

DATA
TERRA

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...

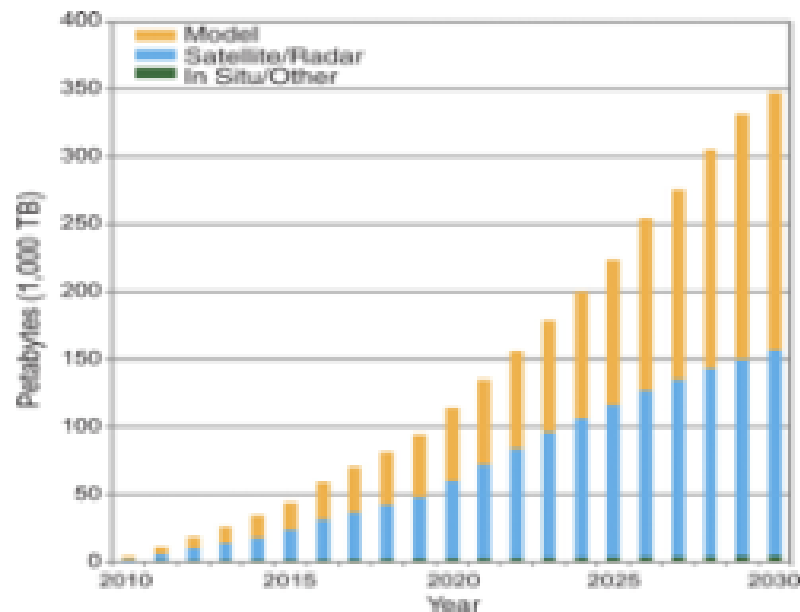
Pourquoi mettre à disposition de nouveaux outils pour la communauté scientifique?





Evolution des besoins des communautés scientifiques:

- Nouvelle perspectives de recherche (Transdisciplinaire)
- Accès à des données **multi-sources, 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



- F_{indable} A_{ccessible} I_{nteroperable} R_{eusable}
-    



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

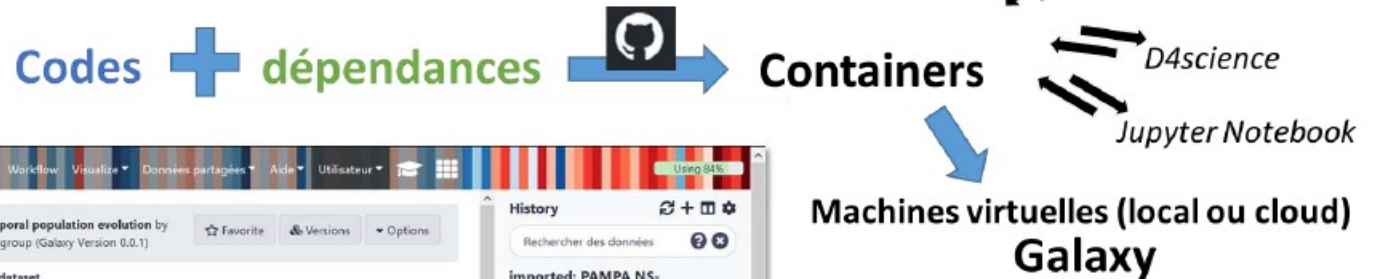
- > Permettent de faire du traitement et de l'analyse de données sans forcément savoir programmer
- > Utilisent des traitements et des routines déjà codées

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*



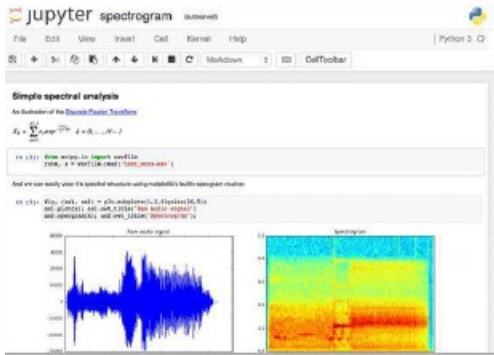
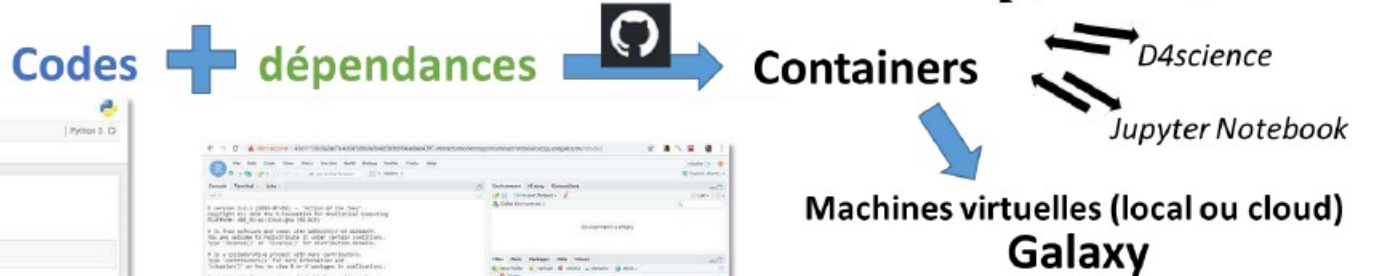
The screenshot shows the Galaxy web interface. The main panel displays a workflow titled 'Estimate temporal population evolution by specialization group (Galaxy Version 0.0.1)'. The workflow includes steps like 'Yearly variation dataset', 'Global tendencies dataset', 'Species file', and 'Specify advanced parameters'. The 'Execute' button is visible at the bottom. The right panel shows a history of previous runs, including 'Imported: PAMPA NS-IBTS G. morhua' and '20: Report'.

The screenshot shows the Galaxy web interface with a workflow titled 'Complete GBIF workflow example from GBIF data'. The workflow includes steps like 'Get Data', 'Collection Operations', 'Filter and Sort', 'Join, Subselect and Group', 'Convert Formats', 'FASTQ Quality Control', 'Assembly', 'NCBI Blast', 'Metagenomic Analysis', 'QIIME', and 'Metagenome in Analysis'. The 'Execute' button is visible at the bottom. The right panel shows a history of previous runs, including '18: GBIF data processing' and '19: Filter on data 14'.

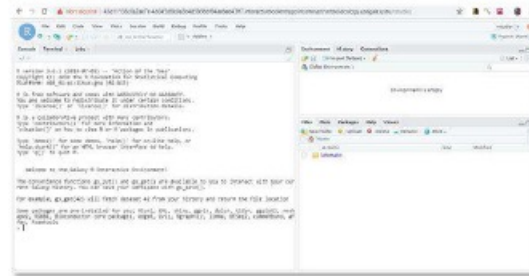
Les VRE : Virtual Research Environment

Volet analyse

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

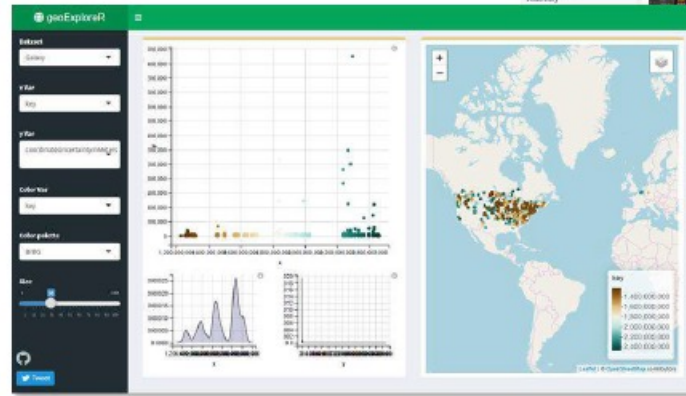
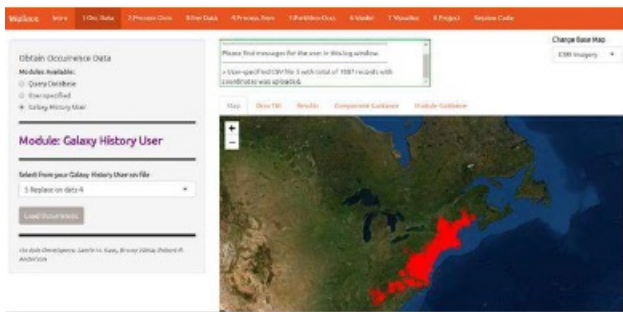


du Jupyter Notebook



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/>

The screenshots illustrate the Galaxy/Ecology interface, which is a virtual research environment for ecological data analysis. The top left panel displays a list of tools categorized into 'Statistics and Visualisation', 'Miscellaneous Tools', and 'Regional Variation'. The top right panel shows a 'History' section with a list of workflows, including 'treed biodiversity data & tuto GBIF data handling' and '20: Species occurrences'. The bottom left panel shows a 'Complete ERV workflow example from GBIF data' with a visual workflow diagram. The bottom right panel shows a 'Your workflows' table with columns for Name, Tags, Owner, Steps, Published, and Panel.

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/>)

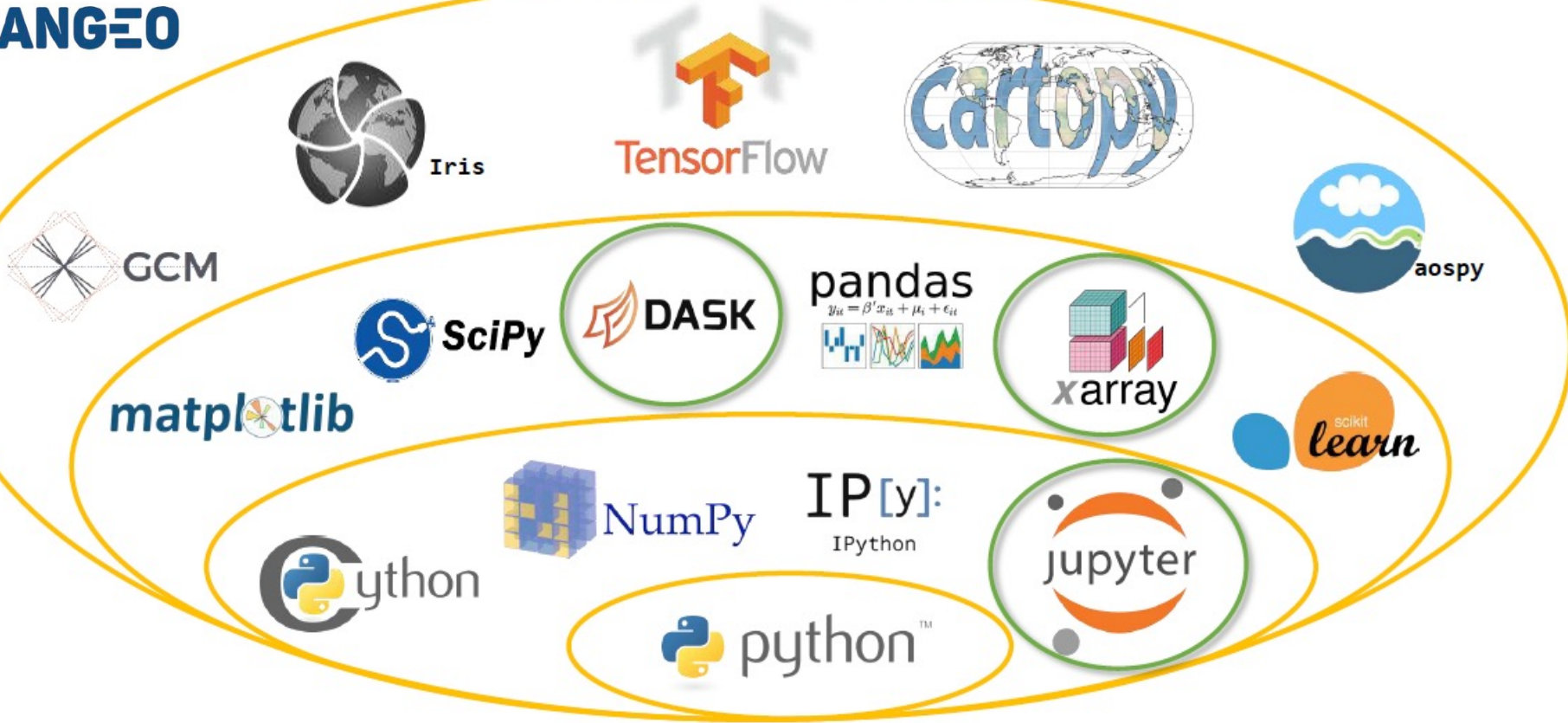
- > écosystème logiciels « big data geoscience »
- > Communauté internationale de développeurs
- > Infrastructure partagée sur le cloud
- > Nombreux codes mis à disposition



Les VAP : Virtual Analysis Platform



PANGEO



Les VAP : Virtual Analysis Platform

Utilisation en live d'une VAP sur un HPC