The French Biodiversity data hub (PNDB) and the Galaxy-E project: A web platform for collaborative data analysis in Ecology





Yvan Le Bras

yvan.le-bras@mnhn.fr

@Yvan2935

Station de Biologie marine Concarneau

<°))))><

—

Jean-Denis Vigne

Laurent Poncet

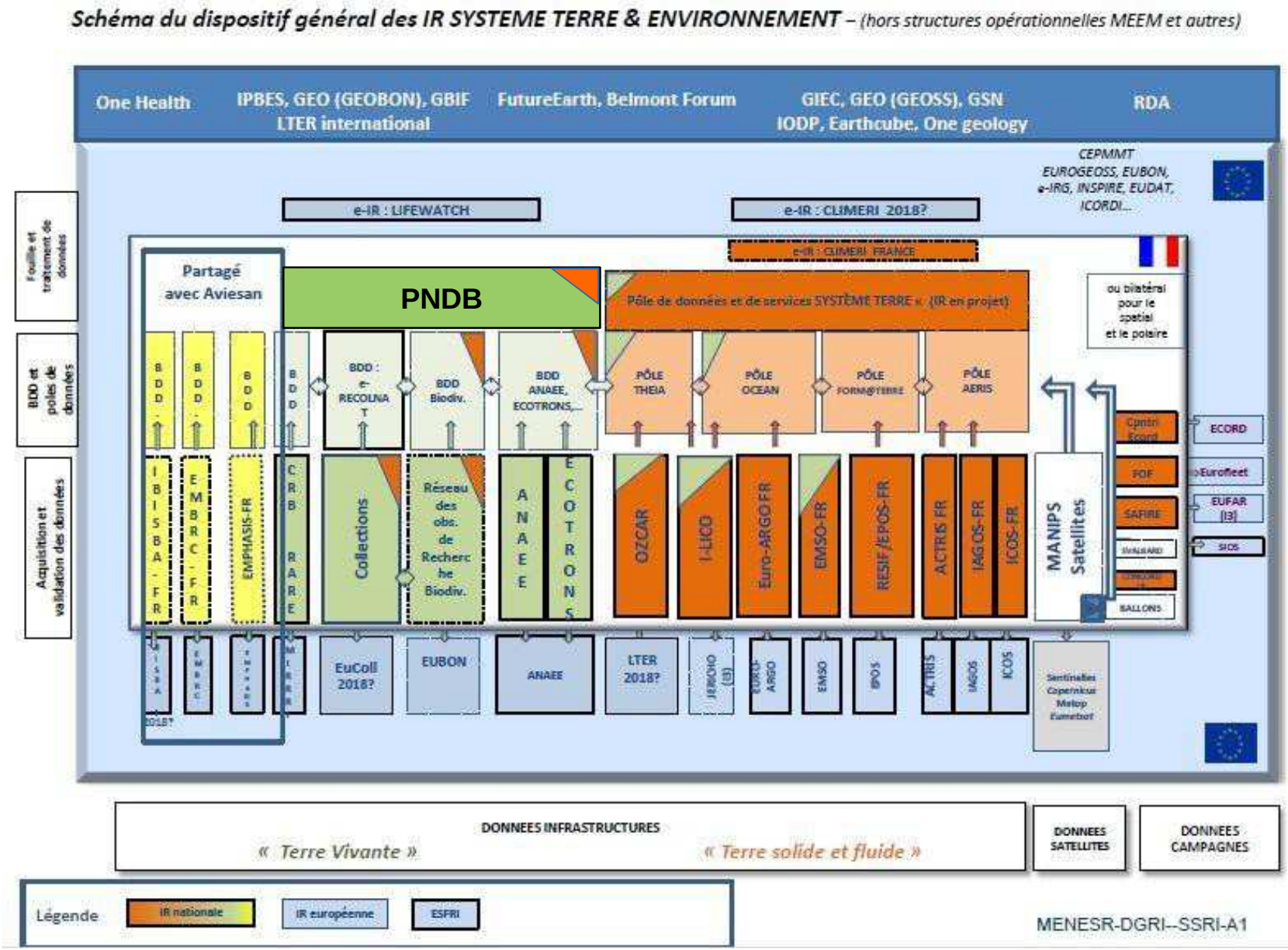
UMS 2006 PatriNat MNHN/CNRS/AFB

Jardin des plantes 75005 Paris

CONTEXT

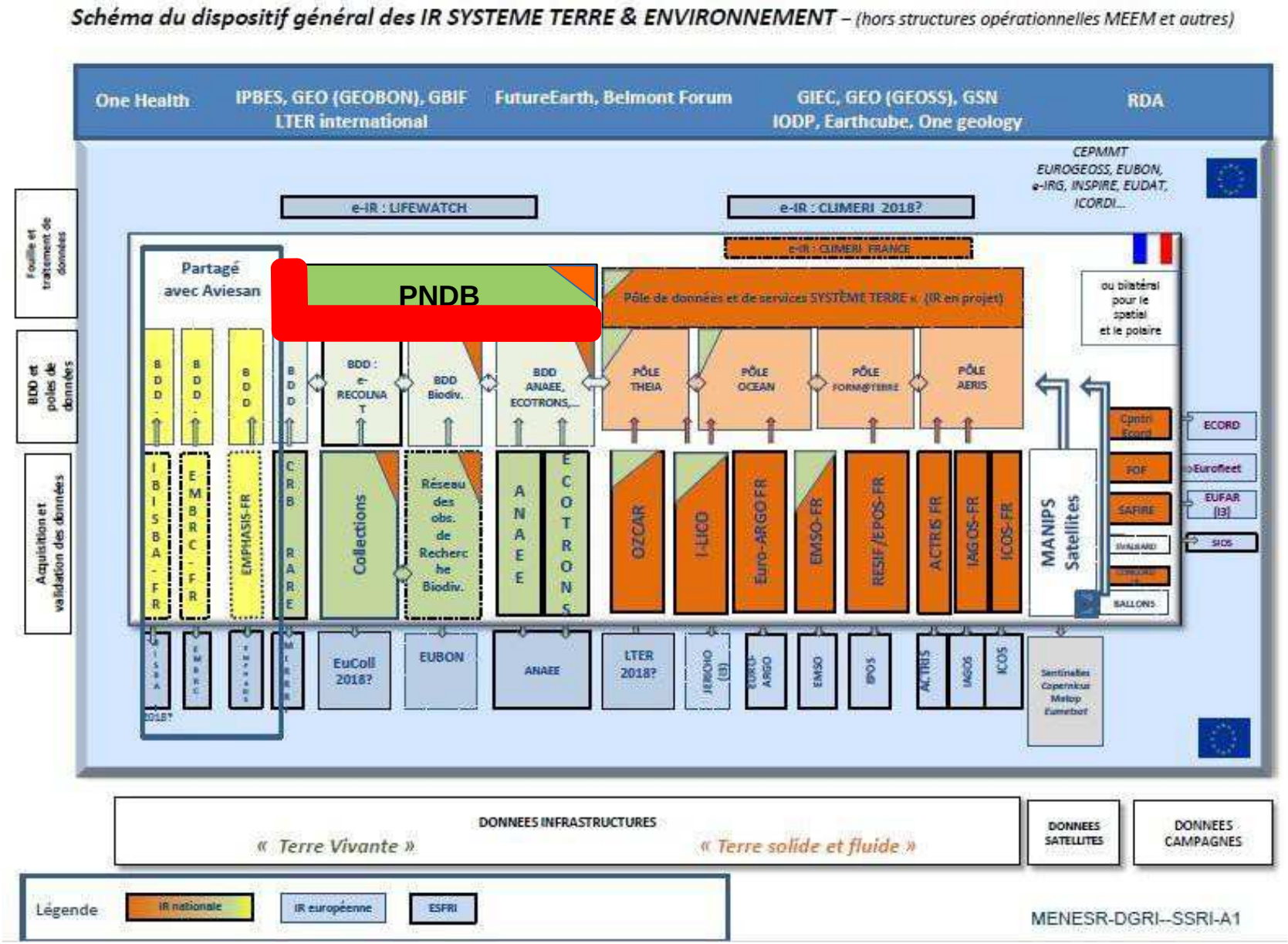
« Pôle national de données de biodiversité »: French Research Infrastructure dedicated to Biodiversity

PNDB ecosystem : A national infrastructure point of view



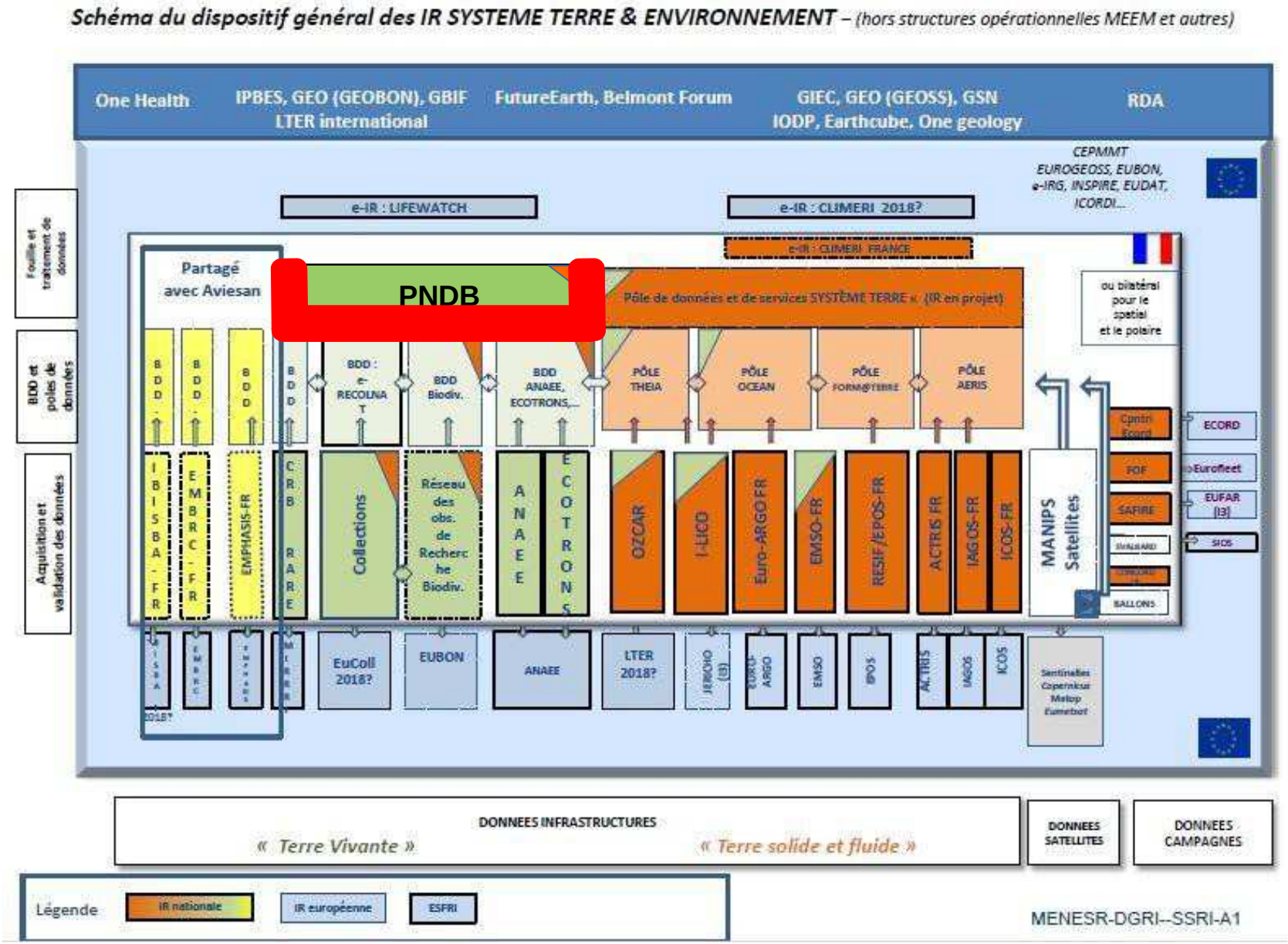
PNDB ecosystem : A national infrastructure point of view

« Biodiv » Interface



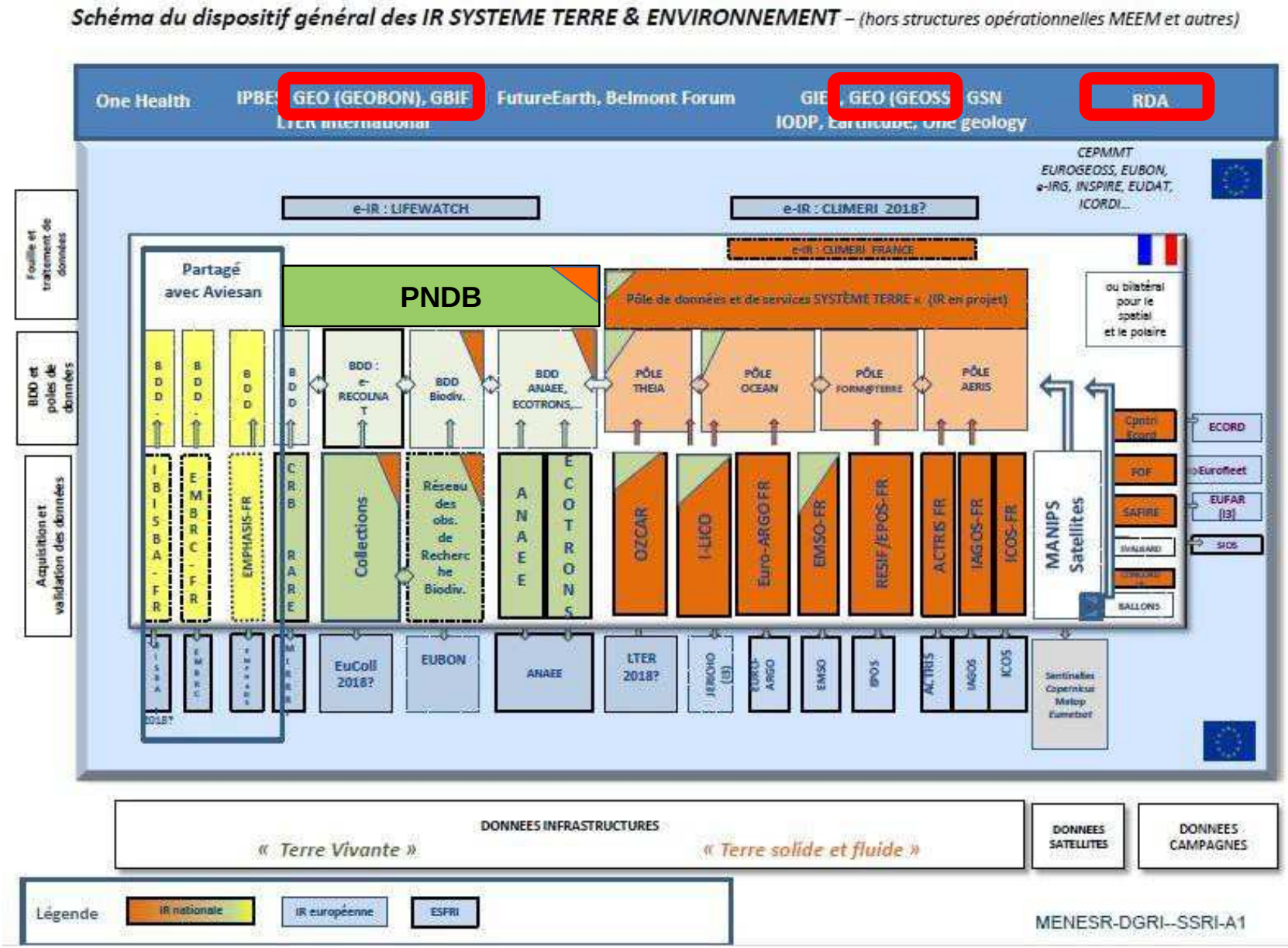
PNDB ecosystem : A national infrastructure point of view

« Earth » interface



PNDB ecosystem : A national infrastructure point of view

International links



PNDB ecosystem : A national infrastructure point of view

International links

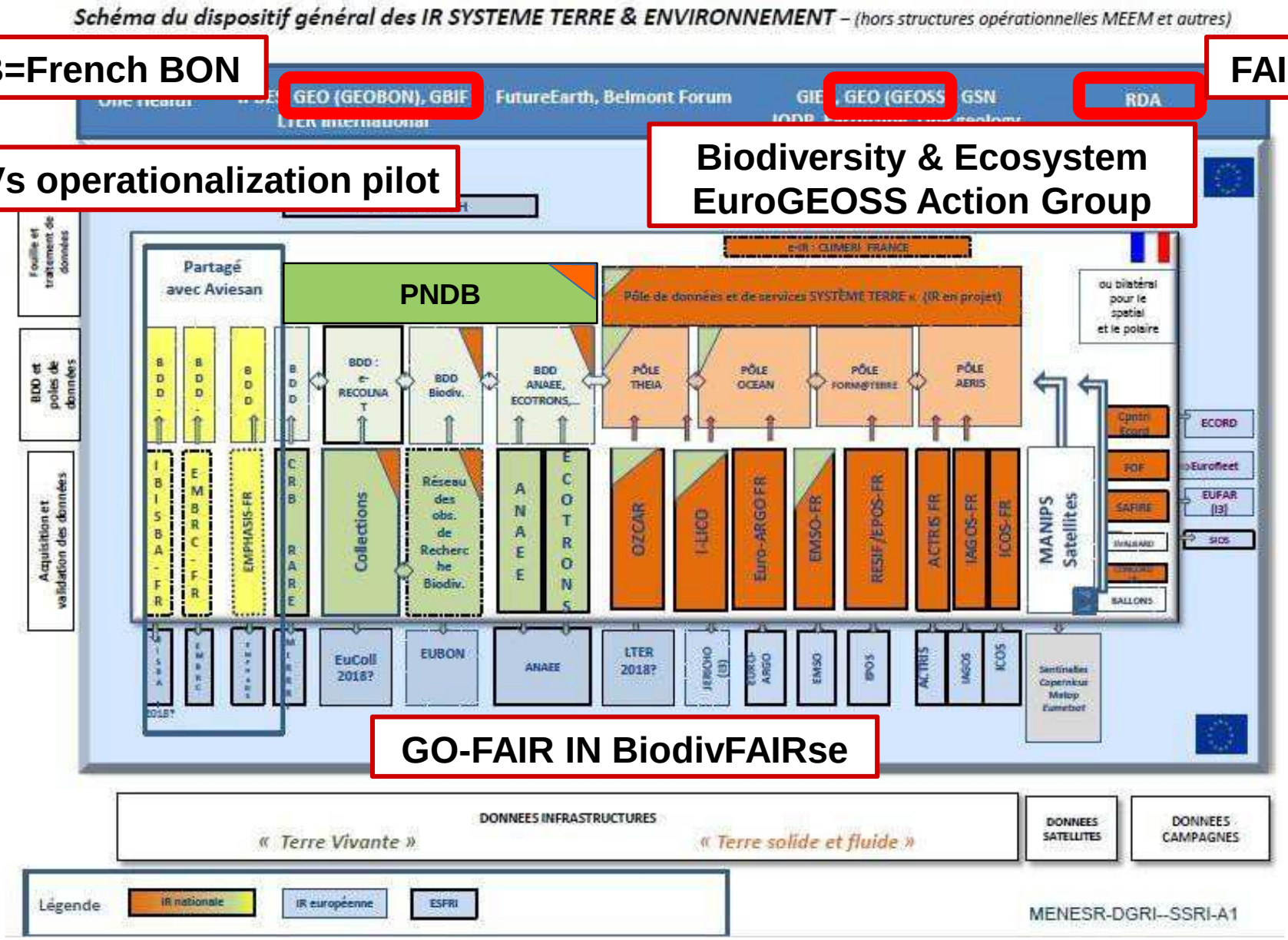
PNDB=French BON

FAIR WGs

EBVs operationalization pilot

Biodiversity & Ecosystem EuroGEOSS Action Group

GO-FAIR IN BiodivFAIRse



PNDB an e-infrastructure point of view

- Promote existing services & products

Data

Metadata

- Provide new services

Metadata Agregation

Analyses

PNDB an e-infrastructure point of view

- Promote existing services & products

Data

Metadata

- Provide new services

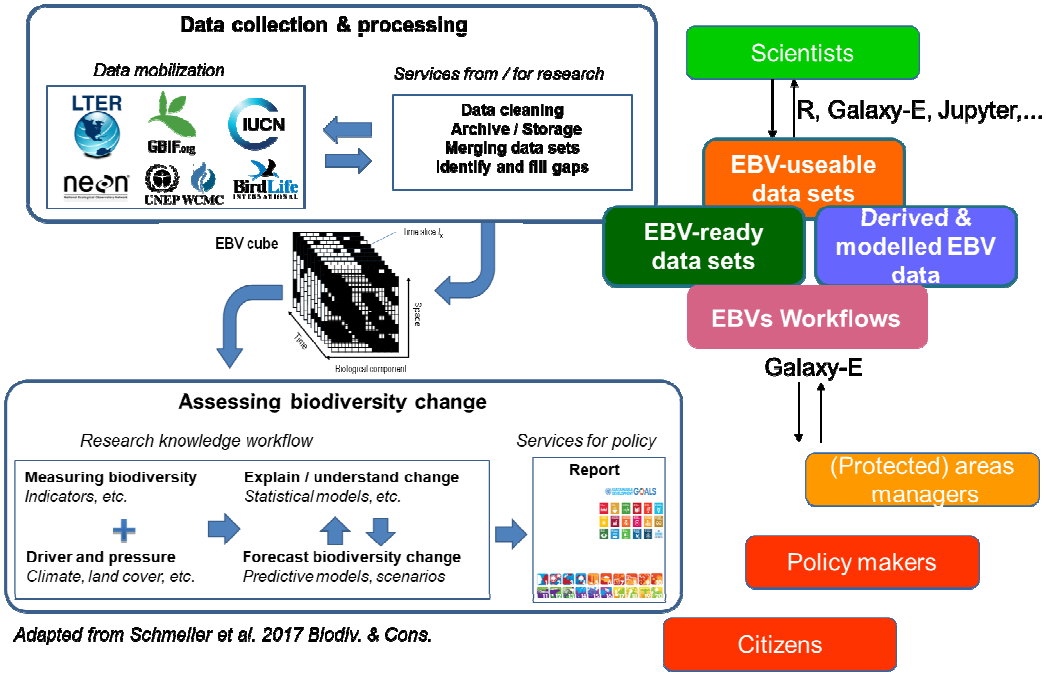
« FAIR »

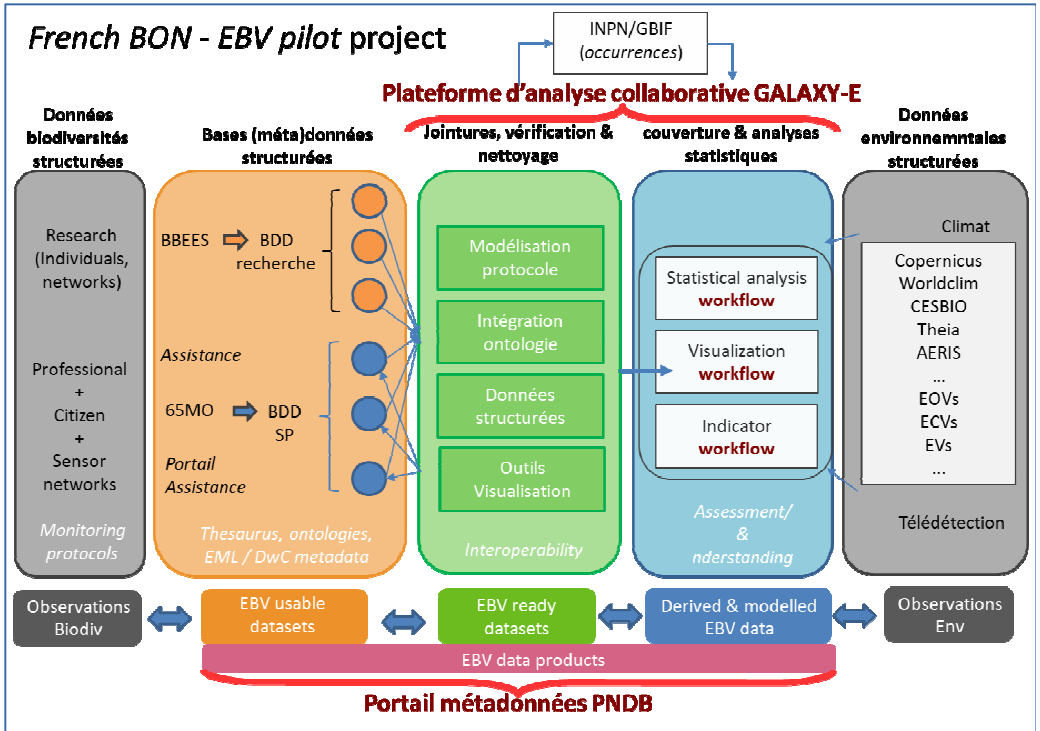
Metadata Agregation

Analyses

Ecology FAIRization following 2 levels

- **A necessary first global FAIRization**
  - Giving access of maximum of transparent and reusable data and related research digital objects (algorithms, tools, workflows, protocols…) to a wide range of users
- **A wonderfull opportunity of « gold standard» FAIRization**
  - Using EBVs concept and biodiversity indicators production!

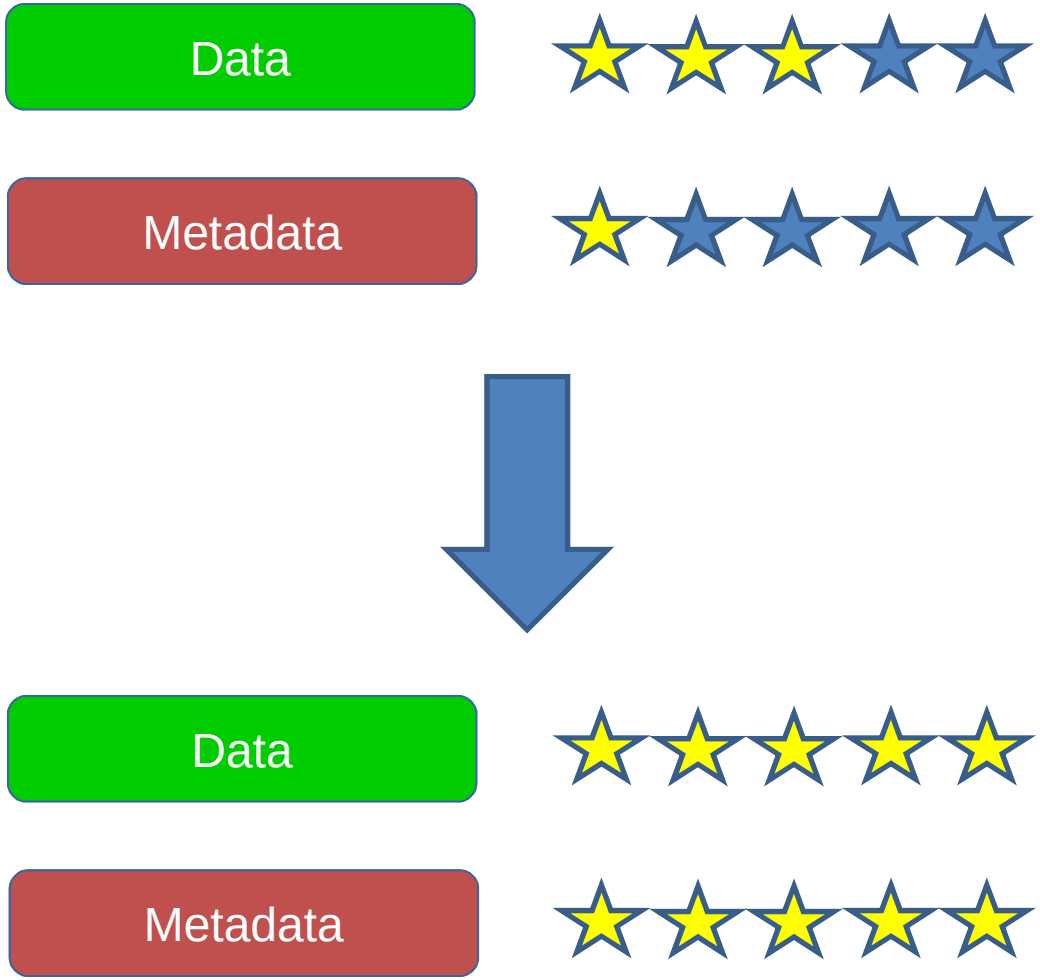




Biodiversity data and FAIR key challenges

Main Difficulty

- **Associate structured and detailled metadata to biodiversity data**



Main Difficulty

- **Associate structured and detailled metadata to biodiversity data**

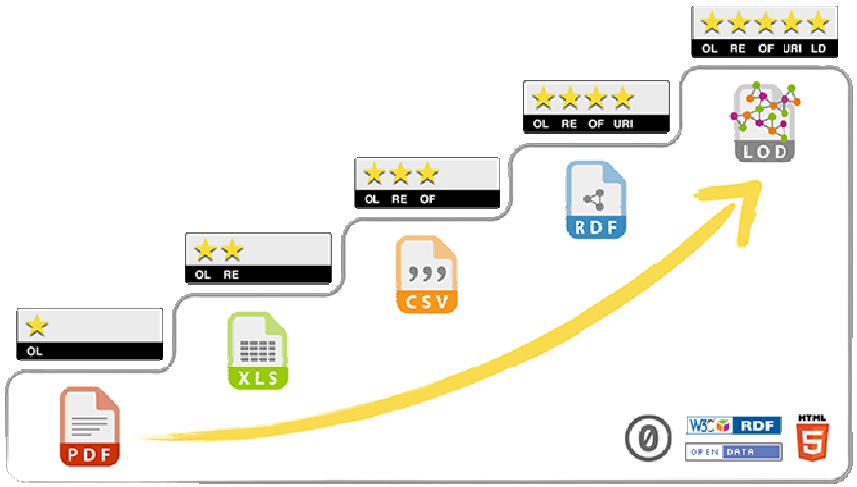
Data

Metadata

- Open Data
- Linked data

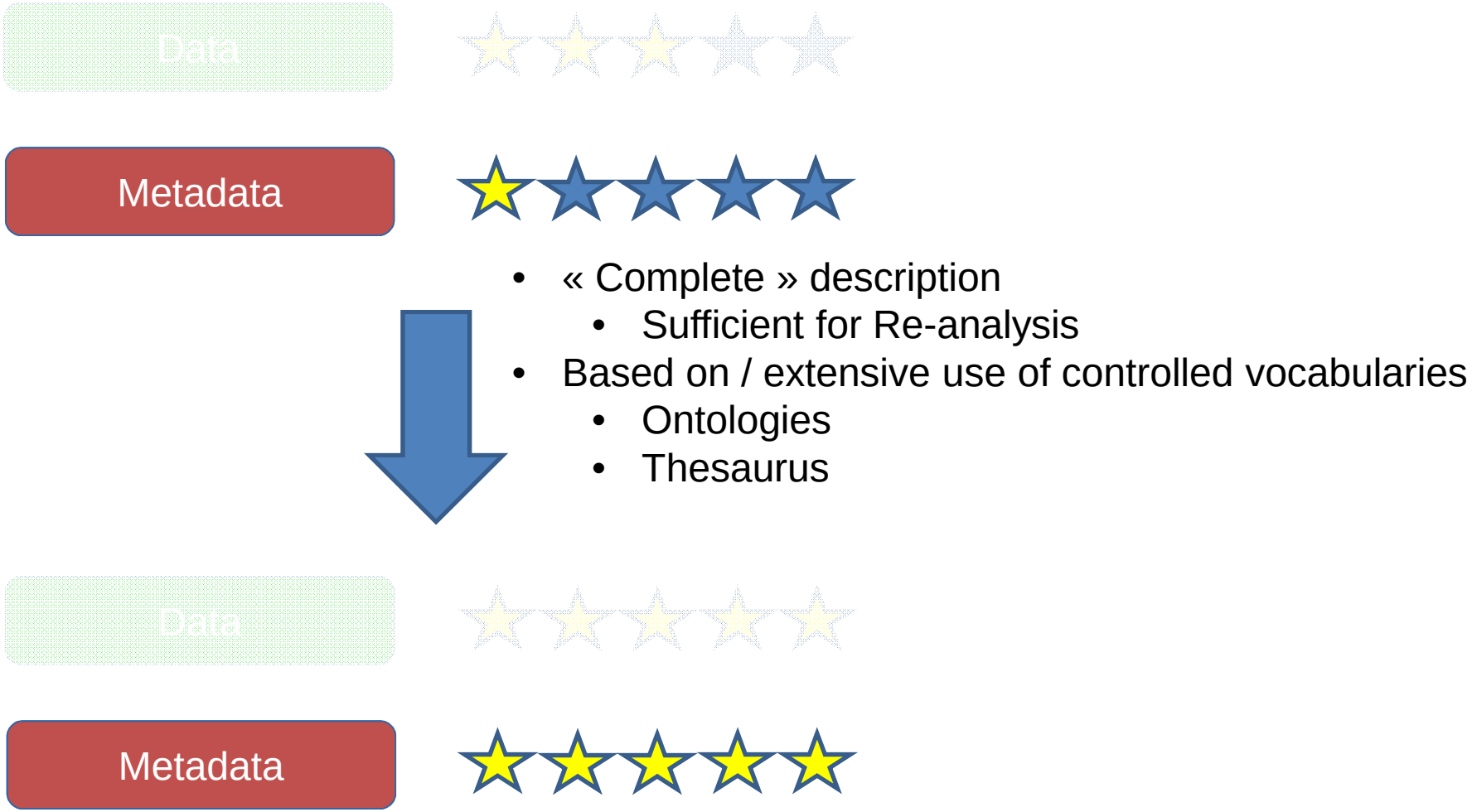
Data

Metadata



Main Difficulty

- **Associate structured and detailled metadata to biodiversity data**



- « Complete » description
  - Sufficient for Re-analysis
- Based on / extensive use of controlled vocabularies
  - Ontologies
  - Thesaurus

Main Lack

- **Open-source and freely available service in a workflow-oriented e-infrastructure that supports curation, sharing and collaboration**

(Gärdenfors *et al.*, 2014; Enquist *et al.*, 2016; Hardisty *et al.*, 2016; La Salle *et al.*, 2016; Hugo *et al.*, 2017, Kissling *et al.*, 2017)

Software/algorithm accessibility ?

Free access ?

Transparency ?

Reproductibility ?

Flexibility on use (local/HPC/Cloud)?

No INSPIRE or webservice-based ?

Easy to use ?

Easy to contribute ?

User communities ?

Contributors communities ?

Main Lack

- **Open-source and freely available service in a workflow-oriented e-infrastructure that supports curation, sharing and collaboration**
  (Gärdenfors *et al.*, 2014; Enquist *et al.*, 2016; Hardisty *et al.*, 2016; La Salle *et al.*, 2016; Hugo *et al.*, 2017, Kissling *et al.*, 2017)

Software/algorithm accessibility ?

Free access ?

Transparency ?

Reproductibility ?

Flexibility on use (local/HPC/Cloud)?

No INSPIRE or webservice-based ?

Easy to use ?

Easy to contribute ?

User communities ?

Contributors communities ?

Coordination OpenAIRE Connect
Open Science as a Service (OSaaS)

i-marine
Data e-Infrastructure Initiative for Fisheries Management and Conservation of Marine Living Resources

BlueBRIDGE

ENVRI Common Operations of Environmental Research Infrastructures

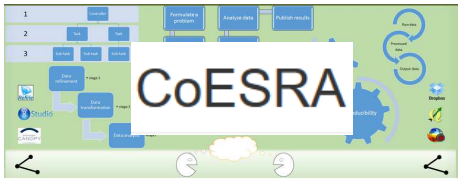
EUBrasil OpenBio

AquaMaps

EU FP7



EU H2020



Australia 2018

Main Lack

- **Open-source and freely available service in a workflow-oriented e-infrastructure that supports curation, sharing and collaboration**
(Gärdenfors *et al.*, 2014; Enquist *et al.*, 2016; Hardisty *et al.*, 2016; La Salle *et al.*, 2016; Hugo *et al.*, 2017, Kissling *et al.*, 2017)

Software/algorithm accessibility ?

Free access ?

Transpare

odu y ?

Flexibility or al/HPC/Cloud)?

No INSPI rvice-based ?

asy to

asy to contri

User communities ?

Contributors communities ?

Coordination OpenAIRE Connect
Open Science as a Service (OSaaS)

iMarine
Data e-Infrastructure Initiative for Fisheries Management and Conservation of Marine Living Resources

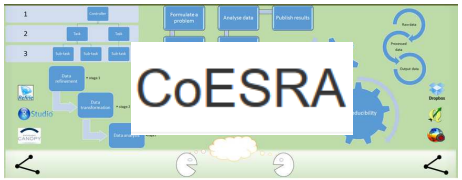
BlueBRIDGE

ENVRI Common Operations of Environmental Research Infrastructures

EUBrasil OpenBio

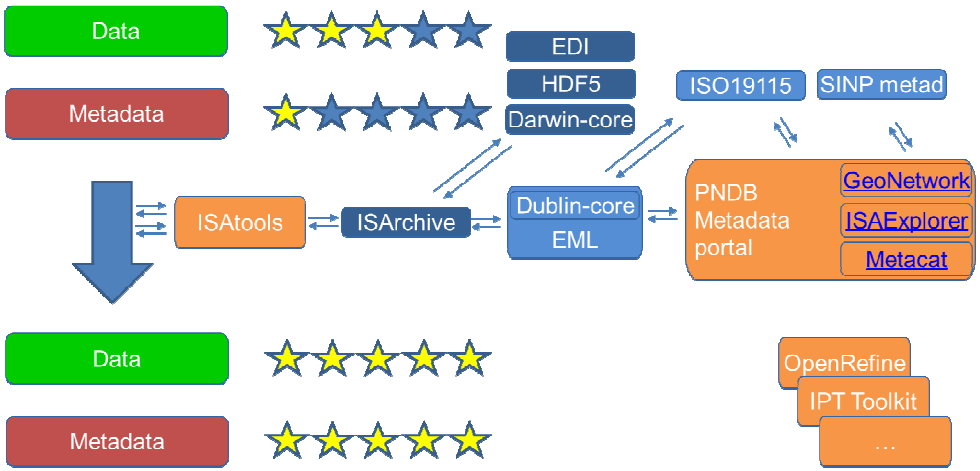
AquaMaps





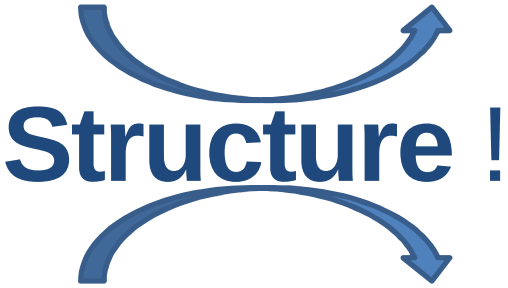
PNDB proposals: give **Access** but also…

Data & metadata



Homemade not sharable & unreproducible R scripts

Packaged tools & dependencies



Metadata poorly filled

Metadata as complete as possible

Private codes

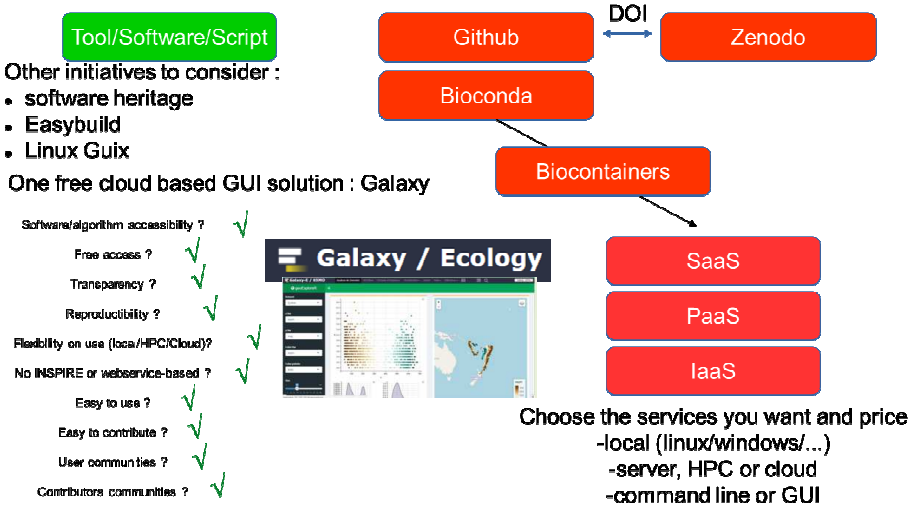
Communities codes



Existing metadata weak

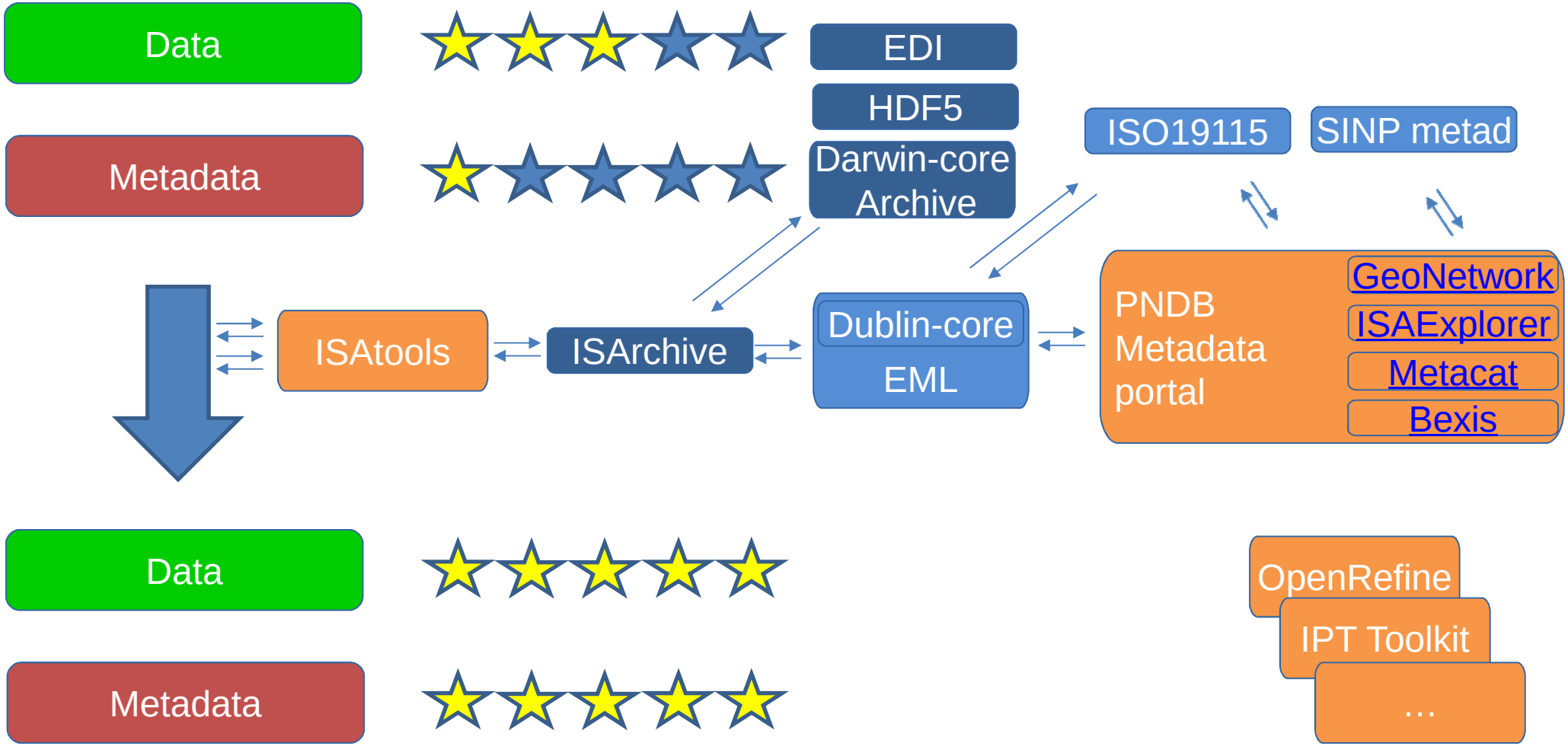
PNDB metadata addition/precision

Analytics tools & platforms



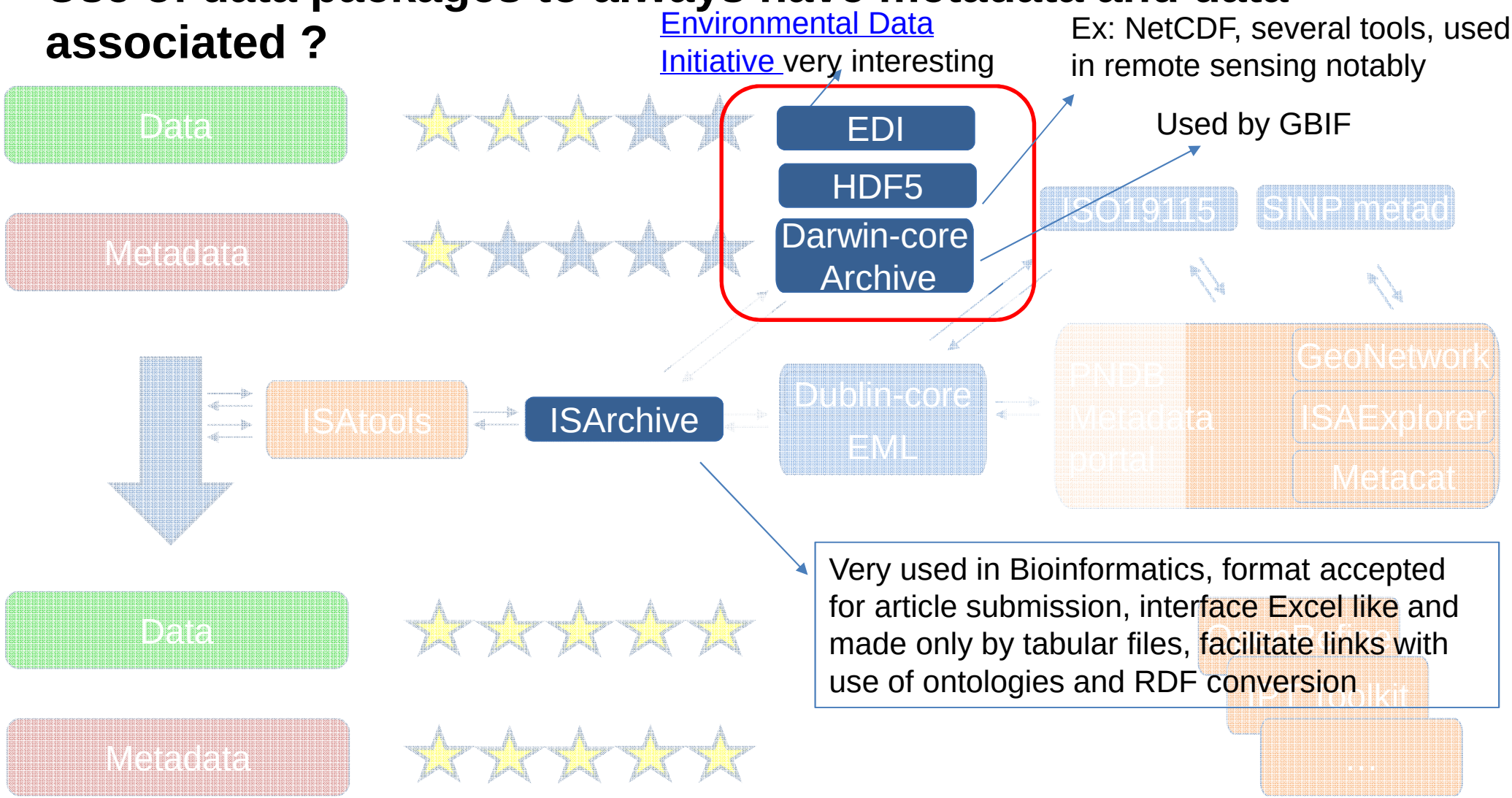
PNDB « data & metadata » proposal

- **Associate structured and detailled metadata to biodiversity data**



PNDB « data & metadata » proposal

- **Use of data packages to always have metadata and data associated ?**



A focus on software accessibility issue

- A comprehensive analysis of the usablity & archival stability of omics computational tools & resources Mangul *et al*., 2018 https://www.biorxiv.org/content/early/2018/10/25/452532
  - 72 hours to install 99 randomly selected tools
  - 28 % of tools impossible to install
  - Only 50 % are easy-to-install / 21 % hard-to-install
  - Conda tools are 100 % installable
  - Majotiry of tools fail automatics installation test
  - Tools possible to install are significantly more cited
  - 24490 URLs of published omics software resources / 26 % broken
    - Github have 20 % of all URLs and only 1 % of broken URLs
  - Take home message
    - Host software on archivally stable services as Github
    - Provide easy-to-use installation interface as Conda and get more citations
    - Tools wrapped as conda are 100 % installable (probably same with easybuild or Guix)

  - Proposal:
    - Host software on archivally stable services as Github + create conda recipes
    - Provide DOI through Zenodo, publish software paper and get more citations

PNDB « analysis tools & platforms » proposal

Tool/Software/Script

Github ⟷ Zenodo (DOI)

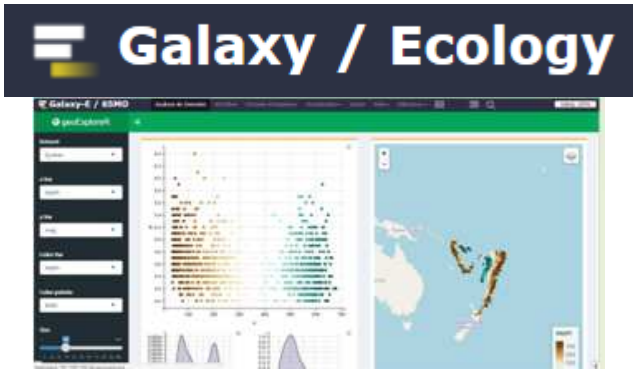
Bioconda

Other initiatives to consider :
- software heritage
- Easybuild
- Linux Guix

Biocontainers

One free cloud based GUI solution : Galaxy

Software/algorithm accessibility ? √
Free access ? √
Transparency ? √
Reproductibility ? √
Flexibility on use (local/HPC/Cloud)? √
No INSPIRE or webservice-based ? √
Easy to use ? √
Easy to contribute ? √
User communities ? √
Contributors communities ? √



SaaS
PaaS
IaaS

Choose the services you want and price
-local (linux/windows/...)
-server, HPC or cloud
-command line or GUI

PNDB « analysis tools & platforms » proposal

A coral reef « PAMPA » proposal

Tool/Software/Script

30 R scripts

- Data selection
- Linear models
- Tracking
- Visualizations

63 packages

Github ⟷ Zenodo (DOI)

Bioconda

Galaxy Toolshed

PAMPA : Indicateurs de la Performance d'Aires Marines Protégées pour la gestion des écosystèmes côtiers, des ressources et de leurs usAges

Galaxy



Galaxy Flavours

OR

Usegalaxy.* instances

Link with Marine BON ?

https://ecology.usegalaxy.eu/

https://ecology.usegalaxy.org.au/

PNDB « analysis tools & platforms » proposal

Data products

« Biodiv » system

« Earth » system

Analytics products

Pre-process

Analysis

Visu



CONDA package

Docker

SLURM

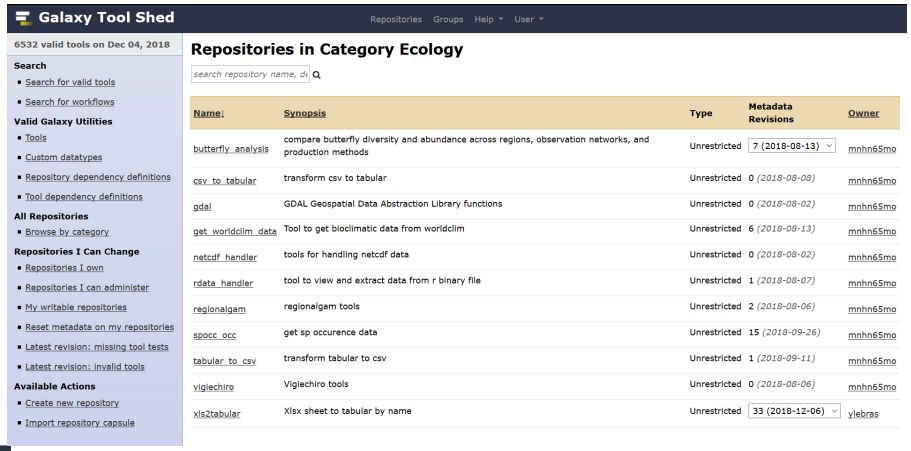
NGINX

High Performance Computing

Cloud

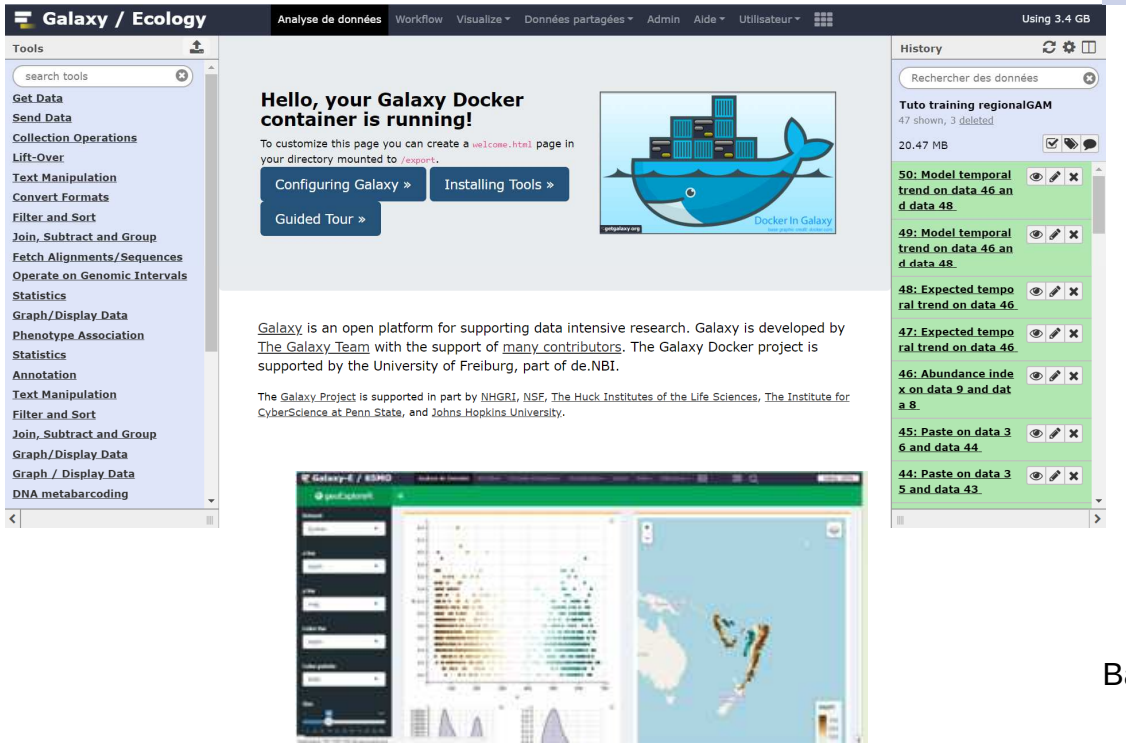
PNDB « analysis tools & platforms » proposal

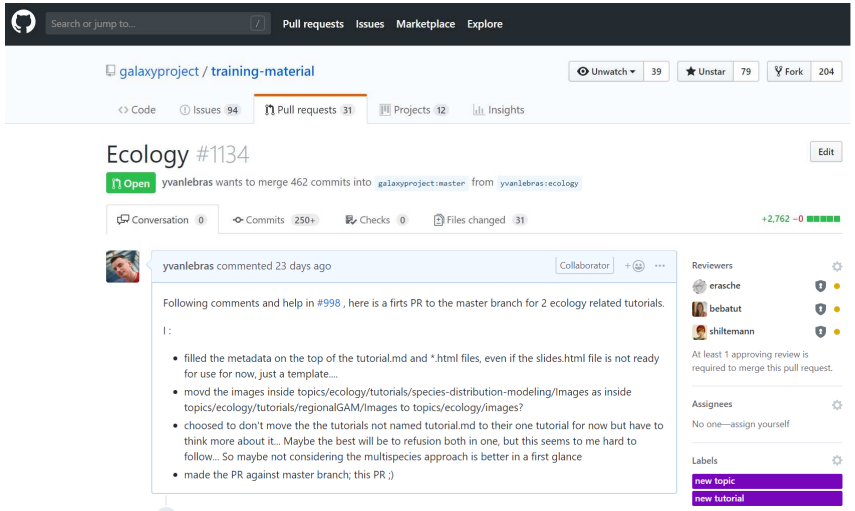
Galaxy « App-store »



Galaxy for Ecology communities (users + contributors)

The **Galaxy-E project**





Batut *et al.*, 2018

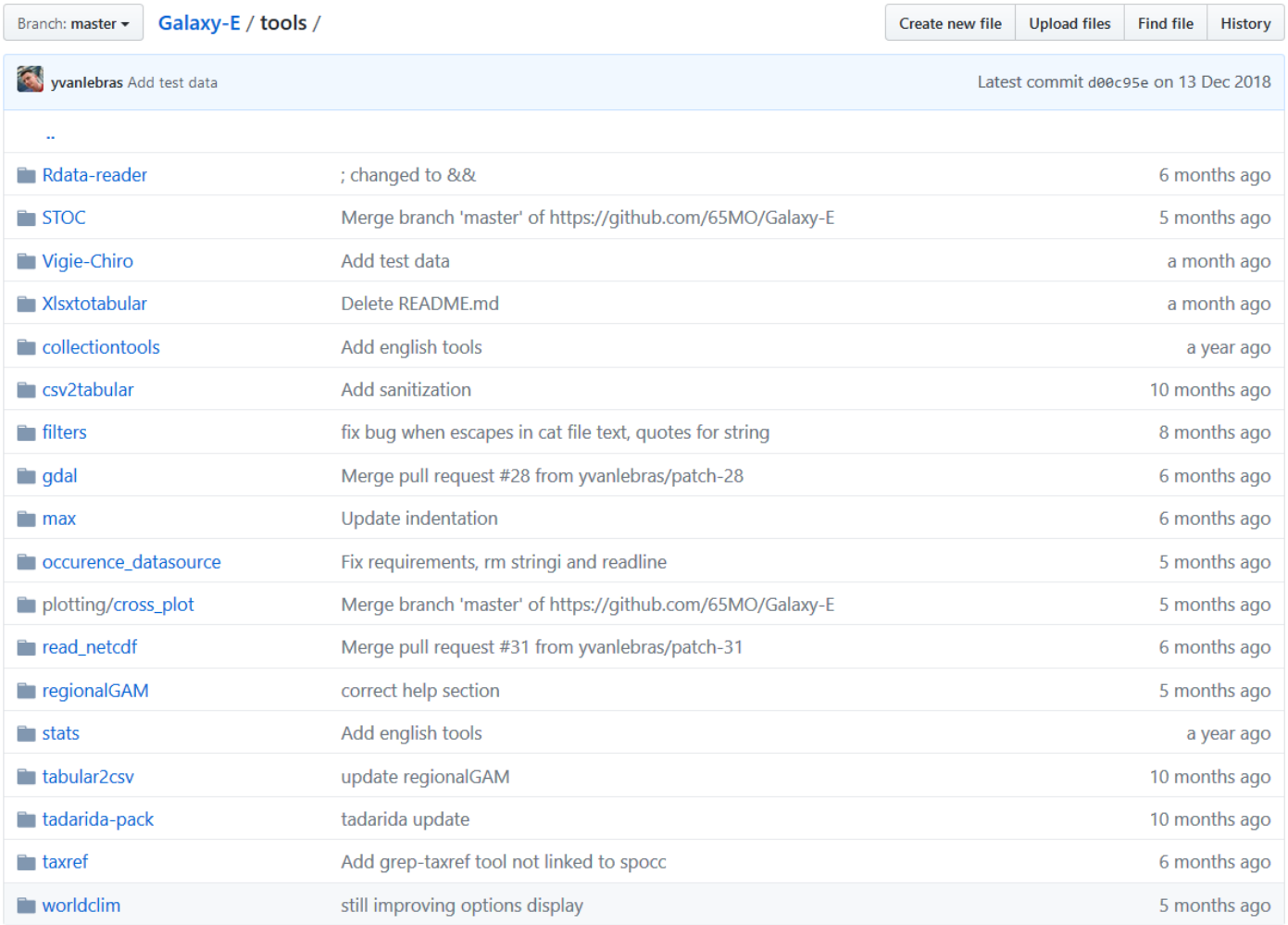
Training as a Service
- Training-material
  - tutorials
  - data, tools & workflows
  - Galaxy instances

PNDB « analysis tools & platforms » proposal

A Galaxy-E dedicated Github repo: https://github.com/65MO/Galaxy-E

27 new Galaxy tools

- Rdata
- Netcdf
- GDAL
- CSV
- Bat accoustic
- GBIF/ALA/… import
- Worldclim data import



Branch: master ▾ Galaxy-E / tools / Create new file | Upload files | Find file | History

yvanlebras Add test data — Latest commit d00c95e on 13 Dec 2018

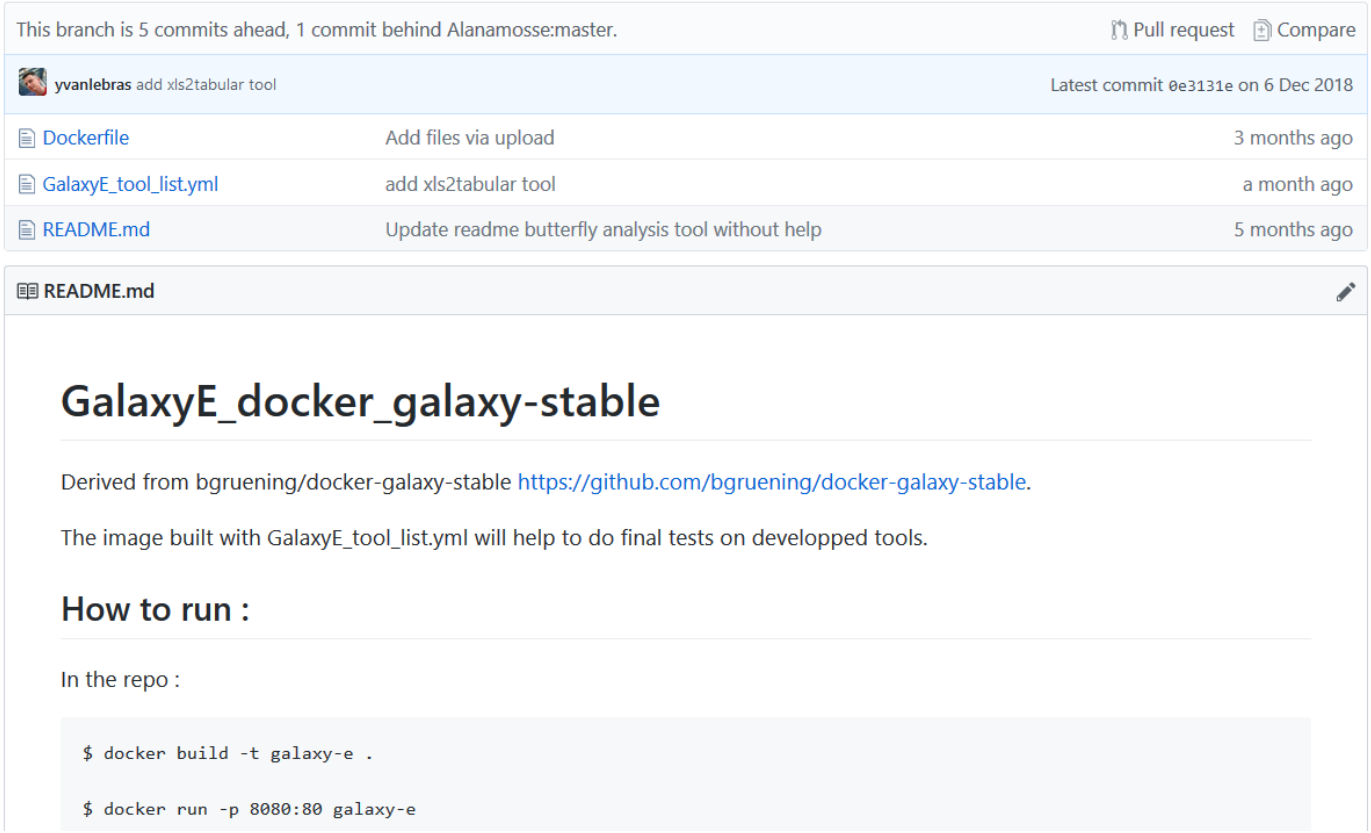
..		
Rdata-reader	; changed to &&	6 months ago
STOC	Merge branch 'master' of https://github.com/65MO/Galaxy-E	5 months ago
Vigie-Chiro	Add test data	a month ago
Xlsxtotabular	Delete README.md	a month ago
collectiontools	Add english tools	a year ago
csv2tabular	Add sanitization	10 months ago
filters	fix bug when escapes in cat file text, quotes for string	8 months ago
gdal	Merge pull request #28 from yvanlebras/patch-28	6 months ago
max	Update indentation	6 months ago
occurence_datasource	Fix requirements, rm stringi and readline	5 months ago
plotting/cross_plot	Merge branch 'master' of https://github.com/65MO/Galaxy-E	5 months ago
read_netcdf	Merge pull request #31 from yvanlebras/patch-31	6 months ago
regionalGAM	correct help section	5 months ago
stats	Add english tools	a year ago
tabular2csv	update regionalGAM	10 months ago
tadarida-pack	tadarida update	10 months ago
taxref	Add grep-taxref tool not linked to spocc	6 months ago
worldclim	still improving options display	5 months ago

PNDB « analysis tools & platforms » proposal

A Galaxy-E Docker image: https://github.com/65MO/GalaxyE_docker_galaxy-stable

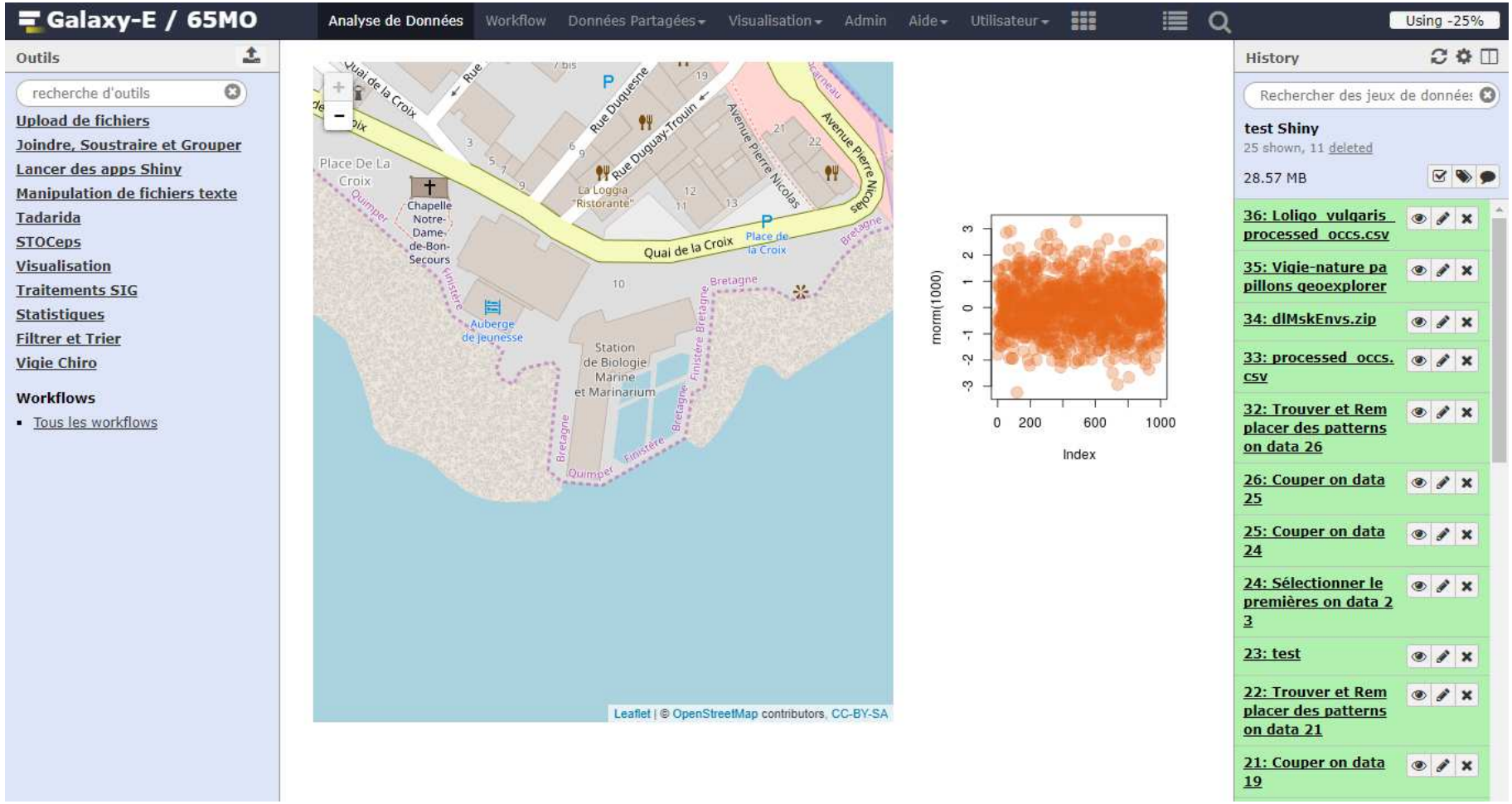
27 new Galaxy tools
- Rdata
- Netcdf
- GDAL
- CSV
- Bat accoustic
- GBIF/ALA/… import
- Worldclim data import

81 existing Galaxy tools
- Text Manipulation
- Filter & Sort
- Graph / Display data
- Statistics
- Metagenomics
- Metabarcoding
- …



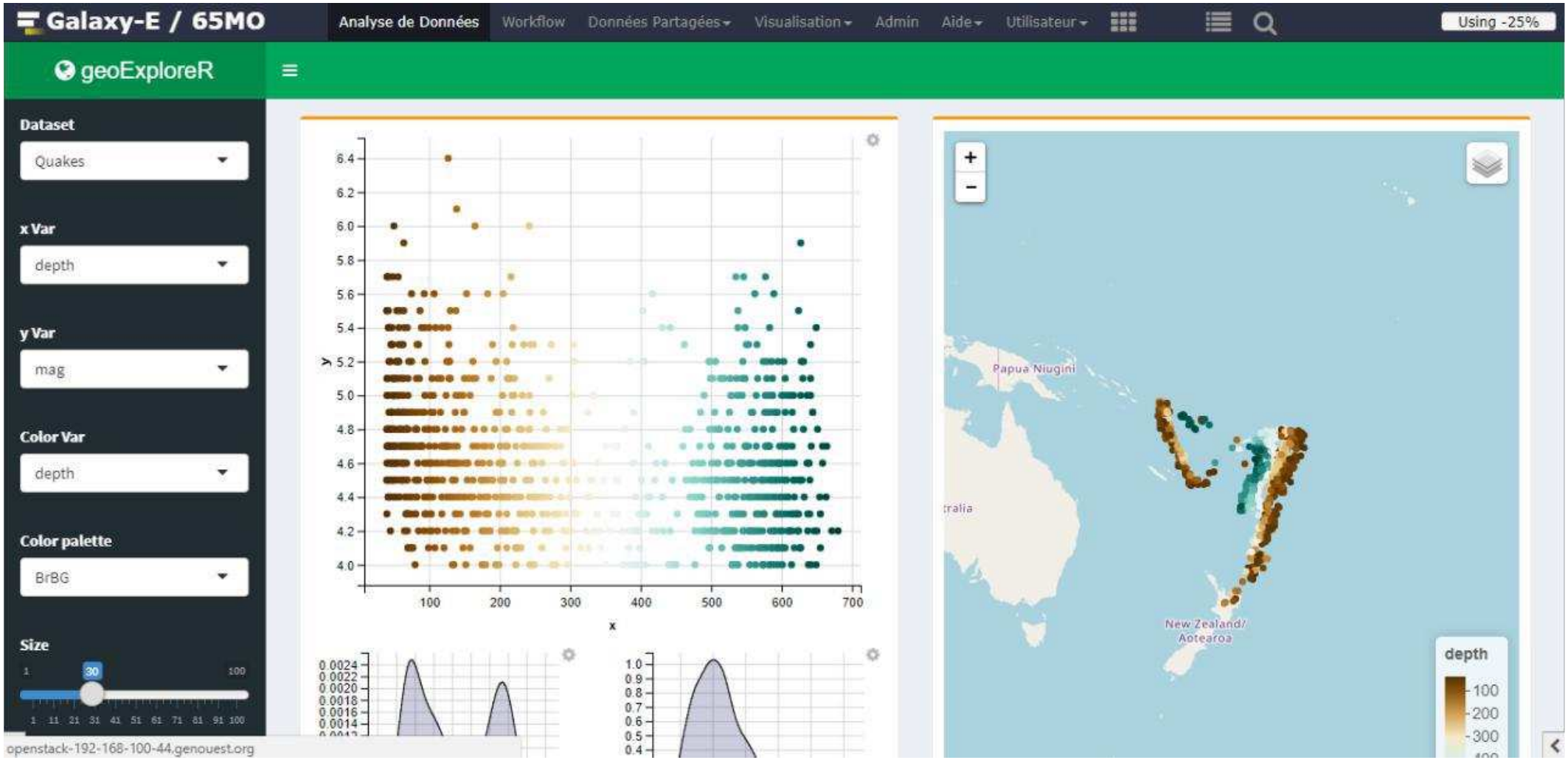
PNDB « analysis tools & platforms » proposal

R Shiny apps integration: GIS



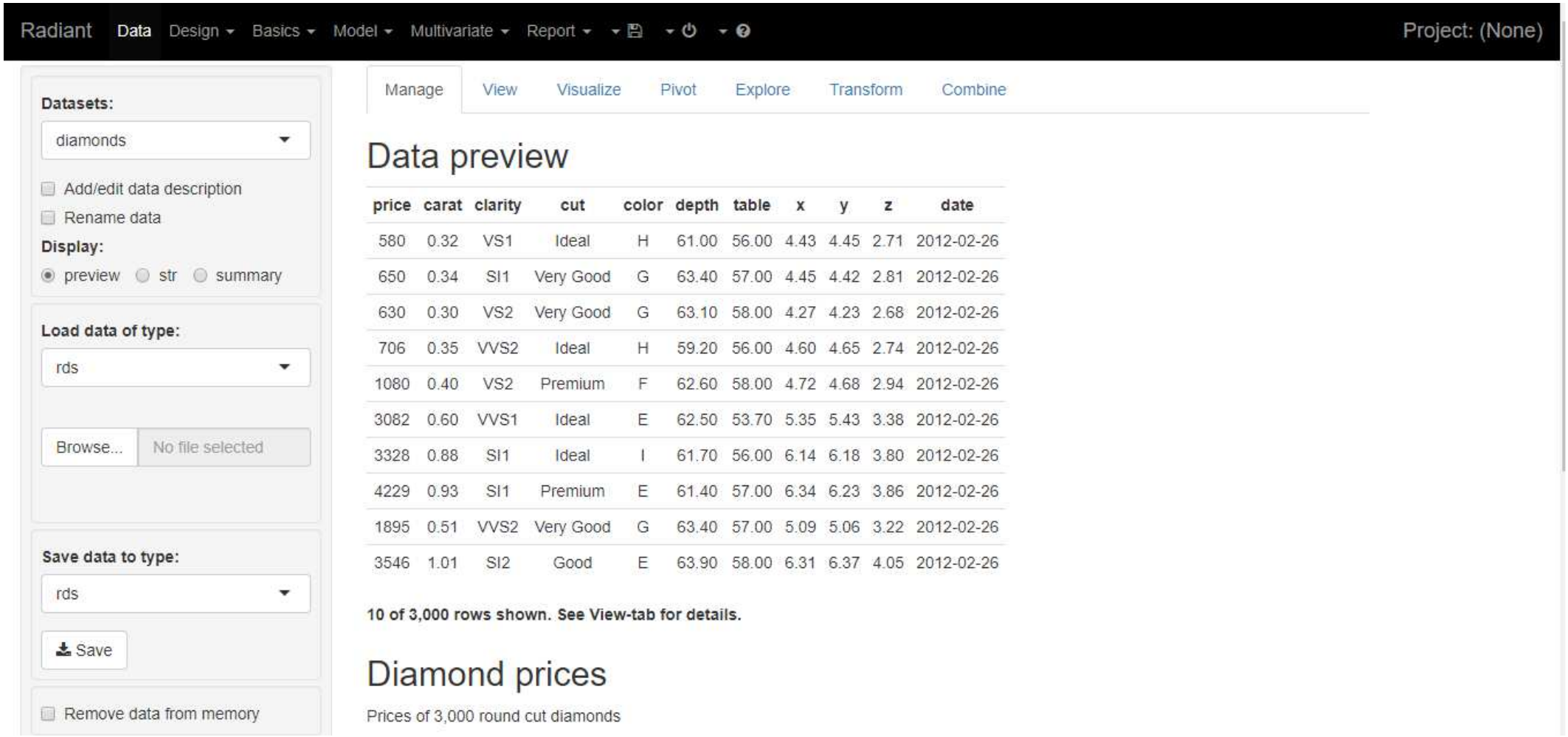
PNDB « analysis tools & platforms » proposal

R Shiny apps integration: GIS + Data exploration



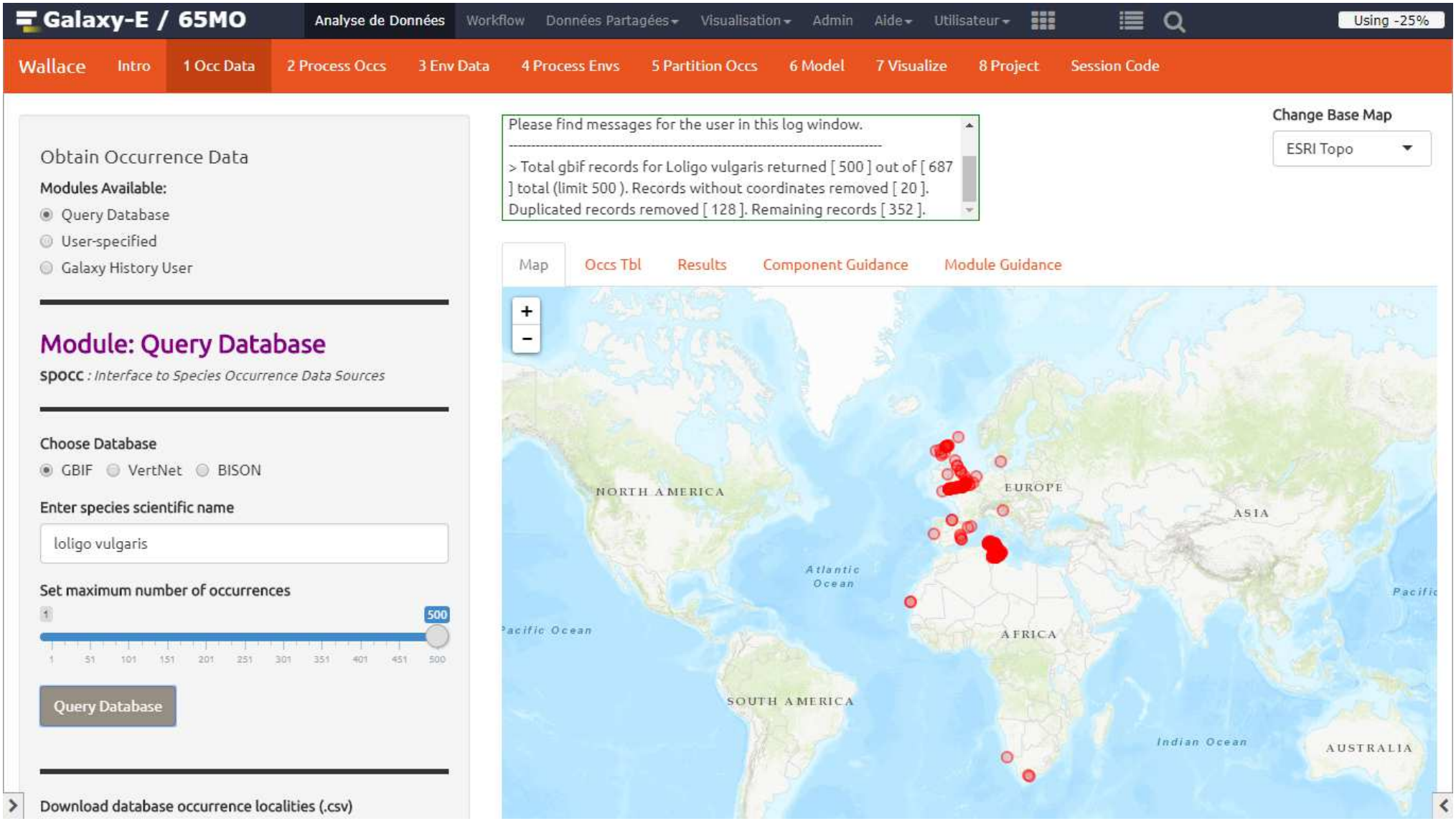
PNDB « analysis tools & platforms » proposal

R Shiny apps integration: Complete data analysis



PNDB « analysis tools & platforms » proposal

R Shiny apps integration: Species Distribution Modeling



References

- RDA FAIR Data Action Plan
- RDA Turning FAIR data into reality

FAIR Data

In research contexts, 'FAIR' or 'FAIR data' should be understood as a shorthand for a concept that comprises a range of scholarly materials that surround and relate to research data. This includes the algorithms, tools, workflows, and analytical pipelines that lead to creation of the data and give it meaning. It also encompasses the standards, metadata, vocabularies, ontologies and identifiers that are needed to provide meaning, both to the data itself and any associated materials.

- RDA FAIR data maturity model wg

The outcomes of the Working Group will be possible to be applied not only to data in the conventional sense but also to data-related algorithms, tools, workflows, protocols and other data-related services produced or managed by the assessed entity.

- FAIRsharing
- Jimenez *et al* (2017) Four simple recommendations to encourage best practices in research software
- Grüning *et al.* (2018) Practical Computational Reproducibility in the Life Sciences
- Mangul *et al.* (2018) A comprehensive analysis of the usablity & archival stability of omics computational tools & resources
- Batut et al. (2018) Community-Driven Data Analysis Training for Biology

Take home message

- **A necessary « FAIRization » process on Ecology**
  - Giving access of maximum of transparent and reusable data and related research digital objects (algorithms, tools, workflows, protocols…) to a wide range of users
- **Adding value to data and metadata**
  - Enhance data sharing and (re-)use
  - Curate / complete metadata
    - Ameliorating existing tools (try to don’t reinvent the wheel)
    - Involving experts + automatic processes
- **Giving access to data, tools and workflows**
  - Through a unique web open-source platform: Galaxy-E
  - Building/gathering communities of users and contributors
  - To different types of users (scientists with or without computing skills, area managers, citizens, policy makers)
  - Paying attention to reproducibility and optimization/mutualization