Ecole Thématique DATA SDUE

Guide de Survie dans la jungle des données
en Sciences de l'Univers et de l'Environnement (SDUE) :
Comment gérer les données pour les valoriser?

Session «*Notebook* et *Virtual Research Environment* »

Joel Sudre

Notebook et Virtual Research Environment

Joël Sudre, IR DATA TERRA / UAR 2013 CPST

Où en sommes-nous dans notre guide de survie?

Où en sommes-nous dans notre guide de survie?

- Cycle de vie des données de la recherche
- Comment mettre en place un plan de gestion de données
- Quels sont les formats d'échange à privilégier
- Comment bien insérer les données et métadonnées
- Où déposer vos données
- Comment valoriser vos données avec un Data Paper

D'après Research data lifecycle – UK Data Service
https://www.ukdataservice.ac.uk/manage-data/lifecycle

Notebook et Virtual Research Environment : De nouveaux outils pour Réutiliser les données…

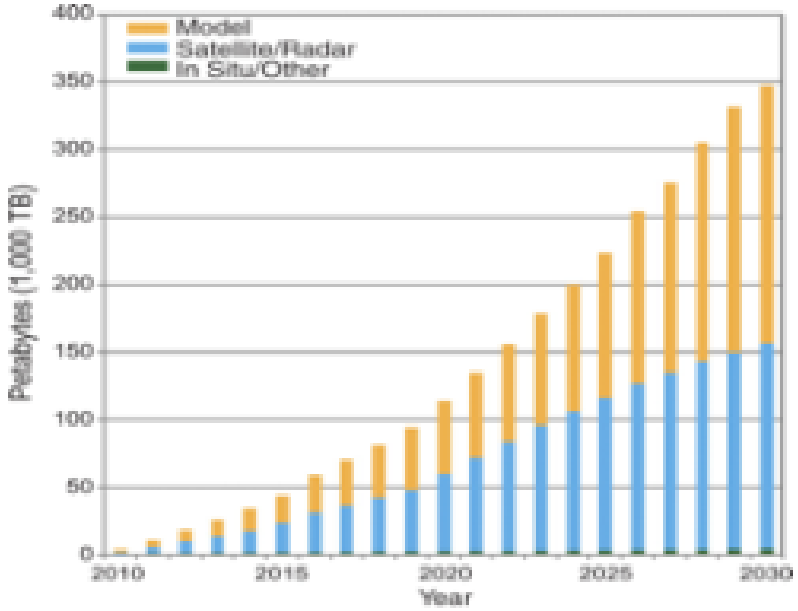
Pourquoi mettre à disposition de nouveaux outils pour la communauté scientifique?

Evolution des besoins des communautés scientifiques:

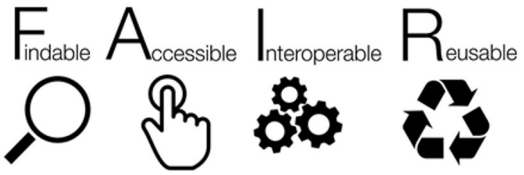
- *Nouvelle perspectives de recherche (Transdiciplinaire)*
- *Accès à des données **multi-source**s, **multi-capteurs***
- *Services d'accès aux données, traitements, analyse/modélisation, IA*

Augmentation exponentielle du nombre de données, diversités des sources, complexités, …

- *Spatiales, in-situ, modèles*
- *besoins d'analyse/réanalyse*
- *traitements intelligents*



Pourquoi mettre à disposition de nouveaux outils pour la communauté scientifique?



- Adaptés aux données bien normalisées
- Permettent accéder à de nombreux jeux de données mis à disposition via des catalogues (Pôle, data-lake, entrepôts,…)
- Permettent d'utiliser des moyens de calcul intensif (HPC/HPDA)
- Permettent d'utiliser des formats de fichiers avec des lecture/écriture beaucoup plus rapide que les formats classiques

Faciliter le travail des scientifiques

Deux catégories d'environnement virtuel

Les VRE : Virtual Research Environment

Simple d'utilisation :

Adapté aux scientifiques qui ne sont pas des programmeurs

-> Permettent de faire du traitement et de l'analyse de données sans forcement savoir programmer

-> Utilisent des traitements et des routines déjà codées

Environnement déjà déployé et utilisé par le PNDB

Les VRE : Virtual Research Environment

Volet analyse

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

Codes + **dépendances** → **Containers**

RStudio

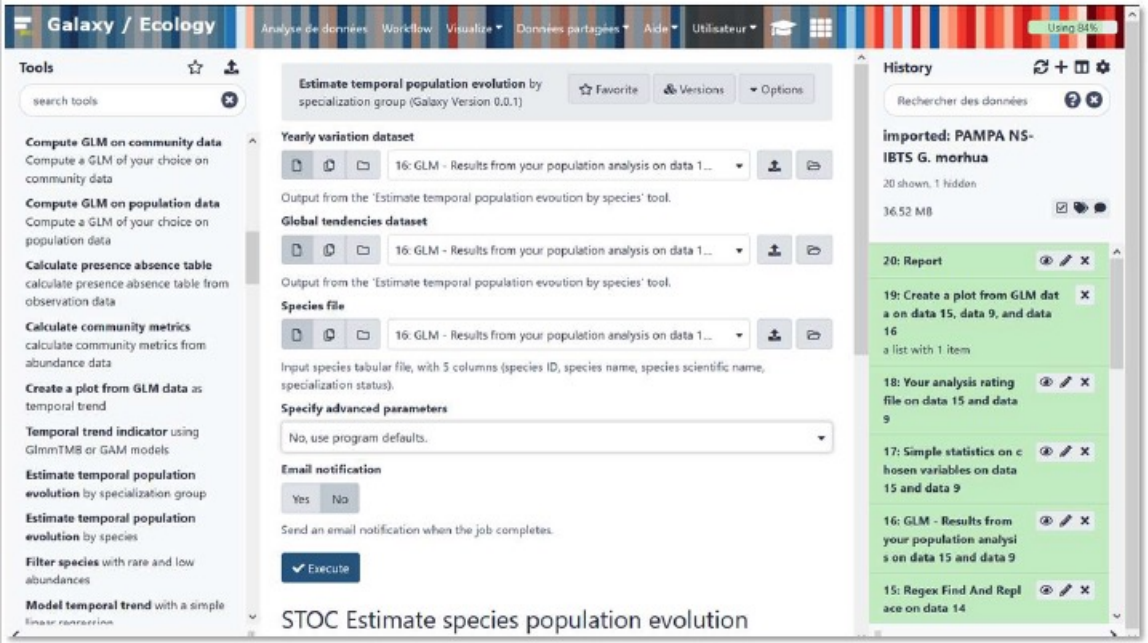
Autres plateformes

D4science

Jupyter Notebook

Machines virtuelles (local ou cloud)

Galaxy



Les VRE : Virtual Research Environment

Volet analyse

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

RStudio

Autres plateformes

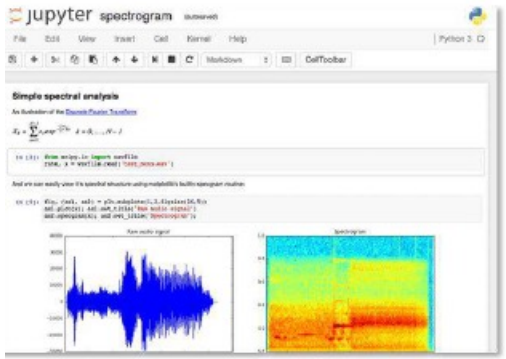
D4science

Jupyter Notebook

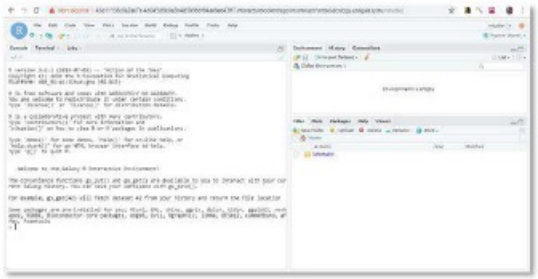
Codes + **dépendances** → **Containers**

Machines virtuelles (local ou cloud)

Galaxy

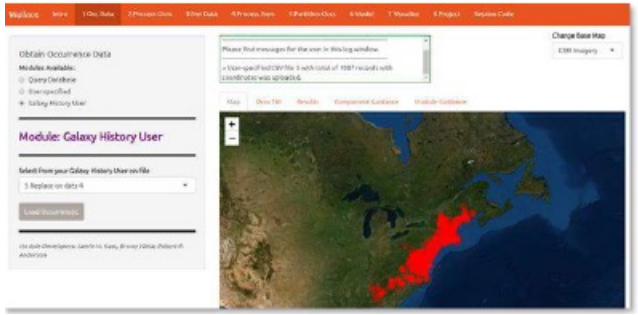


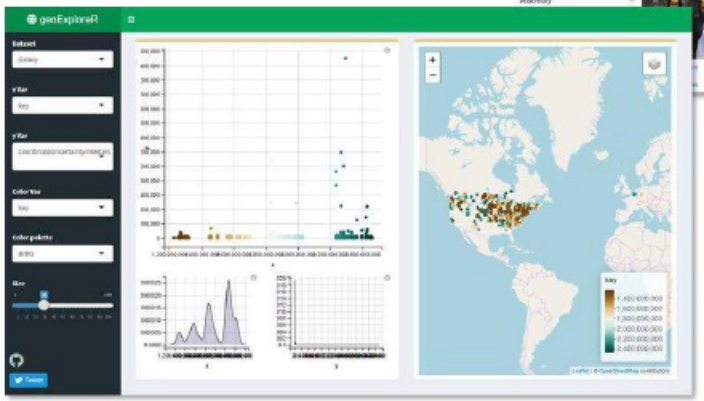
du Jupyter Notebook



du RStudio

Des apps R Shiny

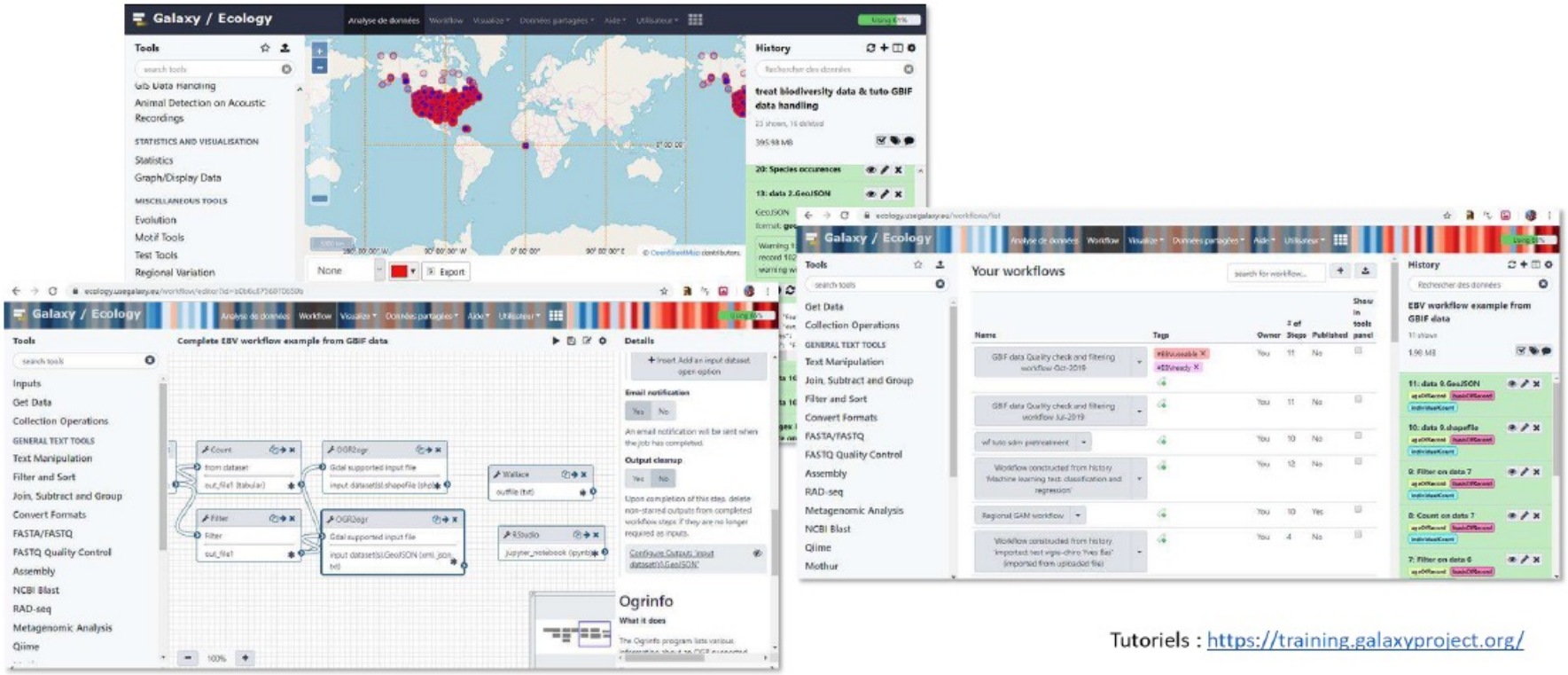




Les VRE : Virtual Research Environment

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

Deux catégories d'environnement virtuel

Les VAP : Virtual Analysis Platform

Demande une connaissance en programmation:

Adapté aux scientifiques qui ont l'habitude de développer des codes dans différents langages

-> Permettent de faire du traitement et de l'analyse de données intensif

-> Permettent de développer de nouveaux codes

-> Permettent d'accéder à des piles logiciels pré-installées sur des clusters, des HPC, etc.

-> L'utilisateur ne se préoccupe plus des installations!

-> Permettent ensuite via des containers de les installer dans une VRE

Les VAP : Virtual Analysis Platform

Environnement Virtuel qui met à disposition des Piles Logiciels avec des langages associés:

-> Python, Julia

-> R Studio, Rshiny

-> Matlab, IDL, Etc.

Jupyter: Environnement de Calcul Interactif avec un format de document reproductible (Code, Texte, Equation (LaTeX), visualisation)

Peut être déployé à la fois sur un PC, un cluster, un HPC/HPDA

Les VAP : Virtual Analysis Platform

Environnement Virtuel qui met à disposition des Piles Logiciels : ex PANGEO (http://pangeo.io/)

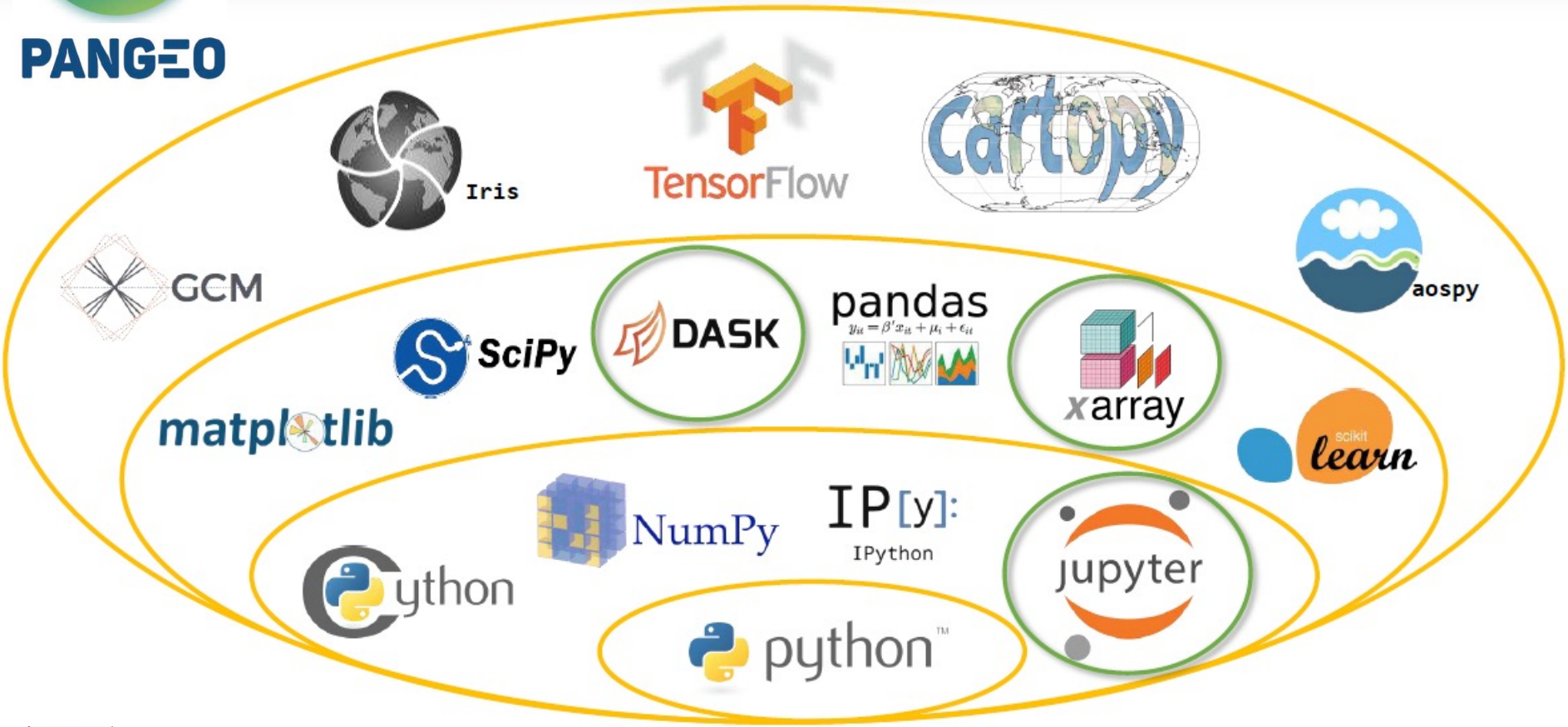
-> écosystème logiciels « big data geoscience »

-> Communauté internationale de développeurs

-> Infrastructure partagée sur le cloud

-> Nombreux codes mis à disposition

Les VAP : Virtual Analysis Platform



Les VAP : Virtual Analysis Platform

Utilisation en live d'une VAP sur un HPC