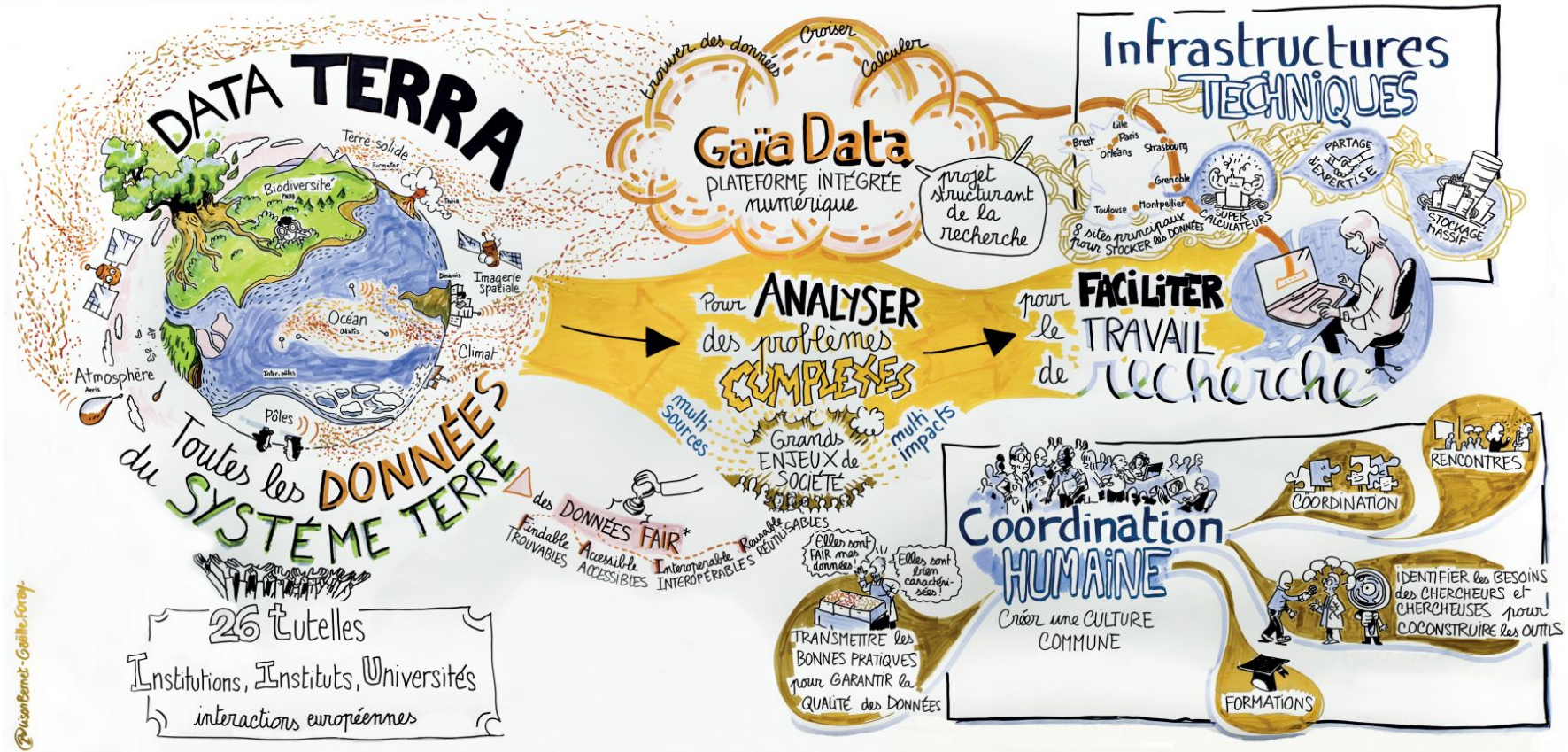


DATA TERRA

Pôle National de Données de Biodiversité

GAIA Data

19-20-21 nov. 2024 — Station CNRS - Moulis

ANNOTATION

Ateliers — Formations

ecoinfoFair2024.sciencesconf.org

EcoinfoFAIR

éco-informatique FAIR* pour et par les communautés

(méta)données de biodiversité sons & images

*Facile à trouver, Accessible, Interopérable, Réutilisable

par
Yvan Le Bras (PNDB-Data Terra, MNHN) &
Olivier Norvez (PNDB-Data Terra, FRB & MNHN)

Data Terra

L'IR DATA TERRA propose des **services** autour des données d'observation du **système Terre interopérables** et **interdisciplinaires** à tous les niveaux

https://www.data-terra.org/

GAIA DATA, Equipex+ PIA 3
un projet structurant

8
Années pour la réalisation du projet

21
Organismes partenaires

62 M €
Budget Total (dont 16, 2 millions d'euros de l'ANR)

400
Ingénieurs, scientifiques et experts

Contribution aux initiatives internationales et européennes en **appui aux politiques publiques de développement durable**

Accès simplifié aux données multi-sources

Interopérabilité des services

Développement de services **sur le cycle complet de la donnée**

Se baser sur les capacités, institutions, structures et **ressources existantes**

Approches multidisciplinaires intégrées pour l'utilisation des données de recherche d'observation de la Terre

www.gaia-data.org

INFRASTRUCTURE GAIA DATA

Architecture de haut niveau




Métadonnées

Métadonnées

Métadonnées

Métadonnées

Pôles de données et Service, CLIMERI, DINAMIS, …

D'autres données d'intérêt pour les communautés (Newspace, ECMWF, …)

Métadonnées

Serveur de vocabulaire

Authentification commune


Utilisateurs

Interface machine 2 machine : API

STAC

RDF

Interaction

Interaction

Centralisation des métadonnées qui sont complétées, harmonisées sur les meilleurs standards (GeoDCAT / STAC)

Métadonnées enrichies, harmonisées et standardisées

Un **portail de découverte et de connaissance** qui fonctionne comme un mélange de Google et d'un site marchand comme fnac.com amazon.com

Panier de collection de données

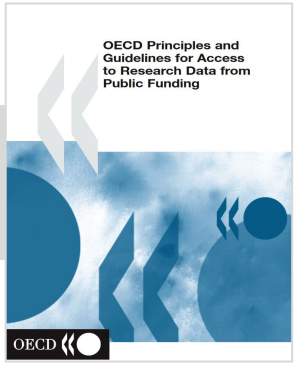
Système d'hypervision

Les outils favoris de l'utilisateur capables d'utiliser les API des collections

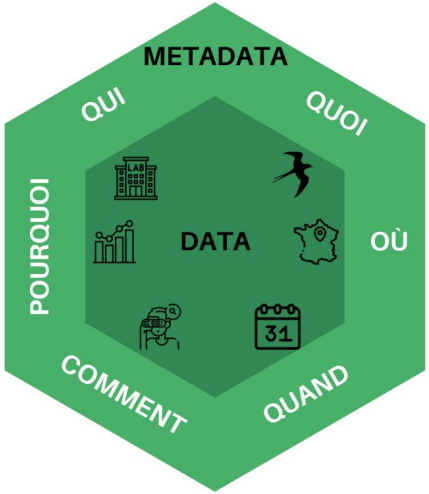
Virtual Research Environment proposé par GAIA DATA pour l'exploration des données

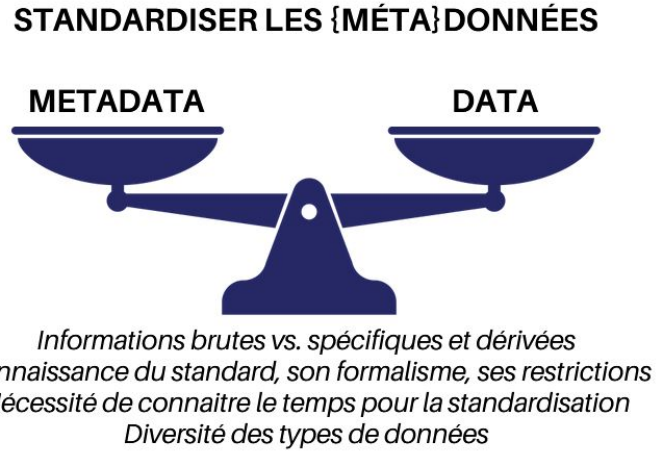
GAIA Dataspace Espace de données pour un projet, un centre, …

Comprendre les données par les métadonnées : *définitions*



"Les données de recherche sont définies comme des enregistrements factuels utilisés comme sources primaires pour la recherche scientifique, et qui sont généralement acceptés dans la communauté scientifique comme nécessaires pour valider les résultats de la recherche." **OCDE, 2007.** https://www.oecd.org/sti/inno/38500813.pdf





"Les métadonnées, que l'on peut définir simplement comme « des données sur les données», sont un moyen de nommer les choses et de représenter les données et leurs relations." **Christine L. Borgman, 2020.** https://books.openedition.org/oep/14692

Comprendre les données par les métadonnées : *définitions*

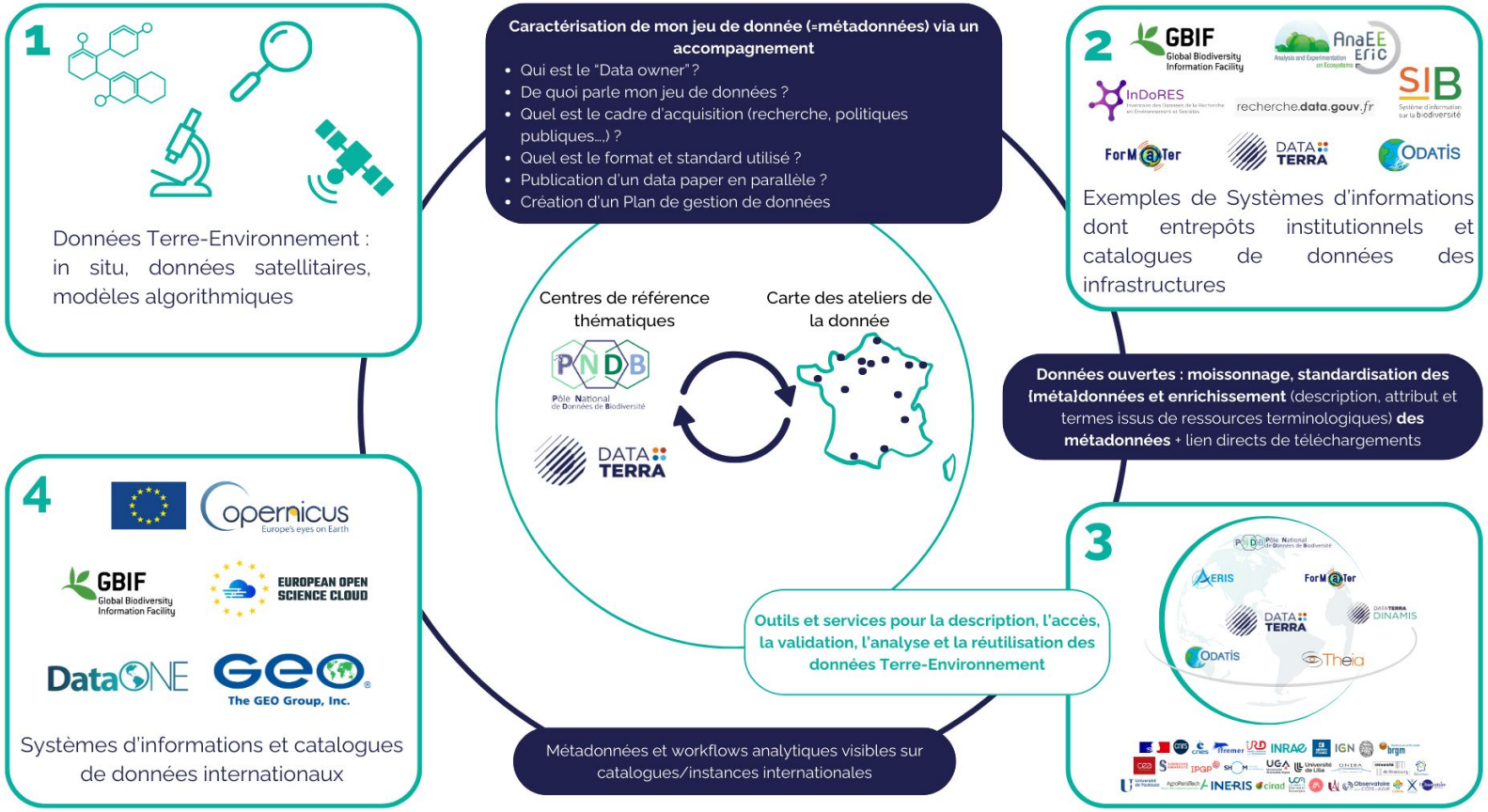
La gouvernance du PNDB, a validé une liste de métadonnées indispensables afin d'avoir un degré de "FAIRitude" (cf. principes FAIR) minimal et relativement élevé

- **Données ouvertes** (CC-BY 4.0 compatible Etalab)
- **Licence obligatoire**
- **Lien direct de téléchargement** des jeux de données brutes
- **Périmètre thématique** (Toute la biodiversité y compris la paléo- et archéo-biodiversité)
- **Périmètre géographique** (Données produites par la France)
- **Couverture temporelle** (à minima une date d'acquisition de données)
- **Résumé**
- **Titre, auteurs et contacts**
- **Cadre d'acquisition** (a minima via un champ texte)
- **DOI / identifiants uniques**
- **couverture taxonomique** (si présence de taxons)
- **mots clés en lien avec Thesaurus**
- **Attributs des données** (Dictionnaire des attributs de données avec unités et descriptions)
- **Annotation sémantique** (Mots-clés et noms d'attributs, ressources utilisables illimitées)

PNDB Pôle National de Données de Biodiversité

Ces informations minimales sont donc communes à toutes les fiches de métadonnées qui seront présentes dans le catalogue du PNDB

flux et stock de {méta}données, pour et par les communautés



FLUX DES DONNÉES & MÉTADONNÉES DE BIODIVERSITÉ

Données issues d'un projet de recherche

Entrepôts thématiques - institutionnels

Métadonnées (format EML) *flux à mettre en place*

Portail Gaia Data

Fiches métadonnées système Terre-Environnement & Climat (format GeoDCAT / STAC) *flux à mettre en place*

Pôle National de Données de Biodiversité - PNDB

données compatibles GBIF (formats DwC et EML) *flux en cours de mise en place*

Portail GBIF

Données occurrence (format SINP / DwC) *flux fonctionnel*

Données issues de l'inventaire national du patrimoine naturel

ou

Plateformes régionales SINP

DepoBio

données & Métadonnées (format SINP) *flux fonctionnel*

Système d'Information nationale du patrimoine naturel - SINP / INPN

Fiches métadonnées (format SINP / GeoNetwork) *flux fonctionnel*

Données issues des politiques publiques

SI métiers du SIB

Métadonnées (format GeoNetwork) *flux en cours de mise en place*

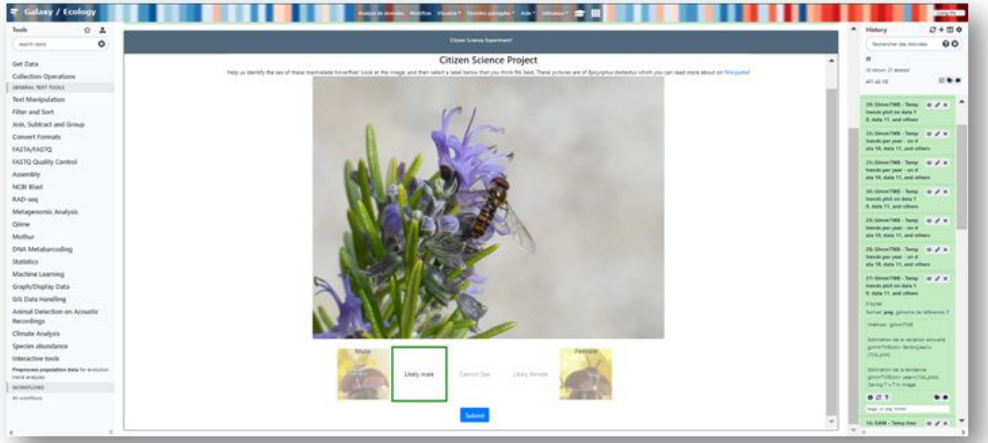
Fiches métadonnées par thématiques ex. aires protégées & cadres ex. EBVs / DPSIR (format EML / GeoNetwork) *flux à mettre en place*

Système d'Information sur la Biodiversité - SIB

issue de et adapté https://mnhn.hal.science/mnhn-04296424/document

Annotations et crowdsourcing : *travaux préliminaires sous Galaxy*

Crowdsourcing image



VIGIENATURE
Un réseau de citoyens qui fait avancer la science

Gamification of participatory science for training and education purposes

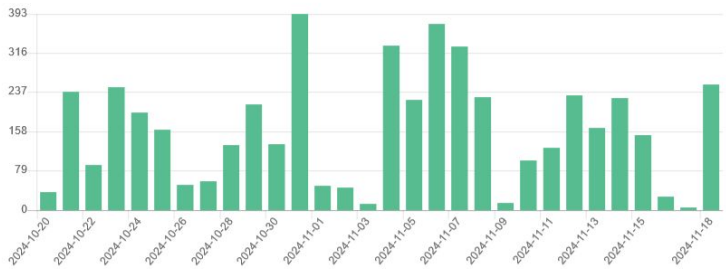
repo

Production / SpiPoll Fly Logout

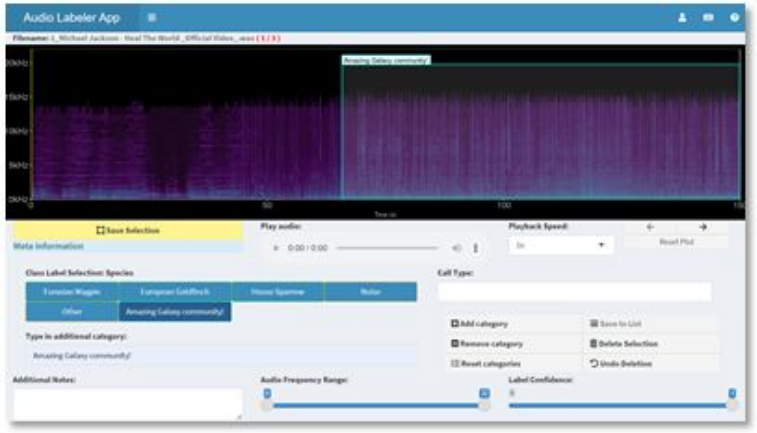
Overview
All time

Batches	Active tasks	Training tasks	Average run	Players	Classifications
6	4.76k	400	3.65	2.00	160.65k

Classifications
Last 30 days

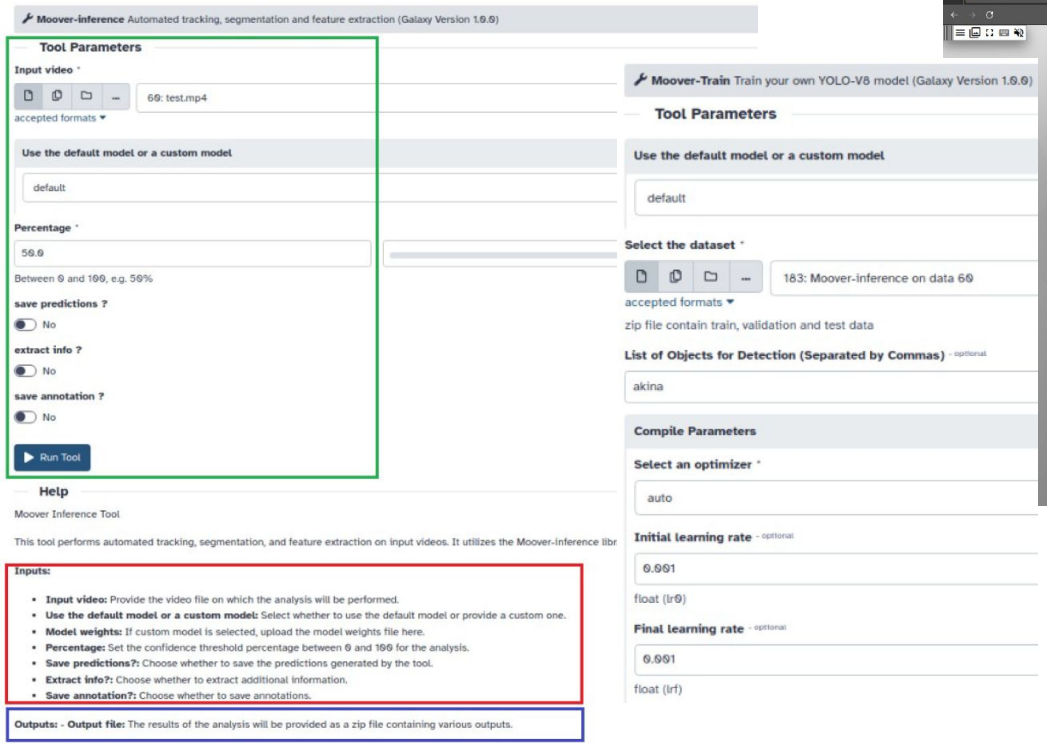


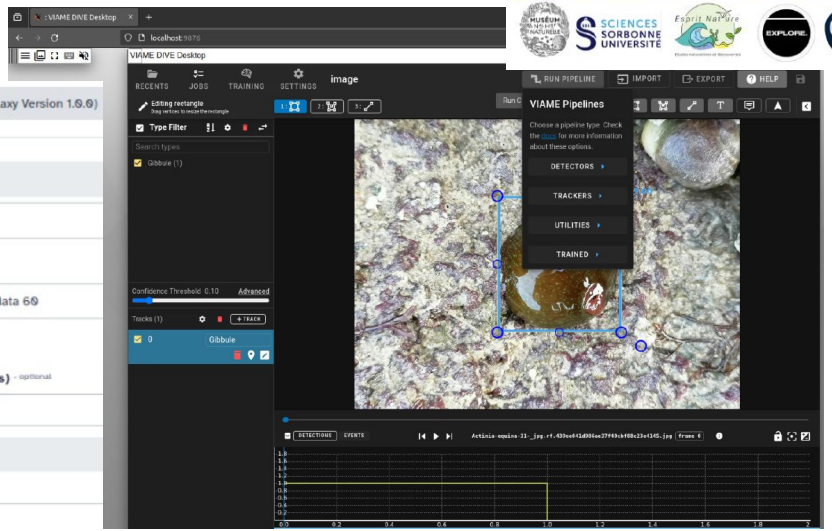
>3400 classifications par mois / >110 par jour



Annotation sons

Annotations et IA : *travaux préliminaires sous Galaxy*



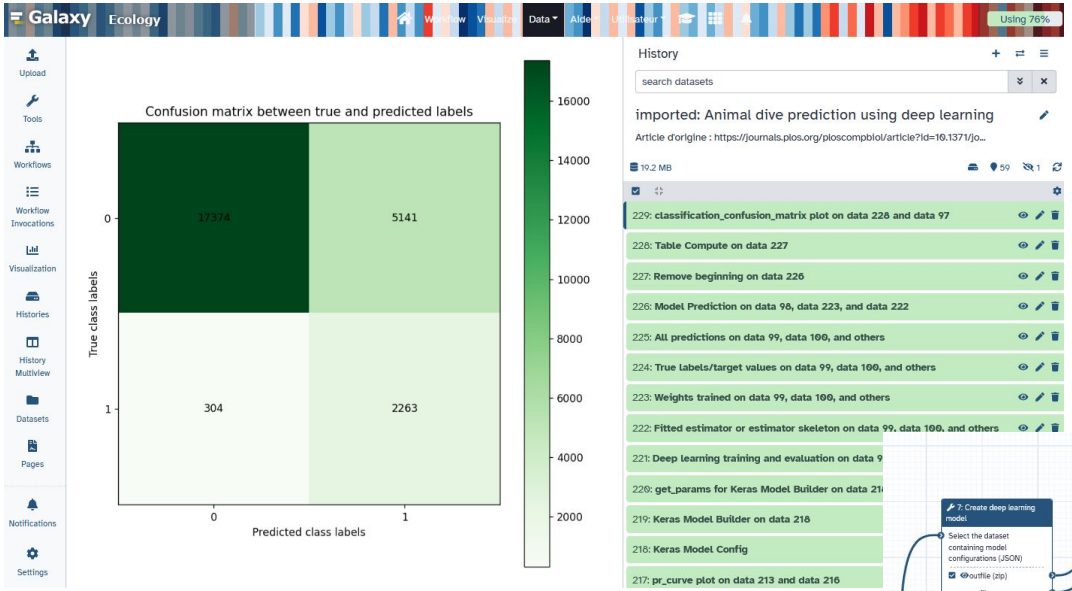


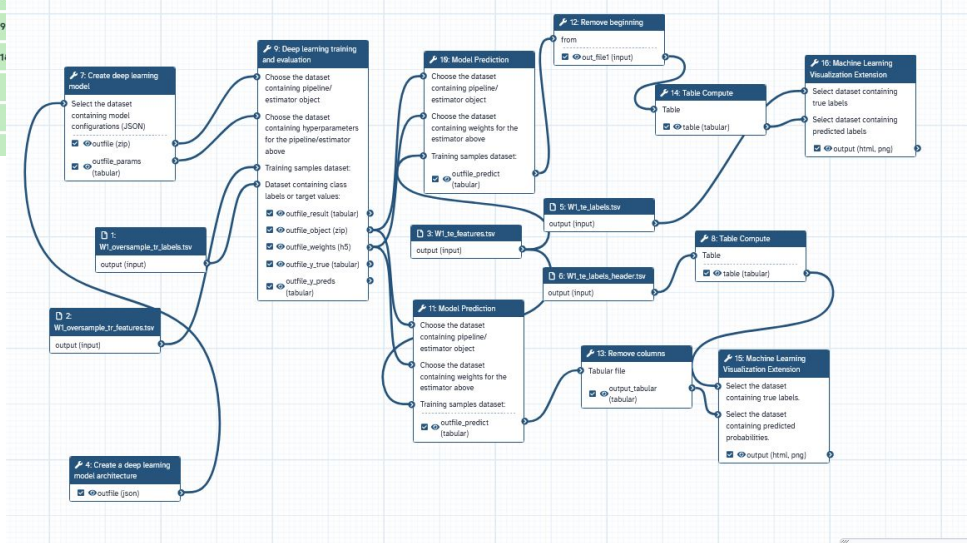
Annotation image ou vidéos

Automated segmentation / tracking / feature extraction

Entrainement

Annotations et IA : *travaux préliminaires sous Galaxy*





PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS PEER-REVIEWED

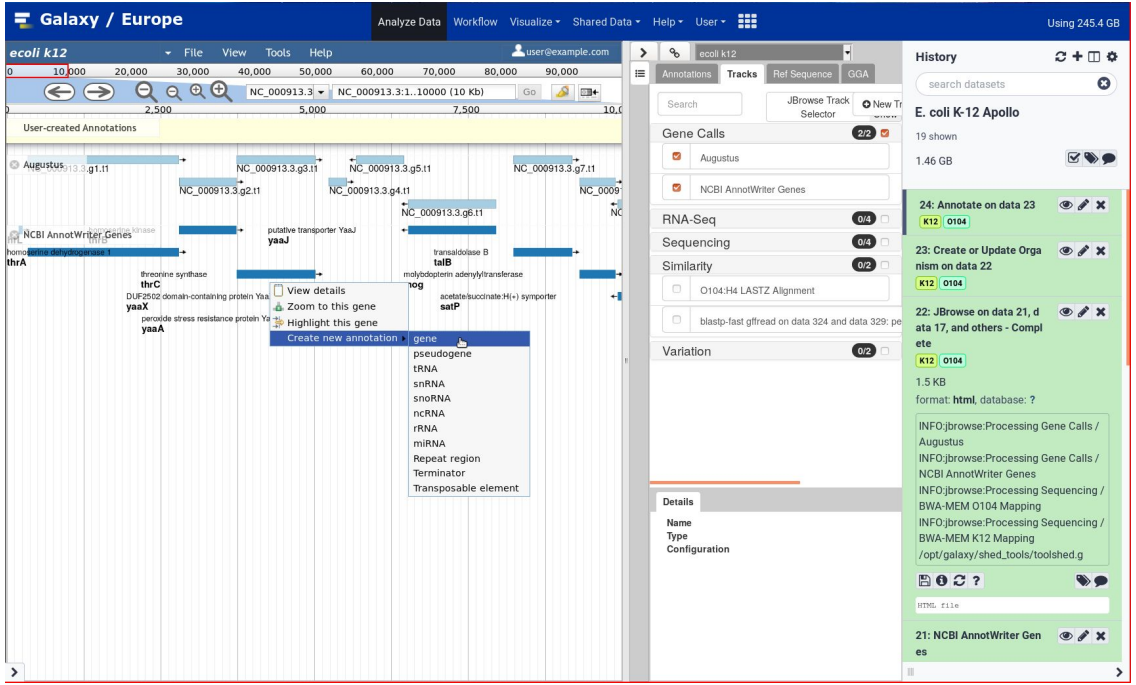
RESEARCH ARTICLE

Deep inference of seabird dives from GPS-only records: Performance and generalization properties

Amédée Roy, Sophie Lanco Bertrand, Ronan Fablet

Version 2 Published: March 11, 2022 • https://doi.org/10.1371/journal.pcbi.1009890

Annotations collaboratives : *Galaxy Genome Annotation as an example*

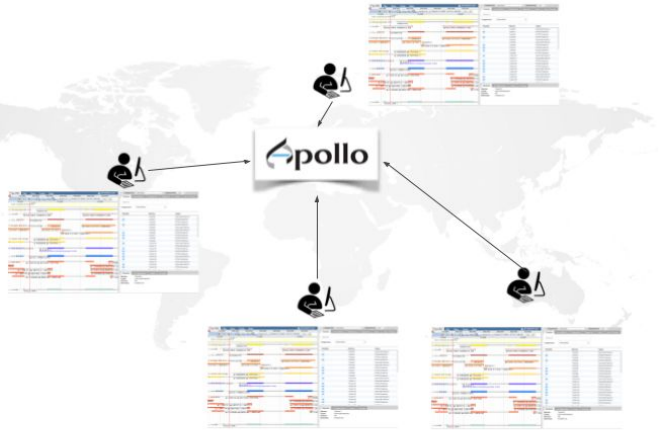


Using Galaxy, any user can upload its own organism, and start curating gene models collaboratively with other users.

Check out the GTN training material to learn how to use it!

GGA

Galaxy Genome Annotation



Classical “workflow”

1. Fetch Data
2. Analyse in Galaxy
3. Send to Apollo
4. Collaboratively Annotate → repeat