle National de Données de Biodiversité

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'INNOVATION

Données et métadonnées du PNDB

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

Mode batch

Données + **Métadonnées enrichies unifiées**

Primary variables description
Semantic enrichment

Via EML

Datapackage

Dataverse
GeoNetwork
Feedback ?

Metacat

OAI-PMH harvesting

MetaSNAKE

API MetaSHARK ?

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

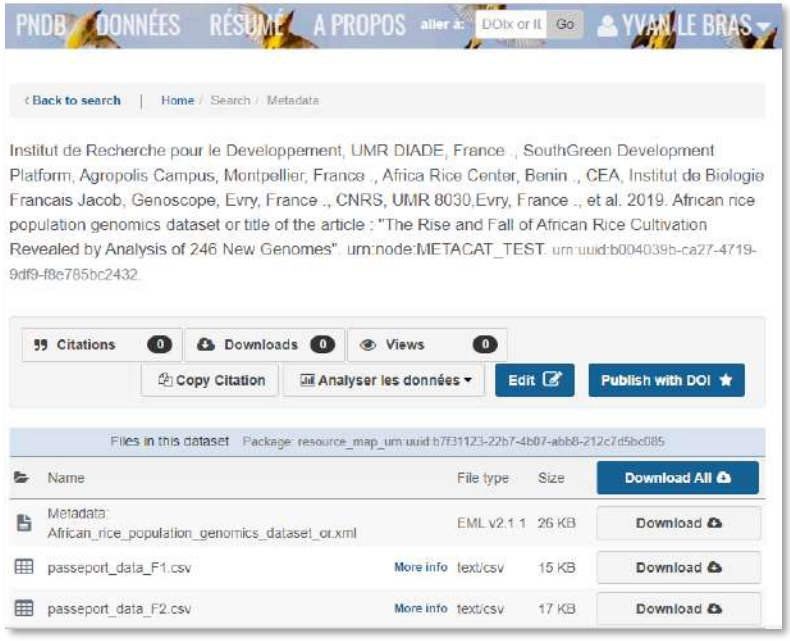
- dataset1
- dataset2
- script1
- article1
- protocol1

SI externes
- Infrastructures
- organismes

Dataverse
GeoNetwork
Solution à façon
...

Dublin core, ISO19115, ...

Métadonnées initiales hétérogènes



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

Pôle National de Données de Biodiversité

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'INNOVATION

Standardisation par la métadonnée vs par un standard de données

- **Pour faciliter l'accès aux données brutes par touts 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