

Science ouverte et principes FAIR dans un projet de bioinformatique

Comment rendre un projet bioinformatique plus reproductible ?

Roscoff, 2-6 juin 2025

Charlotte Berthelier, Pauline François,
Thomas Denecker & l'équipe formation de l'IFB



Une présentation sous le signe de l'Open



Attribution - Partage dans les Mêmes
Conditions 2.0 France (CC BY-SA 2.0 FR)

This is a human-readable summary of (and not a substitute for) the license. [Avertissement.](#)



Vous êtes autorisé à :

Partager — copier, distribuer et communiquer le matériel par tous moyens et sous tous formats

Adapter — remixier, transformer et créer à partir du matériel pour toute utilisation, y compris commerciale.

L'Offrant ne peut retirer les autorisations concédées par la licence tant que vous appliquez les termes de cette licence.



Selon les conditions suivantes :

Attribution — Vous devez créditer l'Oeuvre, intégrer un lien vers la licence et indiquer si des modifications ont été effectuées à l'Oeuvre. Vous devez indiquer ces informations par tous les moyens raisonnables, sans toutefois suggérer que l'Offrant vous soutient ou soutient la façon dont vous avez utilisé son Oeuvre.

Partage dans les Mêmes Conditions — Dans le cas où vous effectuez un remix, que vous transformez, ou créez à partir du matériel composant l'Oeuvre originale, vous devez diffuser l'Oeuvre modifiée dans les même conditions, c'est à dire avec la même licence avec laquelle l'Oeuvre originale a été diffusée.

Pas de restrictions complémentaires — Vous n'êtes pas autorisé à appliquer des conditions légales ou des mesures techniques qui restreindraient légalement autrui à utiliser l'Oeuvre dans les conditions décrites par la licence.



<https://creativecommons.org/licenses/by-sa/2.0/fr/>

**Un contenu trouvable simplement,
accessible, décrit et réutilisable**

Les slides

<https://doi.org/10.6084/m9.figshare.23275349.v3>

Le code

https://github.com/IFB-ElixirFr/FAIR_EBAII_n2

DOI [10.5281/zenodo.4800637](https://doi.org/10.5281/zenodo.4800637)

Site web

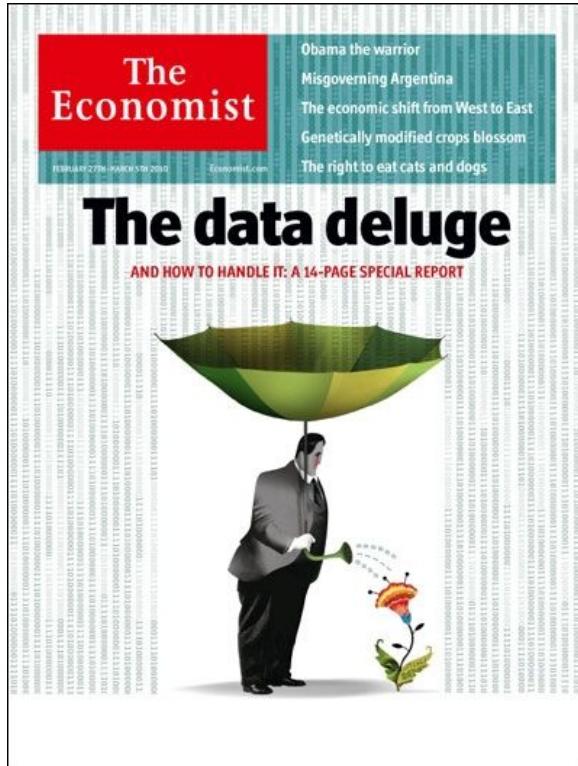
https://ifb-elixirfr.github.io/FAIR_EBAII_n2/



Contexte

Contexte

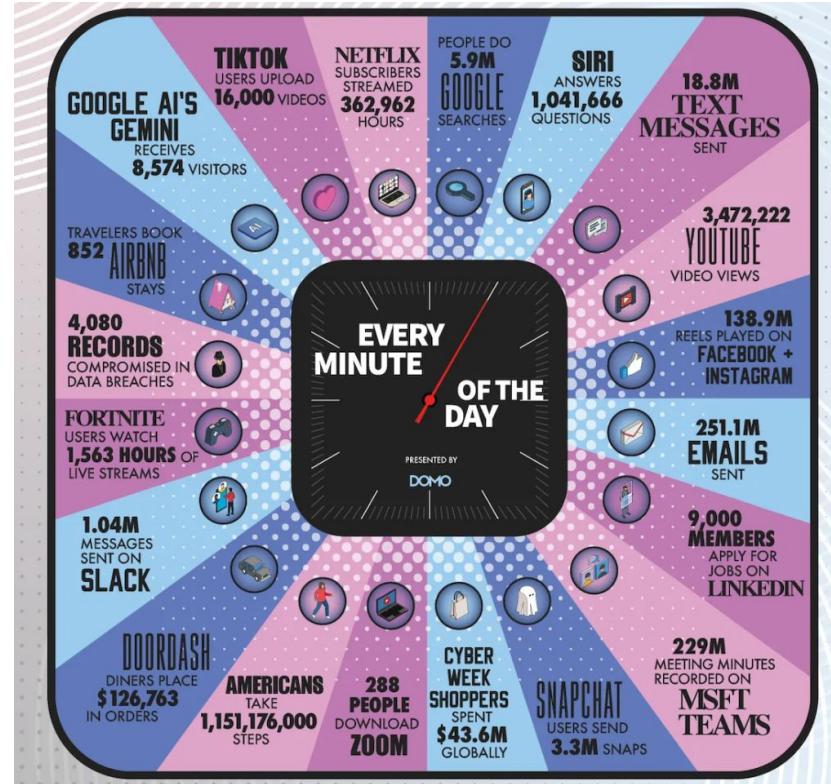
De plus en plus de données



Data is the new oil
Clive Humby

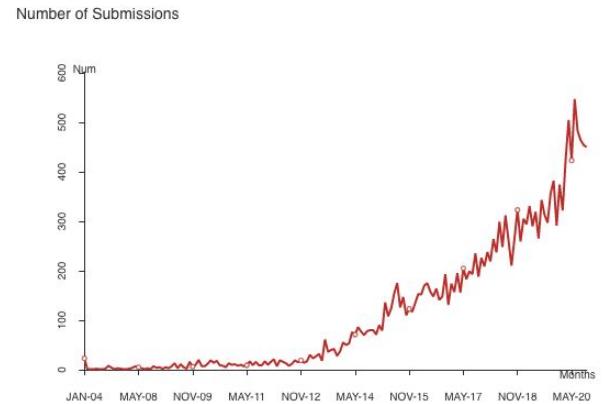
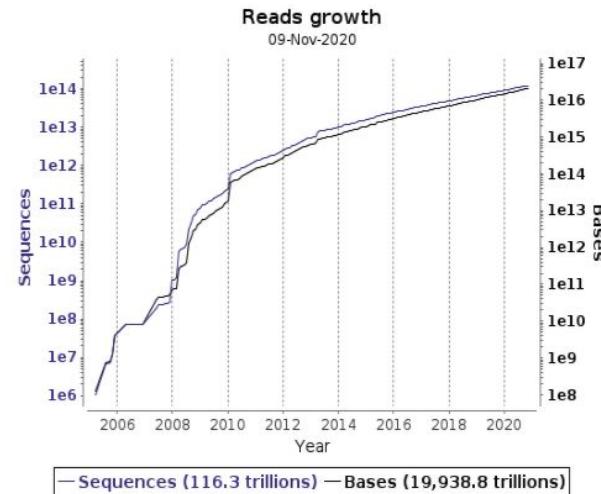
Data is the new oil? No: Data is the new soil.
David McCandless

Quelques chiffres au quotidien



<https://content-technology.com/media-business/domestic-releases-12th-annual-data-never-sleeps-report/>

En biologie, c'est pareil !



Data deluge en biologie

Type de données	Base de données	Volumes de données
Mesures de l'expression des gènes	ArrayExpress	72 938 experiments 2 429 810 assays 56,68 TB of archived data
Mesures de l'expression des gènes	GEO	5 570 704 échantillons
Structure 3D des protéines	PDB	177 009 Structures
Séquence nucléotidiques	GenBank	241 830 635 séquences et 1 731 302 248 418 bases
Données d'identification ou de quantification des protéines	Pride	580 917 268 spectres de masse
Séquence nucléotidiques (COVID-19)	GISAID (EpiCoV)	15 158 725 séquences virales

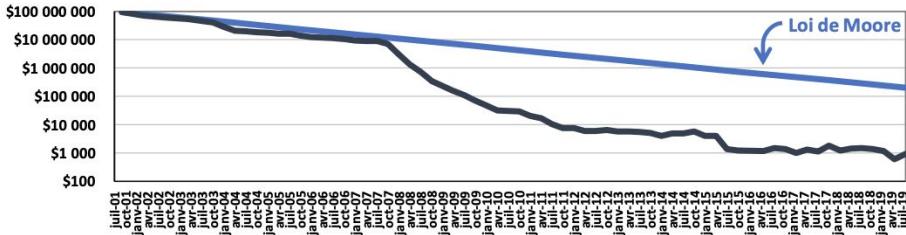
MAJ : Mars 2023

Contexte

Comment ?

Cout

Prix par génome humain



Vitesse

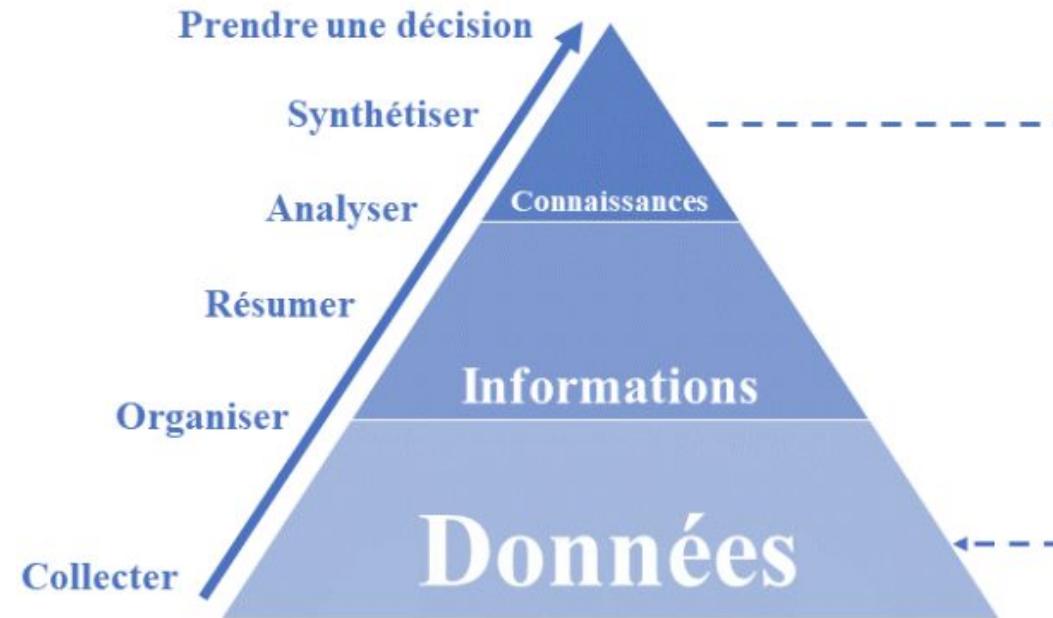
Dans les années 90

4 ans pour le premier milliard de nucléotides du génome humain

Aujourd'hui

Le génome complet en moins de 24h

Pourquoi ? Créeer de la connaissance !



Pourquoi générer toujours plus de données ?

Pourquoi ne pas simplement exploiter les données déjà disponibles ?

- La description des données est encore trop souvent incomplète ;
- Les données ne sont pas facilement récupérables ;
- Il n'y a souvent pas de contrôle systématique des erreurs par des experts ;
- Les données ne sont pas générées exactement de la façon souhaitée ;
- Une question de confiance.

Plus il y a de données, mieux c'est !



Conclusion : Plus simple ? Plus rapide ? Plus sûr ?

Une question de confiance

Lin, D., Crabtree, J., Dillo, I. et al. The TRUST Principles for digital repositories. Sci Data 7, 144 (2020).

<https://doi.org/10.1038/s41597-020-0486-7>



Les 3 “R” de la confiance

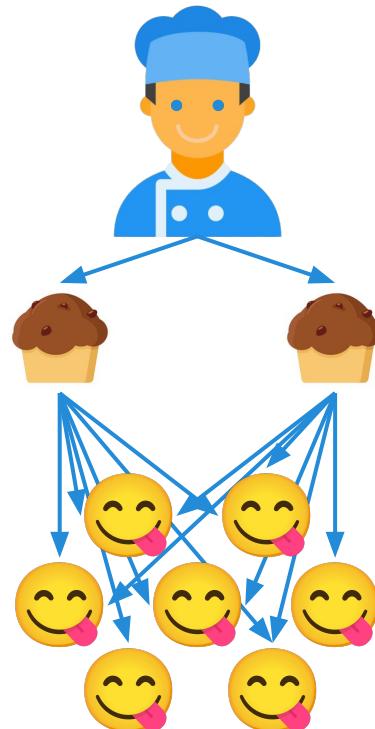
Souvent utilisées mais souvent confondues
et notamment par la langue (Plesser, 2018)



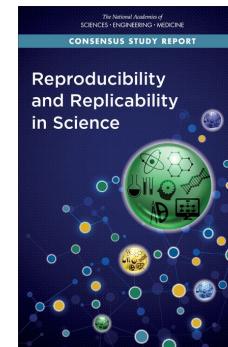
éplicabilité
épétabilité
eproductibilité

Réplicabilité

Jour J



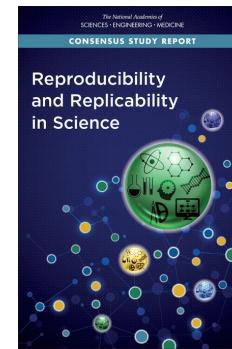
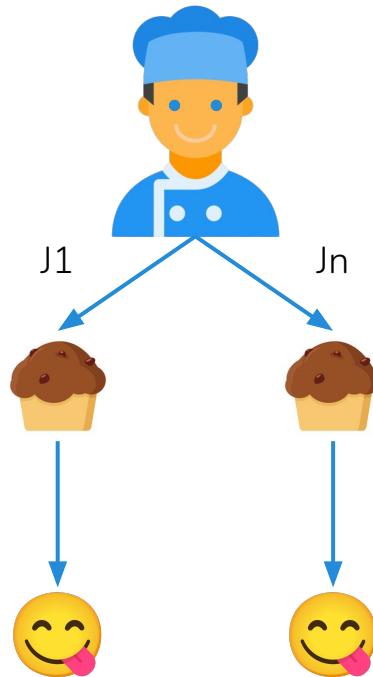
“L’extrême de l’accord entre les résultats individuels successifs obtenus sur le même échantillon soumis à l’essai dans le même laboratoire et dans les conditions suivantes : même analyste, même appareil, même jour ”



<https://doi.org/10.17226/25303>

“Obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data ”

Répétabilité

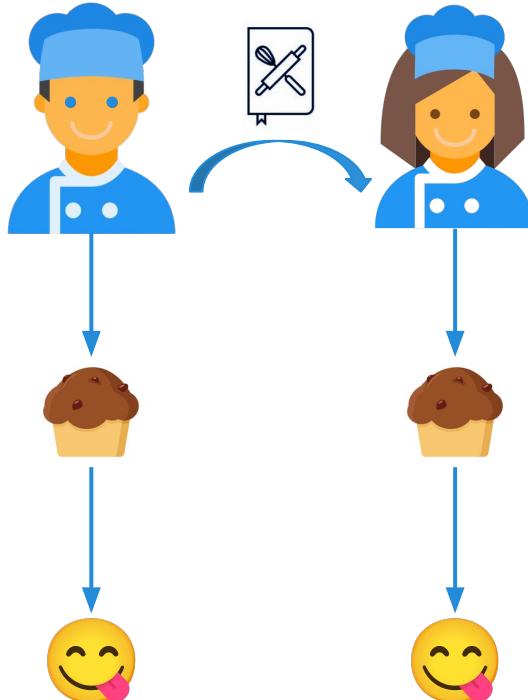


<https://doi.org/10.17226/25303>

“L’étroitesse de l’accord entre les résultats individuels obtenus sur le même échantillon soumis à l’essai dans le même laboratoire et dont au moins l’un des éléments suivants est différent : l’analyste, l’appareil, le jour”

“The measurement can be obtained with stated precision by the same team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same location on multiple trials. For computational experiments, this means that a researcher can reliably repeat her own computation ”

Reproductibilité



“L’étroitesse de l’accord entre les résultats individuels obtenus sur le même échantillon soumis à l’essai dans des laboratoires différents et dans les conditions suivantes : analyste différent, appareil différent, jour différent ou même jour”

“Obtaining consistent computational results using the same input data, computational steps, methods, code, and conditions of analysis”



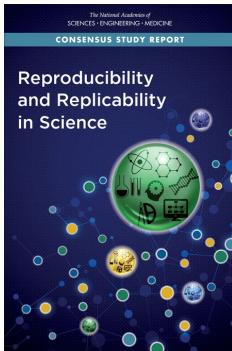
Contexte

En résumé

		Data	
		Same	Different
Code	Same	Reproducible	Replicable
	Different	Robust	Generalisable

<https://doi.org/10.6084/m9.figshare.5443201.v1>,

Recommandations pour être reproductible



<https://doi.org/10.17226/25303>

Description de la partie expérimentale

Méthodes, instruments, procédures, mesures, conditions expérimentales

Description de la partie computationnelle

Etapes de l'analyse des données et choix techniques

Description de la partie statistique

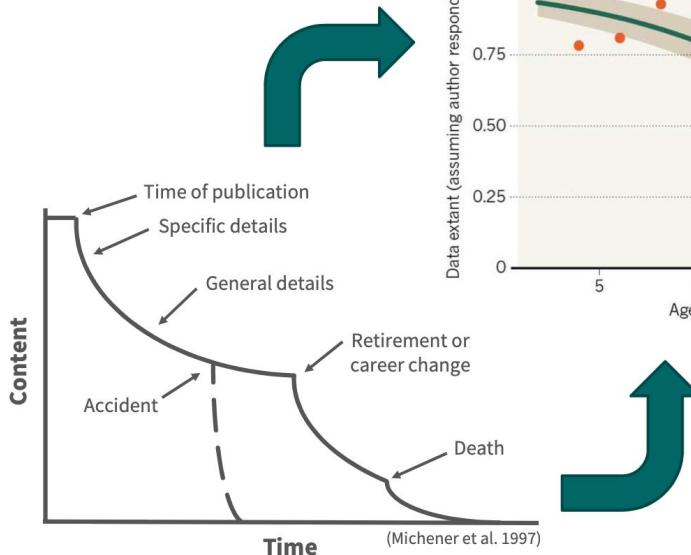
Décisions analytiques : quand, comment, pourquoi

Discussion des choix et des résultats obtenus

Et dans les faits ?

Les données face aux ravages du temps

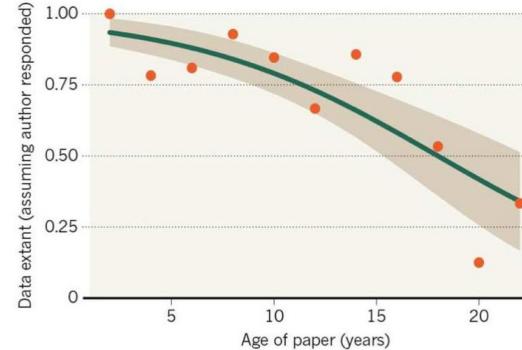
Data Entropy



3

MISSING DATA

As research articles age, the odds of their raw data being extant drop dramatically.



Vines, T. H. et al. Curr. Biol. <http://dx.doi.org/10.1016/j.cub.2013.11.014> (2013).

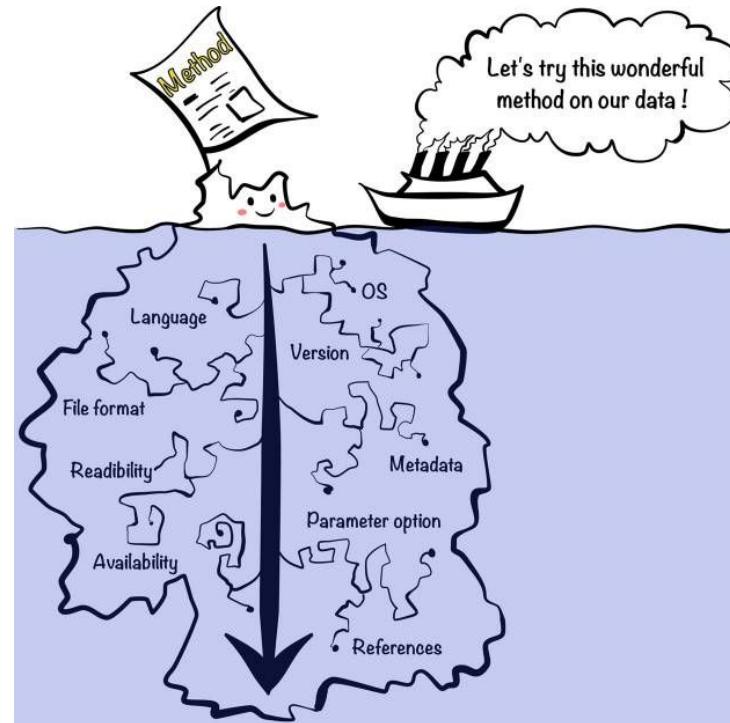
DataONE

Récupérer les données

https://youtu.be/66oNv_DJuPc



Reproduire les données



Kim et al, 2018

<https://dx.doi.org/10.1093%2Fgigascience%2Fgjy077>

En biologie

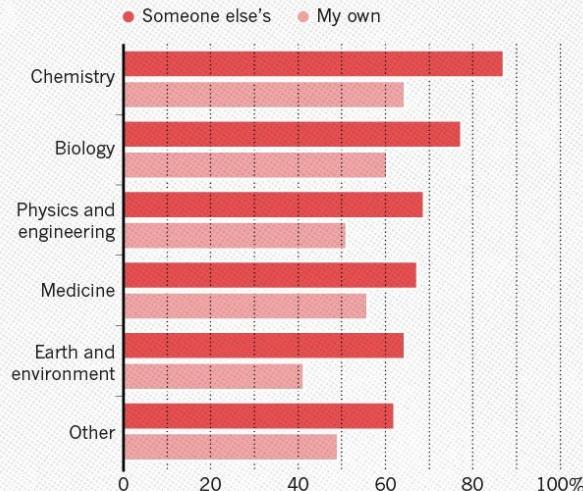
70 %

des analyses en biologie expérimentale
ne sont
pas reproductibles

Ten Years Reproducibility Challenge : êtes vous capables de refaire vos analyses d'il y a 10 ans ?

HAVE YOU FAILED TO REPRODUCE AN EXPERIMENT?

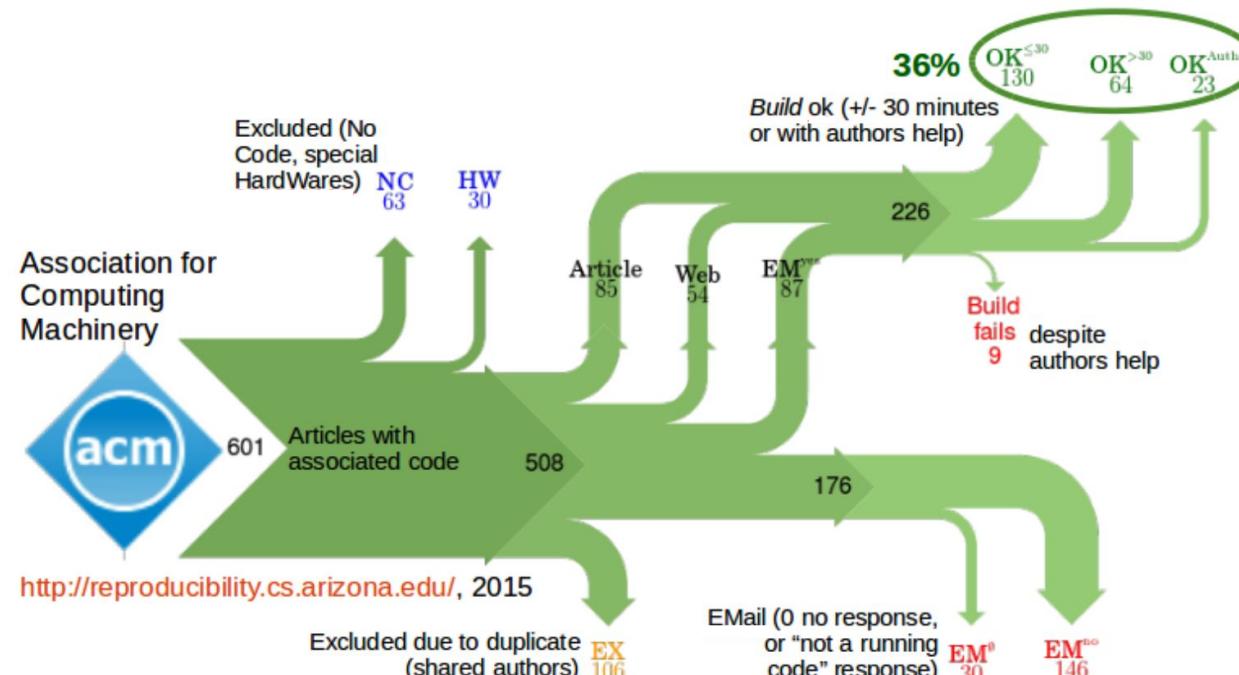
Most scientists have experienced failure to reproduce results.



Monya Baker, 2016

Crise de la reproductibilité

En informatique



(Collberg et al. 2015)

Crise de la reproductibilité

La bioinformatique n'y échappe pas...

Impossibilité d'installer des outils

- OS non compatible
- Dépendance plus disponible / plus valide

Mise à jour de l'outil rendant inutilisable les codes

- Python 2 et Python 3 !
- Changement des arguments des fonctions utilisées (R)

Impossibilité de reproduire les résultats de l'analyse computationnelle

- IDE : version stable du langage différente selon l'OS (Rstudio)
- Version des packages



(Perkel, Nature, 2020)

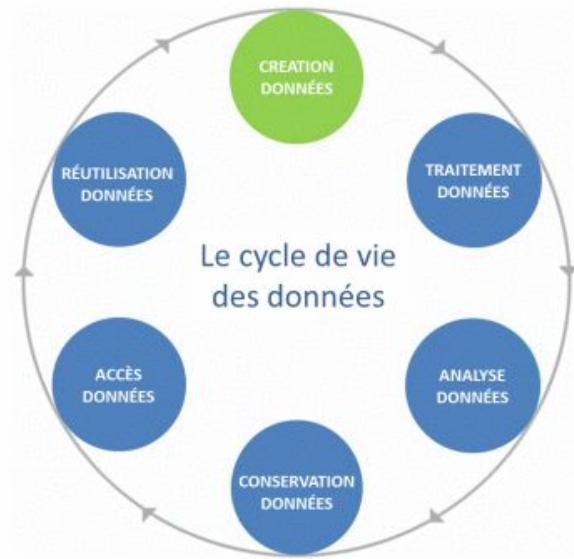
Cycle de vie d'un projet



LA RECHERCHE AGROBIOLOGIQUE
POUR LE DÉVELOPPEMENT

donnees-recherche@cirad.fr

C'est quoi le cycle de vie des données ?



Le modèle de UK Data Archive définit les six étapes suivantes :

- **Création ou collecte des données** (*creating data*) ;
- **Traitements des données** (*processing data*) ;
- **Analyse des données** (*analysing data*) ;
- **Conservation des données** (*preserving data*) ;
- **Accès aux données** (*giving access to data / data discovery*) ;
- **Réutilisation des données** (*reusing data*).

[Une introduction à la gestion et au partage des données de la recherche - Le cycle de vie des données](#)

*Comment rendre un projet bioinformatique plus
reproductible ?*

PARTIE I

Du point de vue de l'analyse des données

Des principes hérités des principes FAIR data

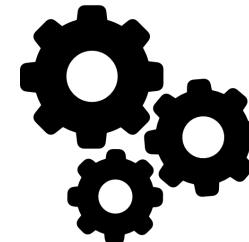
Findable



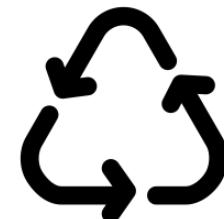
Accessible



Interoperable



Reusable



Des principes hérités des principes FAIR data

🔍 F

👉 A

⚙️ |

⟳ R

Des principes hérités des principes FAIR data

Q Facile à trouver (pour l'Homme et la machine)

- Identifiant unique (un DOI par exemple)
- Métadonnées décrivant l'analyse et les outils ; elles sont FAIR, consultables et indexables
- Protocoles d'analyses simples à trouver (GitHub pages)

👉 A

⚙️ |

↻ R

Des principes hérités des principes FAIR data

🔍 Facile à trouver (pour l'Homme et la machine)

- Identifiant unique (un DOI par exemple)
- Métadonnées décrivant l'analyse et les outils ; elles sont FAIR, consultables et indexables
- Protocoles d'analyses simples à trouver (GitHub pages)

👉 Accessible

- Ressources disponibles (GitHub, Dockerhub) et outils tiers open source (conda)
- Les métadonnées sont accessibles, même lorsque le logiciel n'est plus disponible



|



Des principes hérités des principes FAIR data

🔍 Facile à trouver (pour l'Homme et la machine)

- Identifiant unique (un DOI par exemple)
- Métadonnées décrivant l'analyse et les outils ; elles sont FAIR, consultables et indexables
- Protocoles d'analyses simples à trouver (GitHub pages)

👉 Accessible

- Ressources disponibles (GitHub, Dockerhub) et outils tiers open source (conda)
- Les métadonnées sont accessibles, même lorsque le logiciel n'est plus disponible

⚙️ Interopérable

- Coopération des outils (snakemake, docker) aussi bien en local que sur serveurs (cloud ou cluster)

♻️ R

Des principes hérités des principes FAIR data

🔍 Facile à trouver (pour l'Homme et la machine)

- Identifiant unique (un DOI par exemple)
- Métadonnées décrivant l'analyse et les outils ; elles sont FAIR, consultables et indexables
- Protocoles d'analyses simples à trouver (GitHub pages)

👉 Accessible

- Ressources disponibles (GitHub, Dockerhub) et outils tiers open source (conda)
- Les métadonnées sont accessibles, même lorsque le logiciel n'est plus disponible

⚙️ Interopérable

- Coopération des outils (snakemake, docker) aussi bien en local que sur serveurs (cloud ou cluster)

♻️ Réutilisable

- Protocole rejouable (snakemake) à l'identique (Jupyter) dans un environnement virtuel (docker)

Pour aller plus loin

scientific data



OPEN

ARTICLE

Introducing the FAIR Principles for research software

Michelle Barker^{1,2*}, Neil P. Chue Hong^{3,2}, Daniel S. Katz^{3,2}, Anna-Lena Lamprecht^{3,2}, Carlos Martinez-Ortiz^{3,2}, Fotis Psomopoulos^{4,6}, Jennifer Harrow⁵, Leyla Jael Castro^{3,2}, Morane Grunepet², Paula Andrea Martinez^{3,10} & Tom Honeyman^{3,11}

Research software is a fundamental and vital part of research, yet significant challenges to discoverability, productivity, quality, reproducibility, and sustainability exist. Improving the practice of scholarship is a common goal of the open science, open source, and FAIR (Findable, Accessible, Interoperable and Reusable) communities and research software is now being understood as a type of digital object to which FAIR should be applied. This emergence reflects a maturation of the research community to better understand the crucial role of FAIR research software in maximising research value. The FAIR for Research Software (FAIR4RS) Working Group has adapted the FAIR Guiding Principles to create the FAIR Principles for Research Software (FAIR4RS Principles). The contents and context of the FAIR4RS Principles are summarised here to provide the basis for discussion of their adoption. Examples of implementation by organisations are provided to share information on how to maximise the value of research outputs, and to encourage others to amplify the importance and impact of this work.

Introduction

In 2014 the publication of ‘The FAIR Guiding Principles for scientific data management and stewardship’¹ supported a vision where valuable scientific outputs are made ‘FAIR’ by becoming more Findable, Accessible, Interoperable and Reusable. From the outset, the FAIR Guiding Principles were intended to be applicable to many different types of assets. Increasingly, attention has focused on the management of research software in research has catalysed application of the FAIR Guiding Principles to this type of digital asset.

Community-endorsed FAIR principles for research software were released in 2022 by the FAIR for Research Software (FAIR4RS) Working Group (WG), which was jointly convened by the Research Software Alliance (ReSA), Future Of Research Communications and E-scholarship (FORCE11), and the Research Data Alliance (RDA). This milestone reflects a maturation of the research community in understanding the benefits of having FAIR research software, and coming together as the FAIR4RS WG to achieve this. The FAIR4RS WG is a global, interdisciplinary community whose members share an interest in the application of FAIR principles to research software, such as researchers, software users, developers and maintainers, policy makers, infrastructure support staff, and funders.

The FAIR4RS Principles are relevant to any stakeholder in the research community seeking to increase transparency, reproducibility, and reusability of research. This paper highlights the importance of the FAIR4RS Principles and the positive signals of adoption that demonstrate high levels of community support. It must also be acknowledged that research software and data discoverability is a long-standing challenge and there have

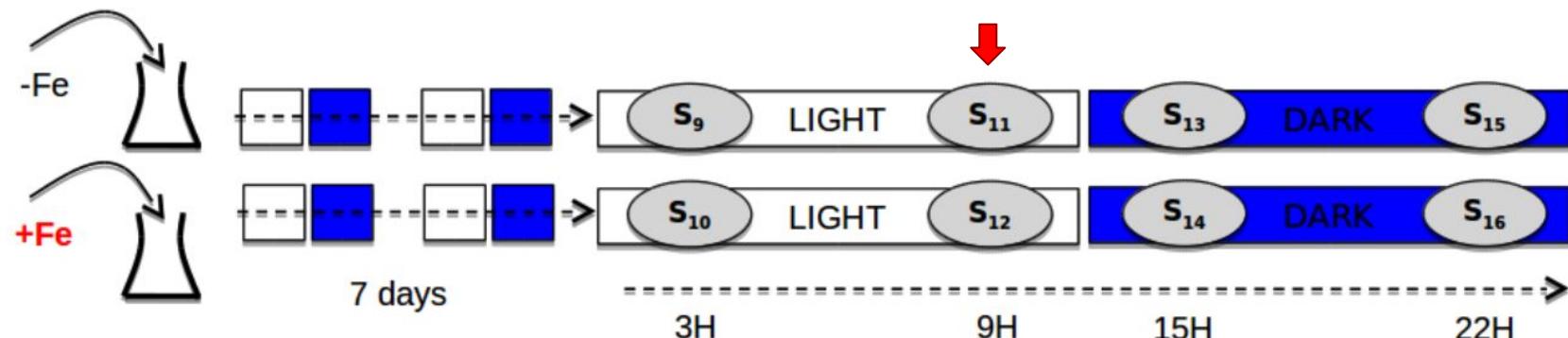
¹Research Software Alliance, Qld 4780, Cairns, Australia. ²Software Sustainability Institute & EPCC, University of Edinburgh, J4 Pottenrow, Edinburgh, EH8 9BT, UK. ³NCSA & CSE & ECE & iSchool, University of Illinois at Urbana-Champaign, 1205 W Clark St., Urbana, IL, 61801, USA. ⁴Institute of Computer Science, University of Potsdam, An der Bahn 2, 14476, Potsdam, Germany. ⁵Netherlands eScience Center, Science Park 140, 1098 XG, Amsterdam, Netherlands. ⁶Institute of Applied Biosciences, Centre for Research and Technology Hellas, Thessaloniki, 57001, Greece. ⁷ELIXIR Hub, South Building, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK. ⁸Semantic Technologies team, ZB MED Information Centre for Life Sciences, Gleueler Strasse 60, 50931, Cologne, Germany. ⁹Software Heritage, Inria, 2 rue Simone Iff, Paris, 75012, France. ¹⁰Research Software Alliance Australian Research Data Commons, Level 6, Dung Tower, The University of Queensland, Brisbane, QLD 4072, Australia. ¹¹Australian Research Data Commons, University of Technology Sydney Library, Ultimo, NSW, 2007, Australia.
*e-mail: michelle@researchsoft.org

Exemple d'utilisation

Exemple d'utilisation

Plan expérimental

Étude de la réponse à une privation en fer chez l'algue verte *Ostreococcus tauri*
16 échantillons de données RNAseq (triplicat, single-end de 100bp)

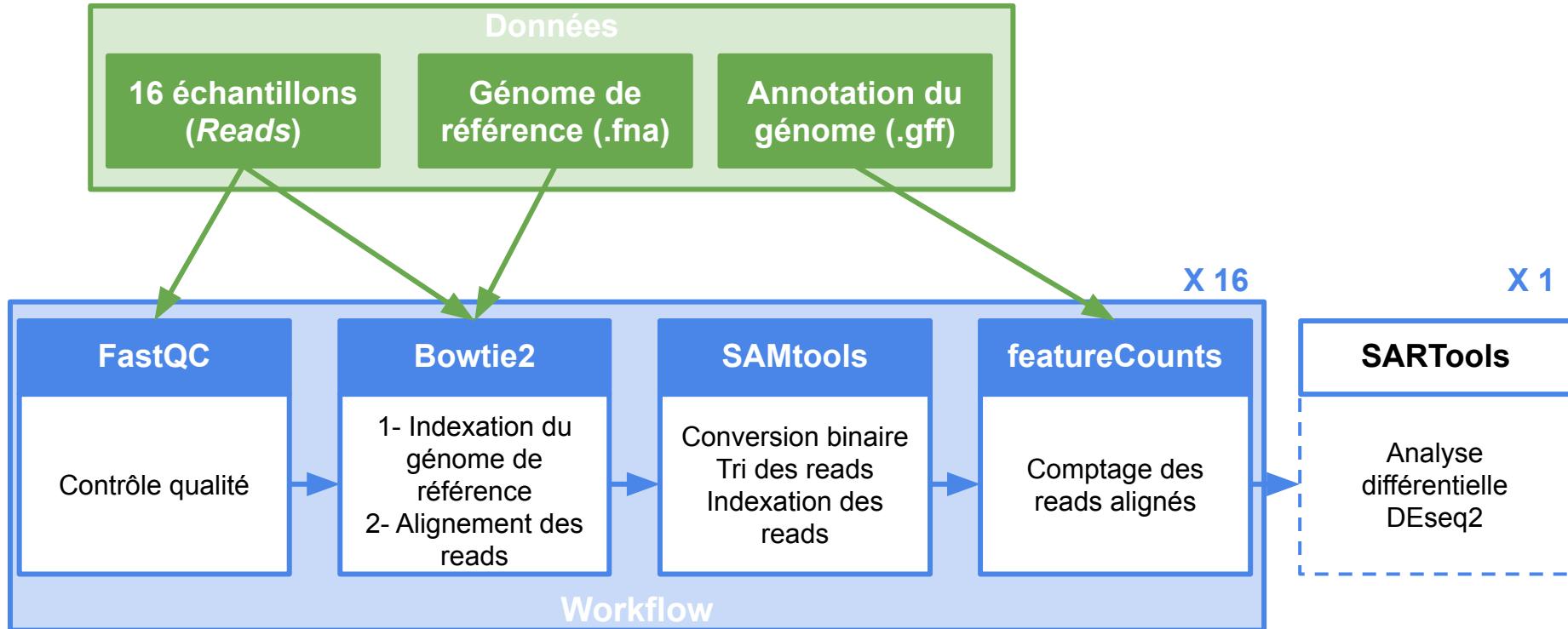


(Lelandais *et al.* 2016)

Données réduites pour la démonstration

Exemple d'utilisation

Analyse de données



Notre question à présent ?

**Comment réaliser cette analyse
de façon reproductible ?**

(et pourquoi pas la rejouer en un click!)

Proposition en 7 étapes



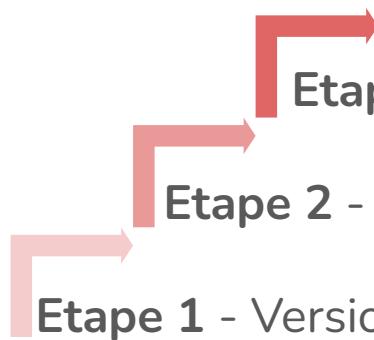
Etape 1 - Versionner le code et le partager

Proposition en 7 étapes

Etape 2 - Création d'un environnement virtuel

Etape 1 - Versionner le code et le partager

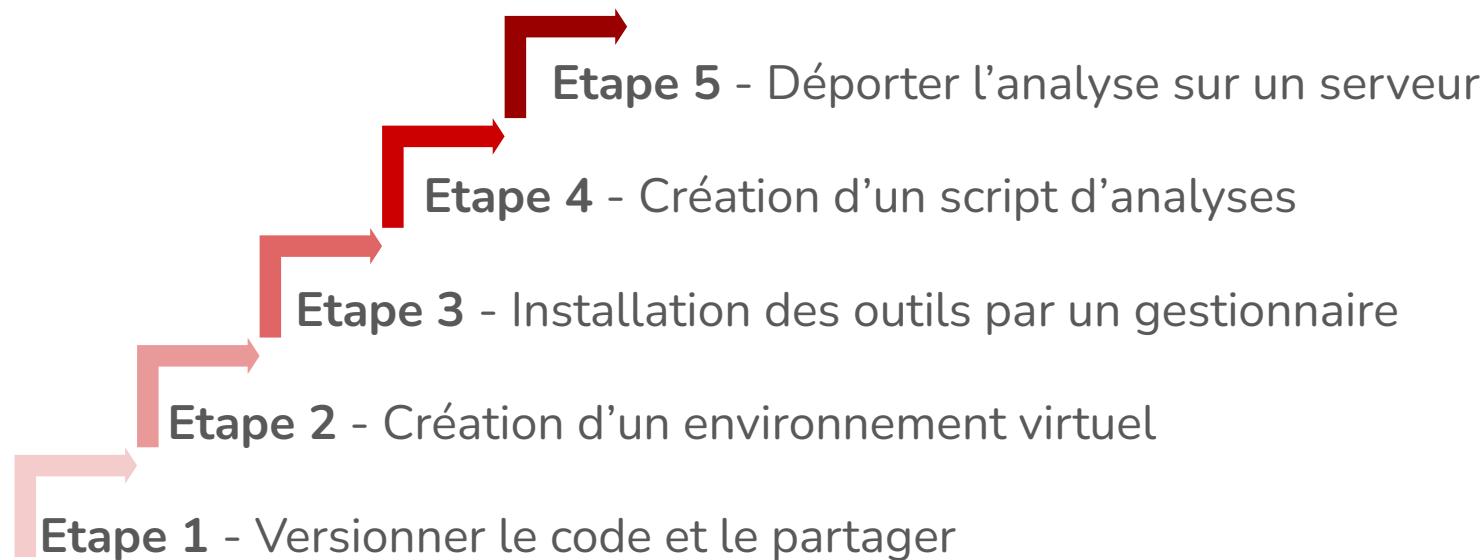
Proposition en 7 étapes

- 
- Etape 1 - Versionner le code et le partager**
 - Etape 2 - Création d'un environnement virtuel**
 - Etape 3 - Installation des outils par un gestionnaire**

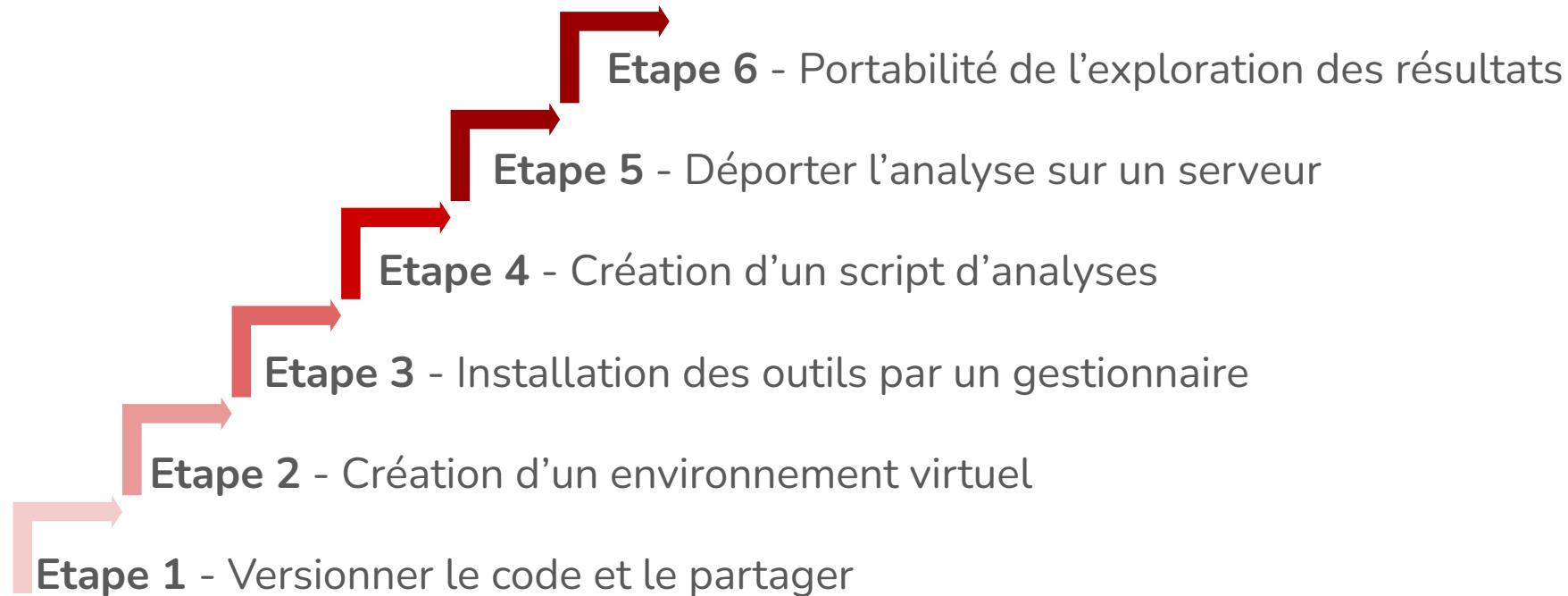
Proposition en 7 étapes

-
- ```
graph TD; E1[Etape 1 - Versionner le code et le partager] --> E2[Etape 2 - Cr閐ation d'un environnement virtuel]; E2 --> E3[Etape 3 - Installation des outils par un gestionnaire]; E3 --> E4[Etape 4 - Cr閐ation d'un script d'analyses]
```
- Etape 1 - Versionner le code et le partager**
  - Etape 2 - Cr閐ation d'un environnement virtuel**
  - Etape 3 - Installation des outils par un gestionnaire**
  - Etape 4 - Cr閐ation d'un script d'analyses**

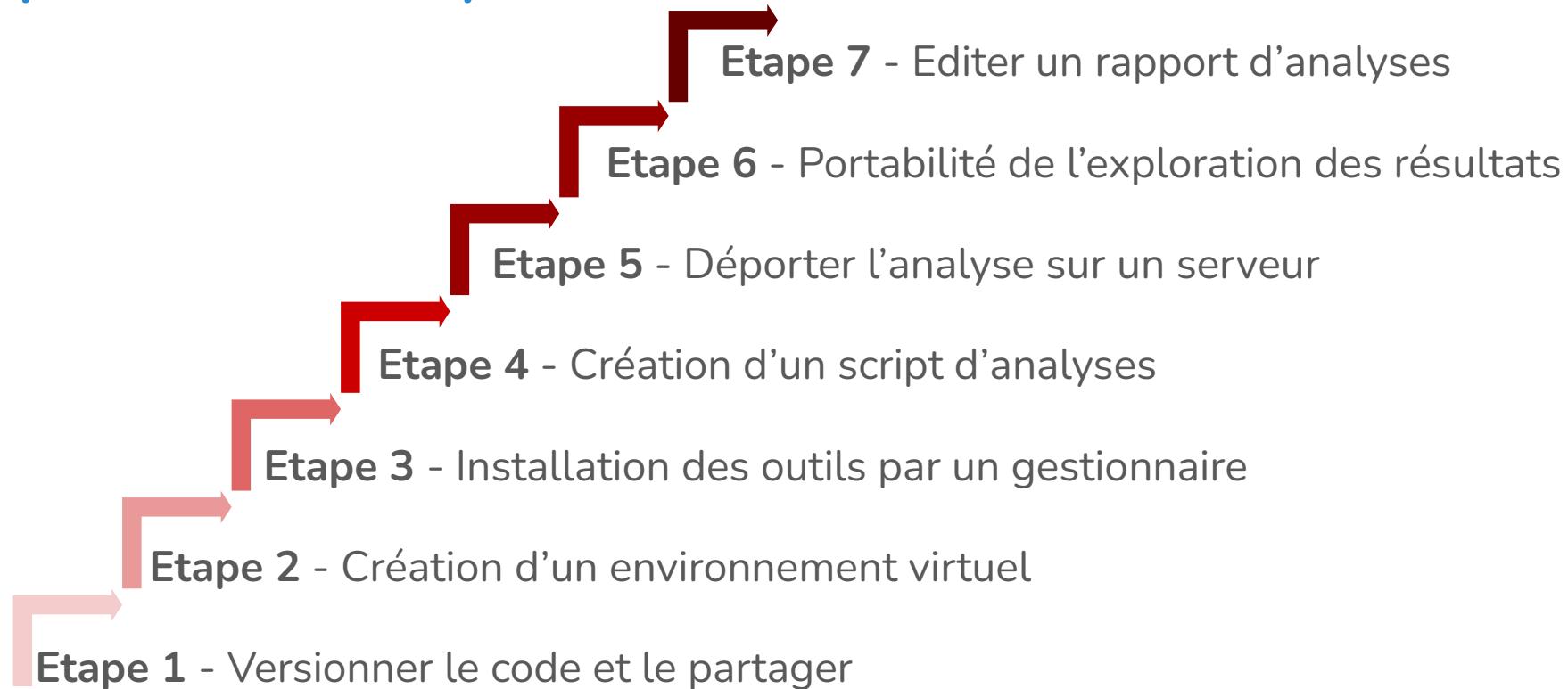
## Proposition en 7 étapes



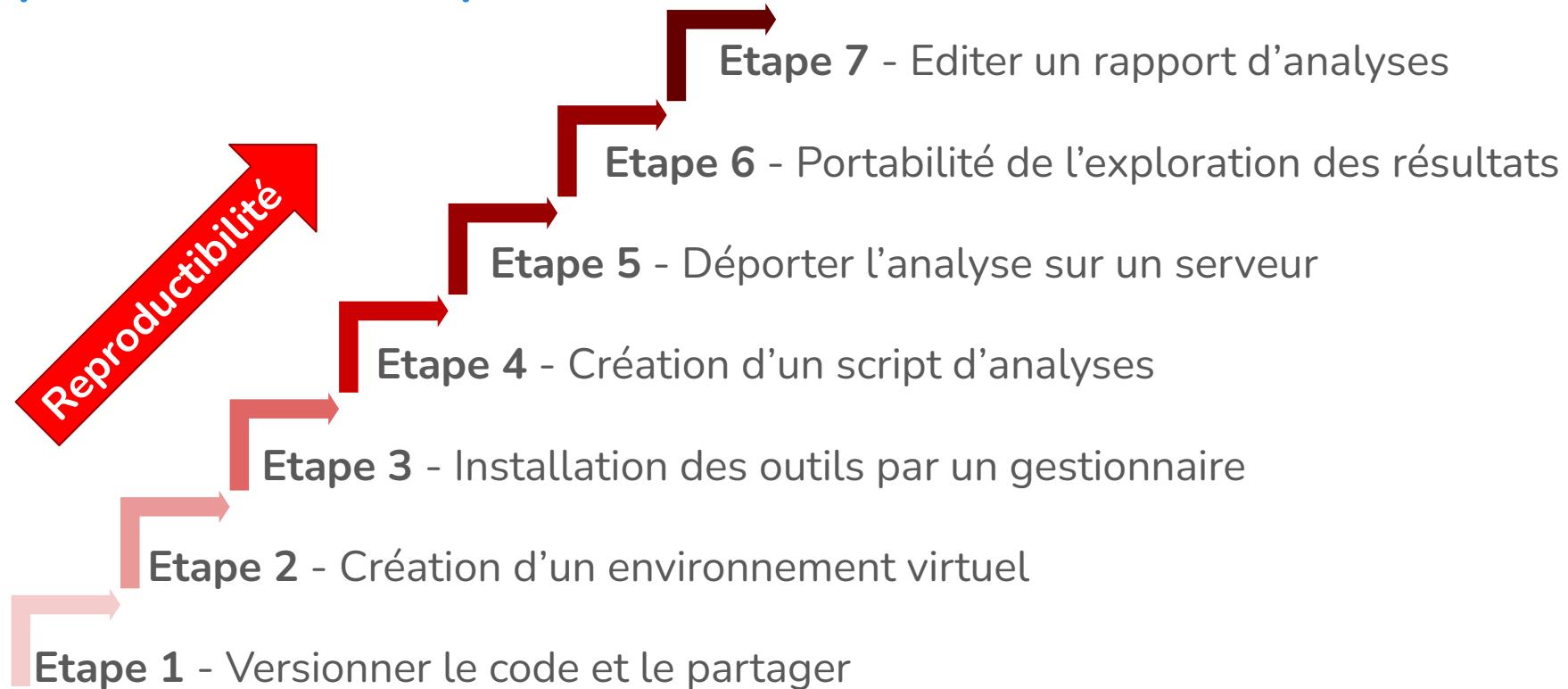
## Proposition en 7 étapes



## Proposition en 7 étapes



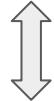
## Proposition en 7 étapes



# IMPORTANT

## Des choix techniques équivalents

GitHub



GitLab

Atlassian  
Bitbucket

# IMPORTANT

## Des choix techniques équivalents

GitHub



GitLab



Atlassian  
Bitbucket



# IMPORTANT

## Des choix techniques équivalents

GitHub



GitLab

Atlassian  
Bitbucket



nextflow



# IMPORTANT

## Des choix techniques équivalents

GitHub



GitLab

Atlassian  
Bitbucket



nextflow



ABIMS<sup>4</sup> aws

Microsoft Azure  
Google Cloud

## Etape 1 - Versionner le code et le partager

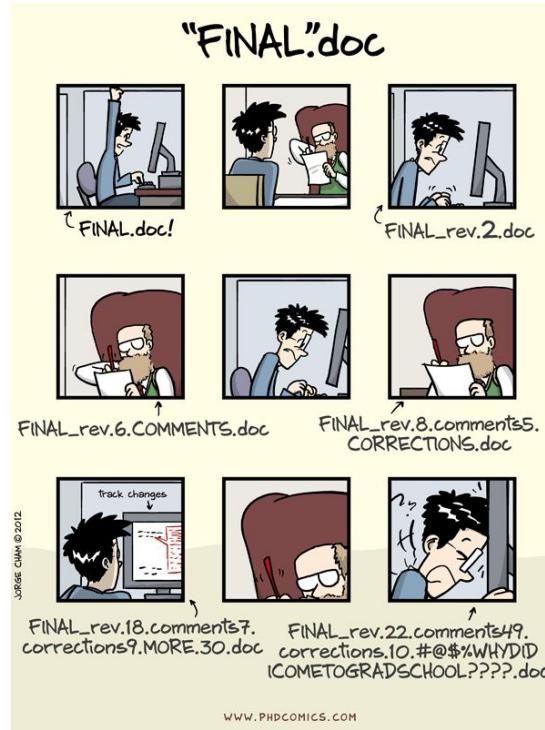
Pourquoi ?

- Figer la version du code
- Vision dans le temps
- Partage & ouverture à la communauté

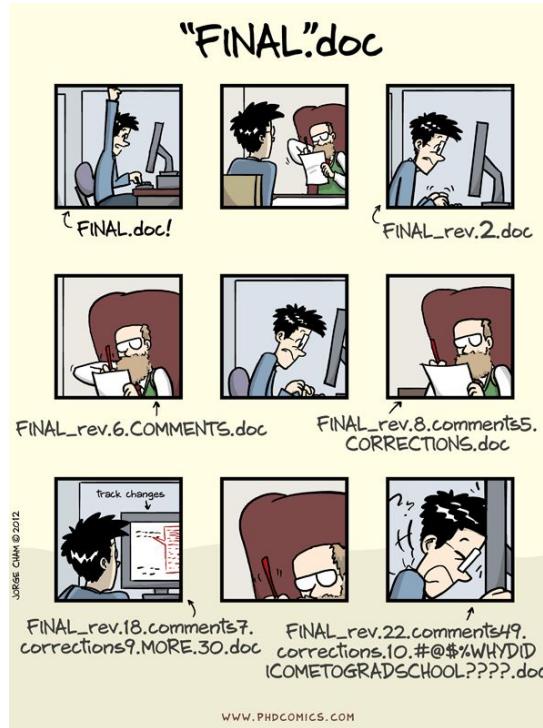


# Toujours plus reproductible !

## Etape 1 - Versionner le code et le partager



## Etape 1 - Versionner le code et le partager



### Avantages

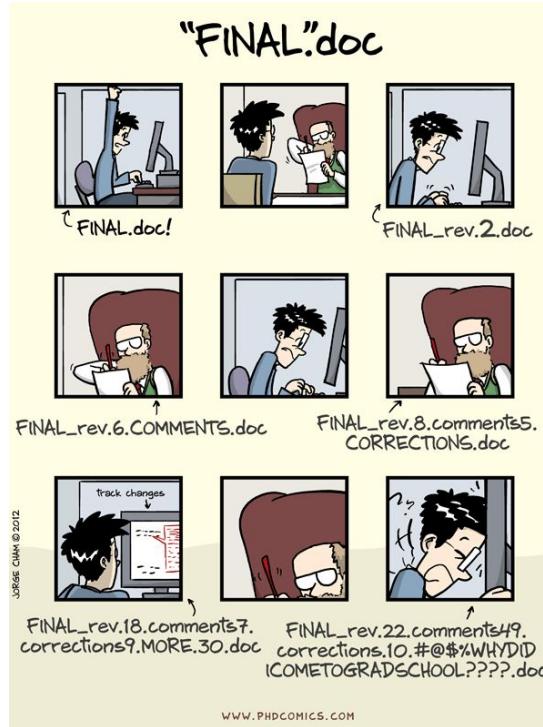
- Sauvegarde du code
- Simple pour partager
- Gestion automatique des versions, fusionne les différentes modifications

### Inconvénients

- Pas simple pour les novices

# Toujours plus reproductible !

## Etape 1 - Versionner le code et le partager



Après

IFB-ElixirFr / IFB-FAIR-bioinfo-training

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

master 2 branches 2 tags Go to file Add file Download Code

thomasdenecker add new links 92e2188 on 2 Sep 61 commits

docs add new links 3 months ago

.gitignore ignore .DS\_Store files 3 months ago

Images.zip to unzip for git-github.md 3 months ago

LICENSE Initial commit 3 months ago

README.md Update README.md 3 months ago

\_config.yml Set theme jekyll-theme-cayman 3 months ago

README.md

Les principes FAIR appliqués à la bioinformatique

L'Institut Français de Bioinformatique (IFB) organise en partenariat avec l'Institut de Biologie Intégrative de la Cellule (IBiC) une formation à destination des bioinformaticiens et biostatisticiens souhaitant mettre en oeuvre les principes "FAIR" (Facile à trouver, Accessible, Interopérable, Réutilisable) dans leurs projets d'analyse et de développement. Les concepts FAIR, initialement définis dans le contexte d'ouverture des données de la recherche, seront ici adaptés pour cadre avec un projet type de développement et/ou analyse bioinformatique/biostatistique. Ainsi, la formation s'abordera par les aspects "FAIR" spécifiques aux données mais introduira plusieurs outils permettant d'améliorer la reproductibilité des analyses.

Pour plus d'informations (programme, slides, ...), un site web de la formation est disponible [ici](#).

Objectifs pédagogiques

A la fin de cette formation, les participants pourront mettre en oeuvre les principes de la science reproductible : encapsuler un environnement de travail, concevoir et exécuter des workflows, gérer des versions de code, passer à l'échelle sur un cluster de calcul, gérer des environnements logiciels et assurer la traçabilité de leur analyse à l'aide de Notebooks.

La formation organisée en deux temps

La formation s'organise en deux temps :

## Etape 2 - Création d'un environnement virtuel

Pourquoi ?

- Figer l'environnement
- Partager l'environnement



# Toujours plus reproductible !

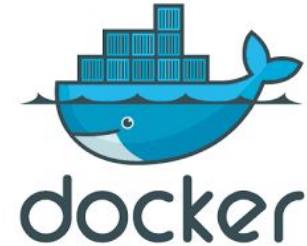
## Etape 2 - Création d'un environnement virtuel

Avant



## Etape 2 - Crédit d'un environnement virtuel

Avant



### Avantages

- Rapide et léger
- Portable
- Simple à partager et déployer

### Inconvénients

- Avec un système à jour
- Accepté dans votre structure ?

## Etape 2 - Crédit d'un environnement virtuel

Avant



Après : figé un outil  
(R & un package)



```
FROM rocker/tidyverse
```

```
MAINTAINER Thomas DENECKER (thomas.denecker@gmail.com)
```

```
RUN Rscript -e
```

```
'devtools::install_github("PF2-pasteur-fr/SARTools",
build_opts="--no-resave-data")'
```

## Etape 3 - Installation des outils par un gestionnaire

Pourquoi ?

- Avoir la bonne version
- Installer simplement



## Etape 3 - Installation des outils par un gestionnaire

Avant : exemple de FastQC

- 1) Télécharger la source
- 2) Décompresser le dossier
- 3) Installer et mettre à jour Java  
(nombreux problèmes)
- 4) Changer les droits

## Etape 3 - Installation des outils par un gestionnaire

Avant : exemple de FastQC

- 1) Télécharger la source
- 2) Décompresser le dossier
- 3) Installer et mettre à jour Java  
(nombreux problèmes)
- 4) Changer les droits



# CONDA

### Avantages

- Gestionnaire simple à installer
- Installation simple des paquets
- Gestion des versions

### Inconvénients

- Peut être lourd (solution miniconda)
- Paquets manquants (R)

## Etape 3 - Installation des outils par un gestionnaire

### Avant : exemple de FastQC

- 1) Télécharger la source
- 2) Décompresser le dossier
- 3) Installer et mettre à jour Java  
(nombreux problèmes)
- 4) Changer les droits



### Après

```
$ conda install -c bioconda -y fastqc=0.12.1
```

Tous les outils utilisés dans le protocole sont disponibles sur Conda (<https://anaconda.org/>) : bowtie2, samtools, htseqcount, aspera, snakemake, ...

Installation simple en une ligne !

## Etape 4 - Création d'un script d'analyse

Pourquoi ?

- Avoir un script d'analyse reproductible
- Ne pas refaire ce qui est déjà fait
- Paralléliser



## Etape 4 - Création d'un script d'analyse

### Avant (script Shell)

```
for sample in `ls *.fastq.gz`
do
 fastqc ${sample}
done
```

## Etape 4 - Crédation d'un script d'analyse

### Avant (script Shell)

```
for sample in `ls *.fastq.gz`
do
 fastqc ${sample}
done
```



### Avantages

- Workflow (gestion des jobs)
- Puissant et rapide
- Capable d'utiliser des environnements Conda
- Parallélisable sur un cluster

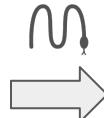
### Inconvénients

- Une logique à apprendre
- Syntaxe moins simple que le script shell

## Etape 4 - Crédation d'un script d'analyse

### Avant (script Shell)

```
for sample in `ls *.fastq.gz`
do
 fastqc ${sample}
done
```



### Après (Snakefile)

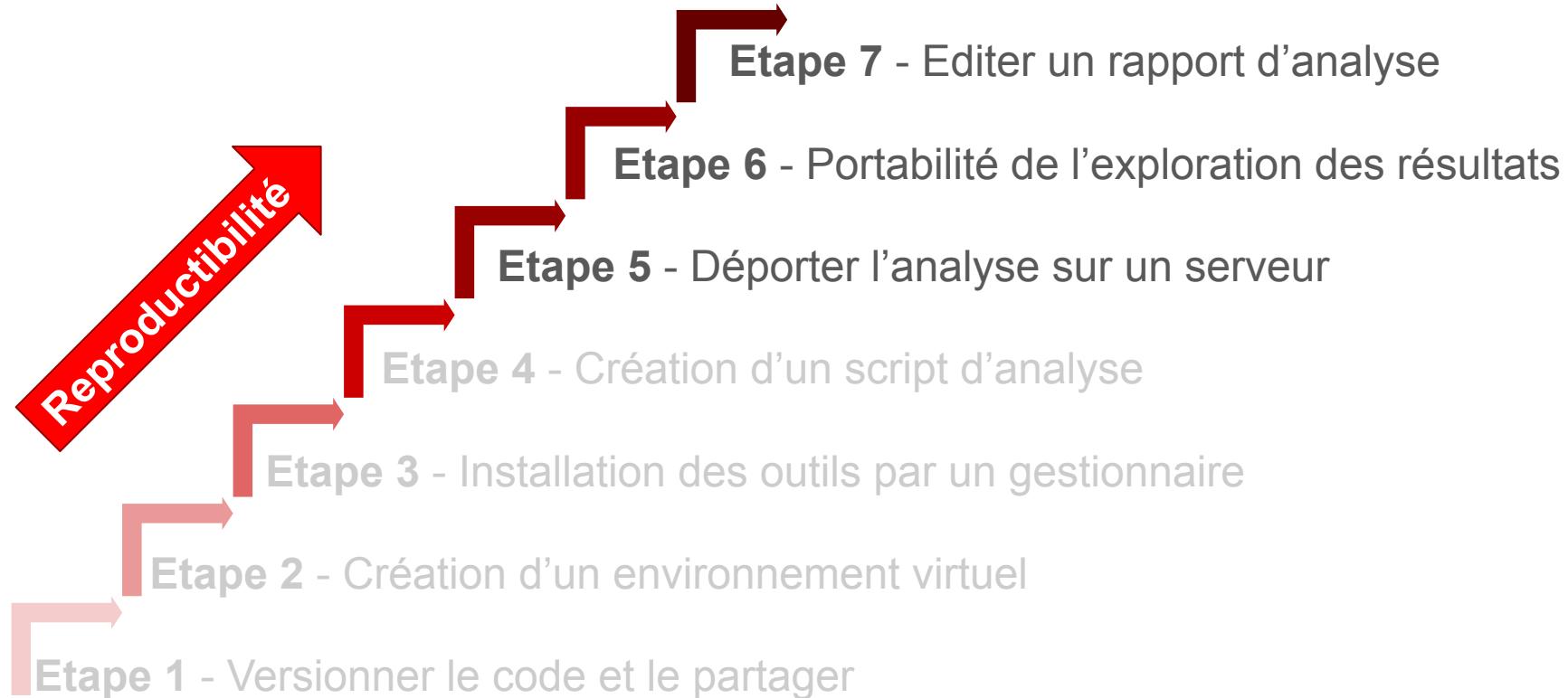
```
$ cat > Snakefile
SAMPLES, =
glob_wildcards("./samples/{smp}.fastq.gz")

rule final:
input:expand("fastqc/{smp}/{smp}_fastqc.zip",smp
=SAMPLES)

rule fastqc:
 input: "samples/{smp}.fastq.gz"
 output:"fastqc/{smp}/{smp}_fastqc.zip"
 message: """Quality check"""
 shell: """fastqc {input} --outdir
fastqc/{wildcards.smp}"""
$ snakemake
```

Plus court à écrire mais pas à exécuter !

## Où en sommes nous ?



## Etape 5 - Déporter l'analyse sur un serveur

Pourquoi ?

- Environnement contrôlé
- Déport de l'analyse



## Etape 5 - Déporter l'analyse sur un serveur

Avant

Adaptation en local et sur les serveurs difficile voire non gérée ...

## Etape 5 - Déporter l'analyse sur un serveur

Avant

Adaptation en local et sur les serveurs difficile voire non gérée ...



### Avantages

- Simple à mettre en place
- Augmentation de la puissance (cloud ou cluster)
- Pour tout le monde

### Inconvénients

- Pas simple pour les novices
- Attention aux données sensibles

## Etape 5 - Déporter l'analyse sur un serveur

Avant

Adaptation en local et sur les  
serveurs difficile voire non gérée ...



Après

```
$ git clone
https://github.com/thomasdenecker/FAIR_Bioinfo

$ cd FAIR_Bioinfo

$ sudo docker run --rm -d -p 80:8888 --name
fair_bioinfo -v ${PWD}:/home/rstudio
tdenecker/fair_bioinfo bash ./FAIR_script.sh
```

**Le protocole est lancé !**

## Etape 6 - Portabilité de l'exploration des résultats

Pourquoi ?

- Rendre simple l'exploration
- Simple à partager



## Etape 6 - Portabilité de l'exploration des résultats

Avant : Terminal R

```
dds <- DESeqDataSetFromMatrix(countData =
cts,colData = coldata, design= ~ batch +
condition)

dds <- DESeq(dds)
resultsNames(dds) # lists the coefficients
res <- results(dds, name =
"condition_trt_vs_untrt")

or to shrink log fold changes
association with condition:
res <- lfcShrink(dds,
coef="condition_trt_vs_untrt", type="apeglm")
```

## Etape 6 - Portabilité de l'exploration des résultats

Avant : Terminal R

```
dds <- DESeqDataSetFromMatrix(countData =
cts,colData = coldata, design= ~ batch +
condition)

dds <- DESeq(dds)
resultsNames(dds) # lists the coefficients
res <- results(dds, name =
"condition_trt_vs_untrt")

or to shrink log fold changes
association with condition:
res <- lfcShrink(dds,
coef="condition_trt_vs_untrt", type="apeglm")
```



### Avantages

- Portable (HTML)
- Accessible partout
- Interactif (paramétrable, graphes dynamiques, ...)

### Inconvénients

- Mélange de R et de HTML

# Toujours plus reproductible !

## Etape 6 - Portabilité de l'exploration des résultats

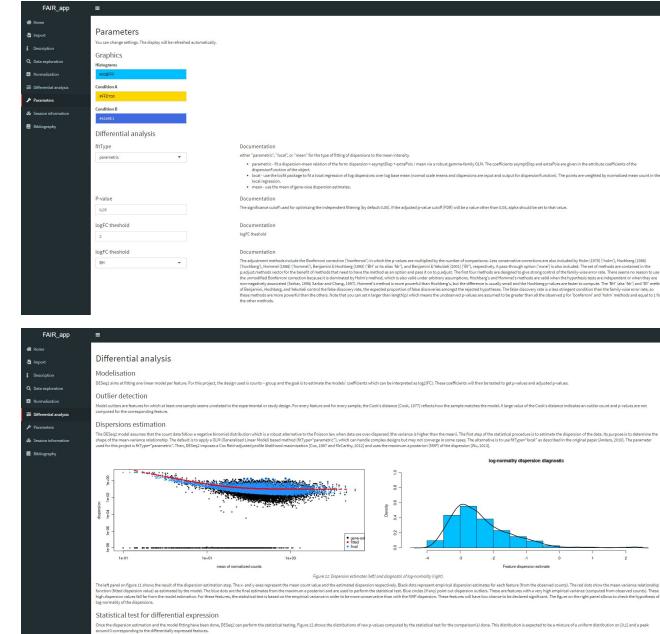


Avant : Terminal R

```
dds <- DESeqDataSetFromMatrix(countData =
cts,colData = coldata, design= ~ batch +
condition)

dds <- DESeq(dds)
resultsNames(dds) # lists the coefficients
res <- results(dds, name =
"condition_trt_vs_untrt")

or to shrink log fold changes
association with condition:
res <- lfcShrink(dds,
coef="condition_trt_vs_untrt", type="apeglm")
```



## Etape 7 - Editer un rapport d'analyse

Pourquoi ?

- Avoir une trace de l'analyse  
(date, heure, paramètres, ...)
- Stocker les versions des outils



*Toujours plus reproductible !*

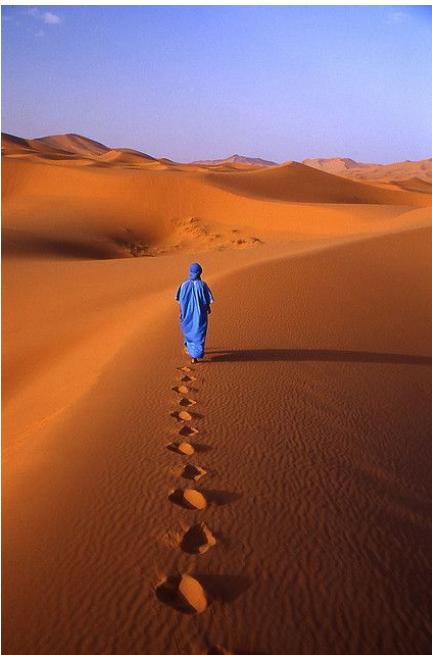
## *Etape 7 - Editer un rapport d'analyse*

Avant



## Etape 7 - Editer un rapport d'analyse

Avant



### Avantages

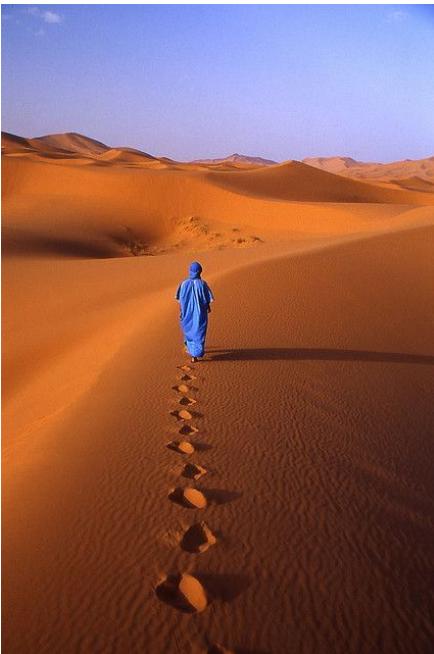
- Syntaxe simple (Markdown)
- Partage (PDF, HTML, ...)

### Inconvénients

- Rares problèmes de visualisation en  $\text{\LaTeX}$

## Etape 7 - Editer un rapport d'analyse

Avant



Après

### Statistical report of project Demo: pairwise comparison(s) of conditions with DESeq2

Thomas Denecker  
2020-11-12

The SARTools R package which generated this report has been developed at P2 - Institut Pasteur by M.-A. Dillies and H. Varet (<https://hugo.varet@pasteur.fr>). Thanks to cite H. Varet, L. Brillat-Guiguen, J.-Y. Coppee and M.-A. Dillies, SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data, PLoS One, 2016, doi: <http://dx.doi.org/10.1371/journal.pone.0157022> when using this tool for any analysis published.

#### 1 Introduction

The analysis reported in this document are part of the Demo project. The aim is to find features that are differentially expressed between STANDARD and DEPLETED. The statistical analysis process includes data normalization, graphical exploration of raw and normalized data, test for differential expression for each feature between the conditions, raw p-value adjustment and export of lists of features having a significant differential expression between the conditions.

The analysis is performed using the R software [1], Bioconductor [2] packages including DESeq2 [3,4] and the SARTools package developed at P2 - Institut Pasteur. Normalization and differential analysis are carried out according to the DESeq2 model and package. This report comes with additional tab-delimited text files that contain lists of differentially expressed features.

For more details about the DESeq2 methodology, please refer to its related publications [3,4].

#### 2 Description of raw data

The count data files and associated biological conditions are listed in the following table.

Table 1: Data files and associated biological conditions.

| label | files                    | iron     |
|-------|--------------------------|----------|
| S1    | SPR3099587_chr18_ftc.txt | STANDARD |
| S2    | SPR3099586_chr18_ftc.txt | STANDARD |
| S3    | SPR3099585_chr18_ftc.txt | STANDARD |
| D1    | SPR3105699_chr18_ftc.txt | DEPLETED |
| D2    | SPR3105698_chr18_ftc.txt | DEPLETED |
| D3    | SPR3105697_chr18_ftc.txt | DEPLETED |

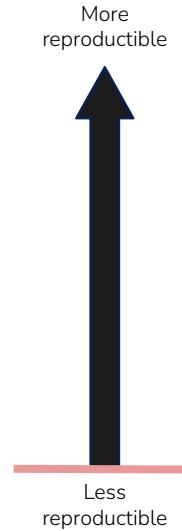
After loading the data we first have a look at the raw data table itself. The data table contains one row per annotated feature and one column per sequenced sample. Row names of this table are feature IDs (unique identifiers). The table contains raw count values representing the number of reads that map onto the features. For this project, there are 7659 features in the count data table.

Table 2: Partial view of the count data table.

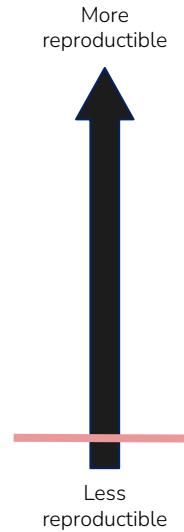
|                | S1 | S2 | S3 | D1 | D2 | D3 |
|----------------|----|----|----|----|----|----|
| costal01g00010 | 11 | 12 | 11 | 7  | 5  | 1  |

## *Conclusion*

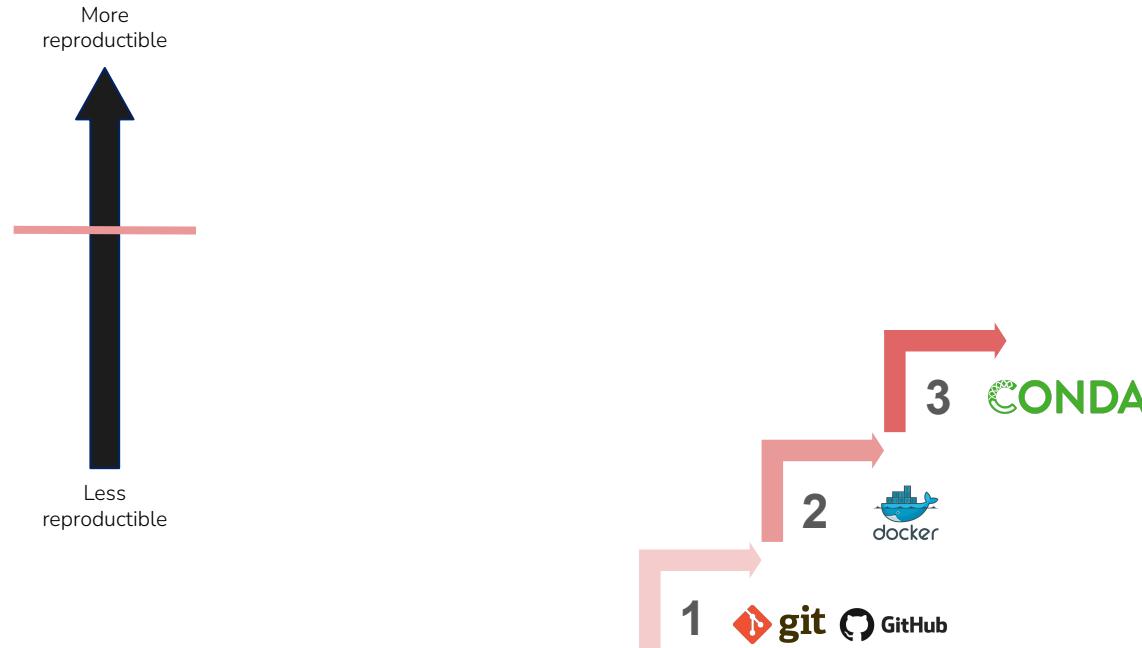
## Quel est notre niveau de reproductibilité ?



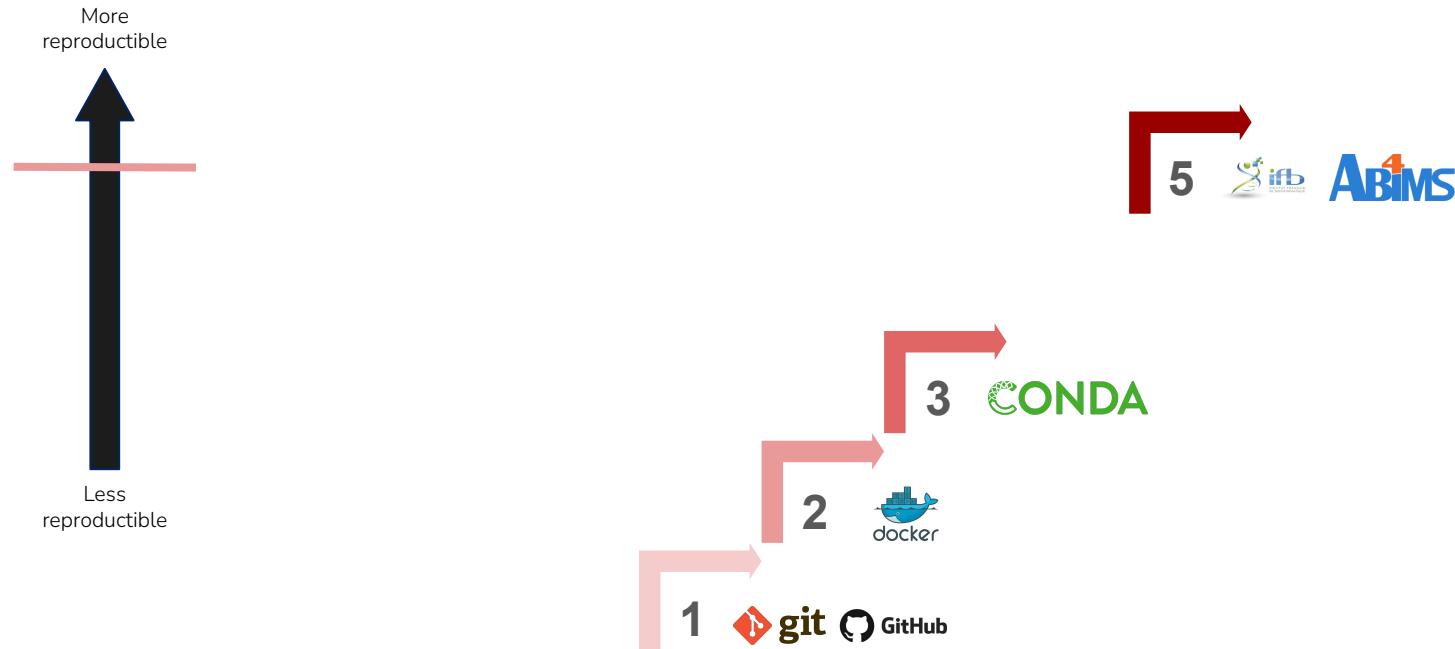
## Quel est notre niveau de reproductibilité ?



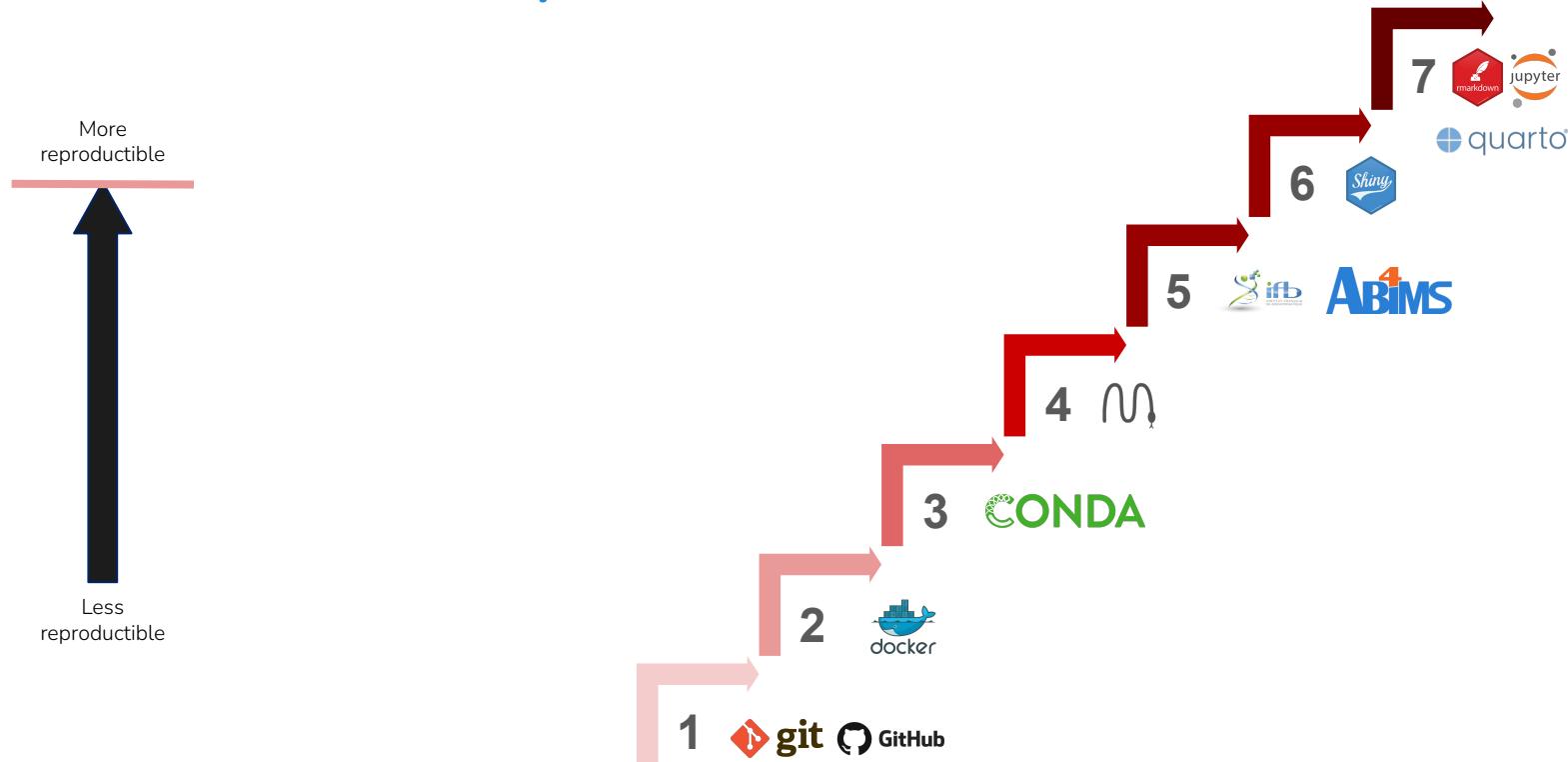
## Quel est notre niveau de reproductibilité ?



## Quel est notre niveau de reproductibilité ?



## Quel est notre niveau de reproductibilité ?



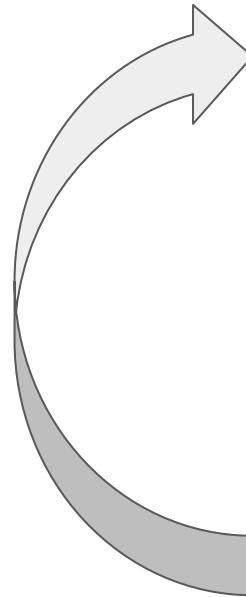
## Take home messages

Une vraie réflexion sur la reproductibilité des analyses en Bioinformatique

Proposition d'une solution qui aide à rendre reproductible n'importe quel protocole d'analyse

**La reproductibilité est une plus value pour la Bioinformatique !**

## Un cercle vertueux



FAIR raw data

+

FAIR\_bioinfo scripts/protocols

=

FAIR processed data

## FAIR\_bioinfo

- Gitlab formations IFB : <https://gitlab.com/ifb-elixirfr/training/fair-bioinfo>
- 2024 : <https://ifb-elixirfr.gitlab.io/training/fair-bioinfo/sessions/2024-10-fair-bioinfo-grand-ouest/programme>
- 2021 : <https://ifb-elixirfr.github.io/IFB-FAIR-bioinfo-training/session2021.html>
- 2020 : <https://ifb-elixirfr.github.io/IFB-FAIR-bioinfo-training/session2020.html>
- 2019 : [https://github.com/thomasdenecker/FAIR\\_Bioinfo](https://github.com/thomasdenecker/FAIR_Bioinfo)

## FAIR & le cluster de l'IFB

- Slurm : [https://ifb-elixirfr.github.io/IFB-FAIR-bioinfo-training/assets/pdf/Session2020/04\\_cluster.pdf](https://ifb-elixirfr.github.io/IFB-FAIR-bioinfo-training/assets/pdf/Session2020/04_cluster.pdf)
- Snakemake + Slurm :  
[https://ifb-elixirfr.github.io/IFB-FAIR-bioinfo-training/assets/pdf/Session2020/04\\_tp1\\_snakemake.pdf](https://ifb-elixirfr.github.io/IFB-FAIR-bioinfo-training/assets/pdf/Session2020/04_tp1_snakemake.pdf)
- Docker/Singularity :  
[https://ifb-elixirfr.github.io/IFB-FAIR-bioinfo-training/assets/pdf/Session2020/04\\_tp2\\_singularity.pdf](https://ifb-elixirfr.github.io/IFB-FAIR-bioinfo-training/assets/pdf/Session2020/04_tp2_singularity.pdf)

## Données

- Originale : <https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR4026187>
- Réduite : <https://doi.org/10.5281/zenodo.3997237>

# Conclusion

## Une belle équipe !

2019



T. Denecker



C. Toffano-Nioche

2020-2023



C. Hernandez



H. Chiapello



G. Le Corguillé



P. Lieby



Y. Mahmah



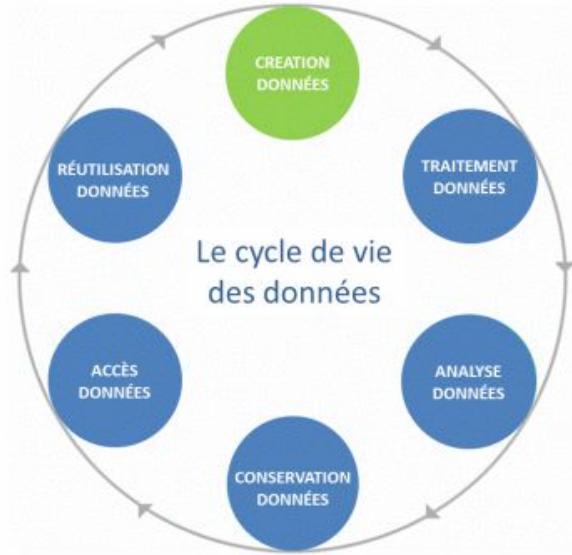
J. Seiler



J. van Helden

# *PARTIE II*

*Du point de vue des données*



Le modèle de UK Data Archive définit les six étapes suivantes :

- **Création ou collecte des données** (*creating data*) ;
- **Traitement des données** (*processing data*) ;
- **Analyse des données** (*analysing data*) ;
- **Conservation des données** (*preserving data*) ;
- **Accès aux données** (*giving access to data / data discovery*) ;
- **Réutilisation des données** (*reusing data*).

[Une introduction à la gestion et au partage des données de la recherche - Le cycle de vie des données](#)

## *Questions préliminaires*

## Qu'est ce qu'une donnée de recherche ?

In the context of the OECD Recommendation of the Council Concerning Access to Research Data from Public Funding (OECD, 2006), “research data” are defined as **factual records** (numerical scores, textual records, images and sounds) **used as primary sources for scientific research** and that are **commonly accepted in the scientific community** as necessary to validate research findings. A research data set constitutes a systematic, partial representation of the subject being investigated.

Scientific data are very diverse: they include observational data, which record natural phenomena (in fields such as astronomy, geoscience and demography); **experimental data**, which record the outcomes of man-made experiments, such as laboratory experiments in physics, chemistry and **biology**, or clinical trials; computational data, which are generated through large-scale simulations; and reference data, which are highly curated datasets, such as the human genome.

[OECD-ilibrary.org](https://OECD-ilibrary.org)

## Qu'est ce qu'une métadonnée ?

Les métadonnées sont des « données qui décrivent des données » :

- **Information** structurée associée à un "objet", un document ou un jeu de données
- **Documentation** qui permet à l'utilisateur de comprendre, de comparer et d'échanger le contenu du jeu de données décrit

Il existe des **standards** de métadonnées :

- Standards minimaux (ex : Dublin Core)
- Standards métiers (ex : EML, DDI...)

Il est conseillé de produire les métadonnées **au moment de la collecte ou de la création** des données plutôt qu'à posteriori. Les métadonnées seront complétées **tout au long du cycle de vie des données**.

# Questions préliminaires

## En résumé

### Les données



Les métadonnées

# Questions préliminaires

## Et dans la “vraie vie” ?

### Le passé

- Le leg (du doctorant précédent ...)
- La biblio à T0
- Les méthodes pré existantes

# Questions préliminaires

## Et dans la “vraie vie” ?

### Le passé

- Le leg (du doctorant précédent ...)
- La biblio à T0
- Les méthodes pré existantes

### Le présent

- Les manips
- La création de connaissance (méthodes, posters ...)

# Questions préliminaires

## Et dans la “vraie vie” ?

### Le passé

- Le leg (du doctorant précédent ...)
- La biblio à T0
- Les méthodes pré existantes

### Le présent

- Les manips
- La création de connaissance (méthodes, posters ...)

### Le futur

- Le manuscrit
- Les publications

## Et dans la “vraie vie” ?

### Le passé

- Le leg (du doctorant précédent ...)
- La biblio à T0
- Les méthodes pré existantes

### Le présent

- Les manips
- La création de connaissance (méthodes, posters ...)

### Le futur

- Le manuscrit
- Les publications

### Des échantillons

- dans les frigos
- dans les tiroirs

# Questions préliminaires

## Et dans la “vraie vie” ?

### Le passé

- Le leg (du doctorant précédent ...)
- La biblio à T0
- Les méthodes pré existantes

### Le présent

- Les manips
- La création de connaissance (méthodes, posters ...)

### Le futur

- Le manuscrit
- Les publications

### Des échantillons

- dans les frigos
- dans les tiroirs

### Des fichiers

- des petits, des gros
- un peu partout (PC, cloud, cluster)
- des données brutes, du code, des résultats

# Questions préliminaires

## Et dans la “vraie vie” ?

### Le passé

- Le leg (du doctorant précédent ...)
- La biblio à T0
- Les méthodes pré existantes

### Le présent

- Les manips
- La création de connaissance (méthodes, posters ...)

### Le futur

- Le manuscrit
- Les publications

### Des échantillons

- dans les frigos
- dans les tiroirs

### Des fichiers

- des petits, des gros
- un peu partout (PC, cloud, cluster)
- des données brutes, du code, des résultats

### De la connaissance

- des méthodes, du code
- des systèmes d'information
- des publications

# Questions préliminaires

*Mais au fait c'est quoi FAIR appliqué aux datas ?*

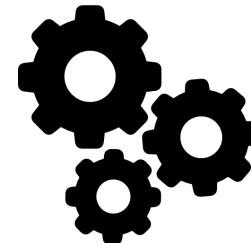
Findable



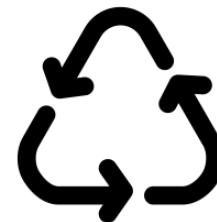
Accessible



Interoperable



Reusable



Sangya Pundir

## Findable -- Faciliter la découverte des données

Findable 

- Les données ont un **PID** (Persistent IDentifier ou identifiant pérenne en français)
- Les données sont décrites par des **métadonnées**
- Ces métadonnées doivent être liées aux PIDs des données
- Les données sont déposées dans un **entrepôt de données**

<https://doranum.fr/enjeux-benefices/principes-fair/>

# Accessible -- Permettre l'accès aux données et leur téléchargement

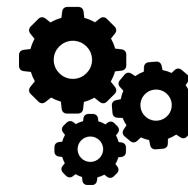
Accessible 

- Les données sont accessibles à travers un **protocole de communication standard**
- Ce protocole est **libre et ouvert**
- Ce protocole permet un accès par **authentification** si besoin
- Les **métadonnées restent accessibles** même si les données ne le sont plus

<https://doranum.fr/enjeux-benefices/principes-fair/>

# Interoperable -- Permettre l'exploitation des données quel que soit l'environnement informatique utilisé

## Interoperable



- Les données sont décrites avec un vocabulaire contrôlé
- Le vocabulaire utilisé respecte les principes FAIR
- Les métadonnées sont reliées à d'autres données

<https://doranum.fr/enjeux-benefices/principes-fair/>

# Reusable -- Permettre la réutilisation des données pour de futures recherches

R<sub>e</sub>usable 

- Les métadonnées ont une **pluralité d'attributs**
- Une **licence de réutilisation** est attribuée aux données
- La description des données indique leur **provenance**
- Le partage des données suit les **standards de la communauté scientifique**

<https://doranum.fr/enjeux-benefices/principes-fair/>

## En résumé



<https://doranum.fr/enjeux-benefices/principes-fair/>

## *En savoir plus*

# scientific **data**

The FAIR Guiding Principles for scientific data management and stewardship

Wilkinson et al., 2016

<https://doi.org/10.1038/sdata.2016.18>



<https://www.go-fair.org/fair-principles/>

# Exemple d'évaluation automatique par FAIR CHECKER

**Check**  
How FAIR is my resource ?

Enter resource identifier (URL/DOI)

Test all metrics

The URL/DOI is valid

Open results

Dataset Database Workflow Publication Dataset Dataset Tool

Radar chart of metrics completion

List of metrics with details and results

| Principle | Name                             | Description       | Comment              | Recommandation                                                                 | Score | Result               | Test               | Details        |
|-----------|----------------------------------|-------------------|----------------------|--------------------------------------------------------------------------------|-------|----------------------|--------------------|----------------|
| F1A       | Unique IDs                       | <span>Info</span> | <span>Comment</span> |                                                                                | 2     | <span>Success</span> | <span>Check</span> | <span>+</span> |
| F1B       | Persistent IDs                   | <span>Info</span> | <span>Comment</span> | To ensure that the used identification scheme is persistent, you should build  | 0     | <span>Failure</span> | <span>Check</span> | <span>+</span> |
| F2A       | Structured metadata              | <span>Info</span> | <span>Comment</span> |                                                                                | 2     | <span>Success</span> | <span>Check</span> | <span>+</span> |
| F2B       | Shared vocabularies for metadata | <span>Info</span> | <span>Comment</span> | You should express all your metadata with properties coming from               | 1     | <span>Success</span> | <span>Check</span> | <span>+</span> |
| A1.1      | Open resolution protocol         | <span>Info</span> | <span>Comment</span> |                                                                                | 2     | <span>Success</span> | <span>Check</span> | <span>+</span> |
| I1        | Machine readable format          | <span>Info</span> | <span>Comment</span> |                                                                                | 2     | <span>Success</span> | <span>Check</span> | <span>+</span> |
| I2        | Use shared ontologies            | <span>Info</span> | <span>Comment</span> | You should express all your metadata with properties coming from               | 1     | <span>Success</span> | <span>Check</span> | <span>+</span> |
| I3        | External links                   | <span>Info</span> | <span>Comment</span> |                                                                                | 2     | <span>Success</span> | <span>Check</span> | <span>+</span> |
| R1.1      | Metadata includes licence        | <span>Info</span> | <span>Comment</span> | You should include information about licence in your metadata using one of the | 0     | <span>Failure</span> | <span>Check</span> | <span>+</span> |
| R1.2      | Metadata includes provenance     | <span>Info</span> | <span>Comment</span> | You should express all your metadata with properties coming from               | 2     | <span>Success</span> | <span>Check</span> | <span>+</span> |
| R1.3      | Community standards              | <span>Info</span> | <span>Comment</span> |                                                                                | 1     | <span>Success</span> | <span>Check</span> | <span>+</span> |

<https://fair-checker.france-bioinformatique.fr>

Thomas Rosnet et al.

*Comment gérer les données ?*

## Comment le gérer ?

# Plan de gestion de données



Planifier et anticiper



Gérer, faire fructifier et ne pas les perdre



Data is the new (s)oil !

## Les objectifs du PGD

1. Assurer la reproductibilité des expériences (comment les données sont obtenues)
2. Respecter le droit et les personnes (clarifier le cadre juridique et éthique)
3. Permettre la réutilisation des données (Garantir la compréhension des données)
4. Éviter les pertes de données (Assurer un stockage adapté)
5. Établir le rôle de chacun (Définir les responsabilités)
6. Clarifier les droits de réutilisation (Spécifier les modalités de partage)

# Un exemple qui finira de vous convaincre : Le Spidergate

NEWS | 07 February 2020 | Clarification [13 February 2020](#)

## 'Avalanche' of spider-paper retractions shakes behavioural-ecology community

Allegations of fabricated data have prompted a university investigation and some soul-searching.

By [Giuliana Viglione](#)



Jonathan Pruitt (McMaster Univerty, CA)

- 1ère alerte en 2020 par Kate Laskowski
- Mise en évidence de nombreuses falsifications dans la donnée brute (patrons de duplications de cellules)
- 19 papiers rétractés dont sa thèse.

→ Manque de publication des données, manque de suivi, manque d'un PGD.

# Plan de gestion des données

## Les outils

Il existe plusieurs outils dont :

### DMP OPIDoR

Solution nationale

The screenshot shows the DMP OPIDoR interface. At the top, there's a navigation bar with links like 'Tableau de bord', 'Créer des plans', 'DMPs publics', 'Modèles de DMP', 'Aide', 'Plus', 'Français', and a user profile. Below the navigation is the CNRS logo. The main area has a title 'Titre du projet' with a placeholder 'projet de test, d'entraînement ou à des fins de formation'. A section titled 'Choisissez un modèle' explains how to choose a model from your organization or another, with a note about the default 'Science Europe - DMP template (english)'. It shows a dropdown menu with 'CNRS (Votre organisme)', 'Autre organisme', and 'Financier'. A message says 'Aucun modèle de disponible.' At the bottom are buttons for 'Créer un plan', 'Suivant', and 'Utiliser le modèle par défaut'.

### DSW - Data Stewardship Wizard

Solution européenne (ELIXIR)

The screenshot shows the DSW interface for document '20210224\_exemple\_canevas\_IFB\_bioimage'. The left sidebar has sections for 'Knowledge Model Editor', 'Knowledge Models', and 'Projects'. The main area is titled 'Chapters' with sections I, II, III, and IV. Chapter IV is selected and titled 'IV. Données de la recherche'. To the right, a large panel displays a questionnaire step with a question 'Quelles mesures de contrôle de la qualité sont prises pour ce jeu de données ?' and a list of options: a. Illumination power, b. Detection system performance, c. Field of view uniformity.flatness, d. Chromatic aberrations, e. Lateral and axial resolution, f. Image quality, and g. Commentaires. At the bottom, a note asks 'Des versions différentes du jeu de données sont-elles créées ?'.

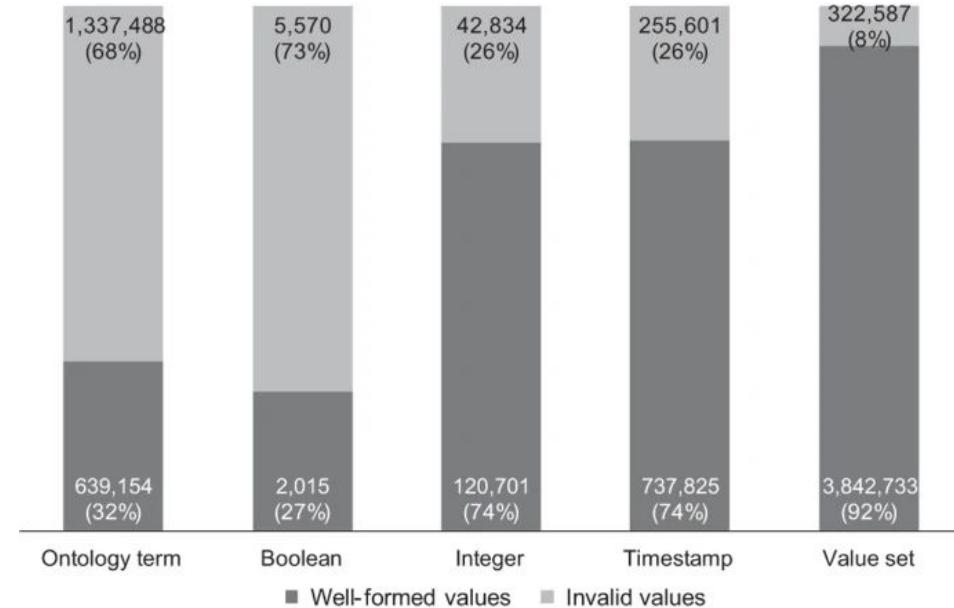
*Comment bien décrire les données ?*

## Bilan de la qualité des métadonnées

Les métadonnées demandées sont différentes entre les bases de données et souvent complexes

La soumission est hétérogène

Les métadonnées sont souvent incomplètes, inconsistantes, redondantes et tout simplement pas assez informatives



Quality of dictionary attributes in NCBI BioSample  
according to their type, in [Gonçalves et al., 2019](#)

## Utilisation de standards

### Définition

A standard provides the **requirements, specifications, guidelines or characteristics** that can be used for the **description, interoperability, citation, sharing, publication, or preservation** of all kinds of **digital objects** such as data, code, algorithms, workflows, software, or papers.

source: <https://fairsharing.org/educational/>

### Comment trouver le bon ?

Sansone, et al. FAIRsharing as a community approach to standards, repositories and policies.

Nat Biotech. 2019 <https://doi.org/10.1038/s41587-019-0080-8>

The screenshot shows the FAIRsharing.org website, a curated resource for data and metadata standards. The header includes a search bar, navigation links for Standards, Databases, Policies, Collections, Add/Claim Content, Stats, and Log in or Register. A main banner states: "A curated, informative and educational resource on data and metadata standards, inter-related to databases and data policies." Below this is a section titled "HOW CAN WE HELP?" with a sub-section for "Research data facilitators, librarians, trainers". It features icons for Researchers, Developers & Curators, Journal Publishers, Librarians & Trainers (highlighted), Societies & Alliances, and Funders. The "Librarians & Trainers" section includes a "Find" button, "Discover" button, and "Learn" button. The "Find" section has a "Recommendations" sub-section with a "Collections" icon. The "Discover" section has a "Collections" sub-section with a "Educational" icon. The "Learn" section has an "About standards, their use in databases and policies, and how we can help you." sub-section. At the bottom is a footer with a search bar, Advanced Search, Search Wizard, and a "Fine grained control over your data" link.

## Exemple de standard : Genomic standards consortium

Producteur de Minimum Information  
Standards utilisés  
par l'ENA (EBI) et SRA (NCBI)

Notion de checklists sur l'ENA

<https://www.ebi.ac.uk/ena/browser/checklists>

|                                                                   |                                                                                                                                                                                                                                                                                     |    |    |                                                                                                                                                |     |             |         |          |                   |
|-------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|----|------------------------------------------------------------------------------------------------------------------------------------------------|-----|-------------|---------|----------|-------------------|
| Specification projects                                            | MIGS                                                                                                                                                                                                                                                                                |    |    |                                                                                                                                                |     | MIMS        | MIMARKS |          | New checklists    |
| Checklists                                                        | EU                                                                                                                                                                                                                                                                                  | BA | PL | VI                                                                                                                                             | ORG | metagenomes | survey  | specimen | e.g., pan-genomes |
| Shared descriptors                                                | collection date, environmental package, environment (biome), environment (feature), environment (material), geographic location (country and/or sea, region), geographic location (latitude and longitude), investigation type, project name, sequencing method, submitted to INSDC |    |    |                                                                                                                                                |     |             |         |          |                   |
| Checklist-specific descriptors                                    | assembly, estimated size, finishing strategy, isolation and growth condition, number of replicons, ploidy, propagation, reference for biomaterial                                                                                                                                   |    |    | target gene                                                                                                                                    |     |             |         |          |                   |
| Applicable environmental packages (measurements and observations) | Air<br>Host-associated<br>Human-associated<br>Human-oral<br>Human-gut<br>Human-skin<br>Human-vaginal                                                                                                                                                                                |    |    | Microbial mat/biofilm<br>Miscellaneous natural or artificial environment<br>Plant-associated<br>Sediment<br>Soil<br>Wastewater/sludge<br>Water |     |             |         |          |                   |

Yilmaz et al, 2011

*Soumettre les données*

## Pourquoi soumettre les données et les métadonnées ?

- Pour l'Open Science et la reproductibilité des expériences
- Pour être FAIR et donner accès aux données
- Pour l'archivage
- Pour les publications
- Pour l'analyse avec par exemple MGinify, GEotoR, ...

The screenshot shows the GEOtoR platform interface. At the top, there's a red banner with COVID-19 information. Below it, the URL is GSE10244. The main search bar contains "GSE10244 Stress response in yeast". To the right, there are links for GEO Publications, FAQ, Help, and Log in. A "Text search" button is highlighted in green. Below the search bar, there's a "Select a platform" dropdown set to "Platform". A "Quick start" section provides instructions for comparing samples. At the bottom, there are "How to use" and "Analyze" buttons.

The screenshot shows the MGinify homepage. It features a dark header with the MGinify logo and a sub-header: "Submit, analyse, discover and compare microbiome data". Below the header are navigation links: Overview, Submit data, Text search, Sequence search, Browse data, Genomes, API, and About. A "Getting started" link is also visible. The main area has two search sections: "Search by Name, biome, or keyword" (with a "Text search" button) and "Sequence similarity" (with a "Sequence search" button). Below these are sections for "Or by data type" and "Or by selected biomes". The "Or by data type" section displays statistics for various study types: amplicon (350874), assemblies (30079), metabarcoding (2050), metagenomes (32855), and metatranscriptomes (2204). The "Or by selected biomes" section shows icons for Human (141077), Digestive system (44755), Aquatic (44755), Marine (32218), and another Digestive system (44755). On the right side, there are links for "Latest", "metab", "cryoconites", "The Ecosystem derived from This project", and "View more -". At the bottom, there's a link for "EMG p Salivar".

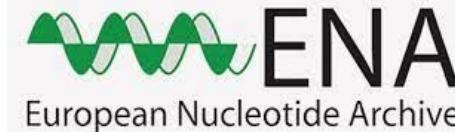
# Soumission des données

## Soumission très hétérogène

Par simple fichier Excel



Un peu plus complexe



MAIS

avec une qualité des métadonnées  
bien supérieure

## Data brokering

Proposer une solution pour fluidifier la soumission des données

Des outils sont proposés dans des branches particulières

L'IFB souhaite offrir une solution nationale divisée en 3 activités

- Développement d'outils
- Formation
- Support aux utilisateurs



## *Questions juridiques*

## Quelles obligations de partage des données ?

Les données de la recherche sont des informations publiques :

- Elles sont soumises à un principe d'**ouverture par défaut** et de **libre utilisation** (Loi Lemaire - Loi République numérique 2016 LPRN)
- Elles sont soumises à un **principe de gratuité** (Loi Valter 2015)
  - Cas particulier de Météo France et IGN
  - Spécificité des brevets et autres formes de valorisation
- Elles sont **protégées contre les risques d'accaparement**

## *Et les principes FAIR ?*

Aussi ouvert que possible, aussi fermé que nécessaire

## *Des exceptions ?*

- Le droit d'auteur comme dans les publications scientifiques, logiciels, ...
- Les projets en partenariat avec le privé
- Les données personnelles soumises à la RGPD (sauf avec un consentement, anonymisation ou dérogations)
- Les données sensibles comme la biodiversité (orchidée)
- Secret médical, secret d'affaires, secret militaire, secret des procédés,...

## Licence

Moyen d'encadrer le partage et la réutilisation des données

Par forcément nécessaire mais fortement recommandé dans tous les cas

Liste des licences et explication : <https://www.data.gouv.fr/fr/licences>

## Modalité de partage

- Considérer les restrictions, embargo et limites de réutilisation
- Se renseigner sur les obligations de partages spécifiques au bailleurs
- Identifier les jeux de données partageables ou non
- Identifier les futurs utilisateurs
- Déterminer quand partager
- Déterminer où partager en fonction des données, des bailleurs, ...

# Partage des données

## Exemple d'entrepôts de données

### Thématique



Trouver le bon ? <https://www.re3data.org/> ou <https://repositoryfinder.datacite.org/>

# Partage des données

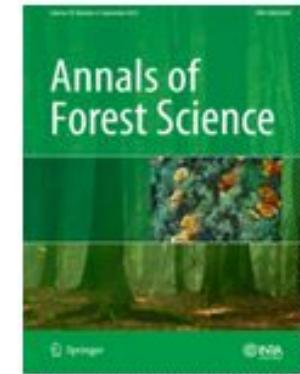
## Toujours chercher à valoriser les données

Publier un **datapaper**

Publier un **article de recherche** (dans un PCI ?)

Rédiger une brève pour un **magazine** spécialisé

Contribuer à un **blog**, ....



BMC Research Notes

Open Data Journal for Agricultural Research

## *Bonnes pratiques*

# Bonnes pratiques

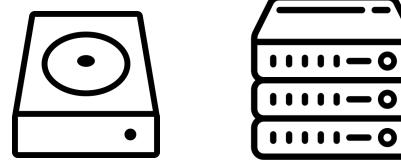
## *Un environnement de travail sûr*

### Sauvegardé

Stratégie 3 2 1



3 copies



2 systèmes



dont 1 distante

### Protégé

## Le stockage

Nombreuses méthodes et technologies de stockage des données

- Disque dur
- Clé USB
- Cloud

### Vérifier l'intégrité des données lors de transfert

Il est possible de contrôler l'intégrité des données avec par exemple le **md5sum**

## Les fichiers : nommage et format

### Nommage

- Bref et explicite
- Sans espace ni caractères spéciaux
- Avec une date au bon format
- Avec l'élément le plus important en premier
- Avec la version du document

### Format

Si possible non propriétaire

Les formats qui perdent le moins de données à la conversion

Le format utilisé par la communauté

#### PUBLIC SERVICE ANNOUNCEMENT:

OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.

THIS IS THE CORRECT WAY TO WRITE NUMERIC DATES:

2013-02-27

THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:

02/27/2013 02/27/13 27/02/2013 27/02/13

20130227 2013.02.27 27.02.13 27-02-13

27.2.13 2013. II. 27. 2 $\frac{1}{2}$ -13 2013.158904109

MMXIII-II-XXVII MMXIII  $\frac{LVII}{CCCLXV}$  1330300900

((3+3)×(111+1)-1)×3/3-1/3<sup>3</sup> 2013 Mississipi

10/11011/1101 02/27/20/13 01237 2-27-13



<https://xkcd.com/1179/>

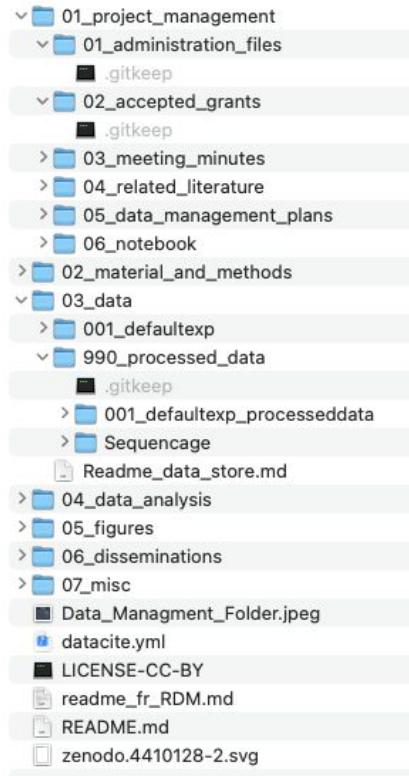
## Organisation des données

### Organisation des dossiers

- Limitez le nombre de dossiers par niveau (5 ou 6 max)
- Allez du général au spécifique
- Choisissez des noms de dossiers explicites

Avec un **README** pour décrire le contenu (txt ou md)

# Focus - Organisation des dossiers



DOI 10.5281/zenodo.4410128

**GOAL of good data management**  
→ optimise the discovery & reuse of data

**Questions to ask yourself**

- Are my files organised in a way that I can easily find what I am searching for?
- What information would I need to understand and use my data in 20 years?
- Could others understand and use my data?

**Folder structure**

- number your main folders
- keep your raw data separately
- ref 1: pipeline structure
- separate ongoing & completed work
- Readme.txt project description

**experimental data can be sorted by date**

**File naming**

P05\_RNAseq-bat3\_v03-02\_20210121\_KH.csv

- major change
- minor change
- YYYY-MM-DD
- YYYY-YYYY
- project number/acronym\*
- describing name
- version\*
- date
- initials creator
- \*if applicable

**General naming tips for folders & files**

- use unique, meaningful names
- not too long (not >30-40 characters)
- no spaces, dots, or special characters ('%', '\*' & '^'|+=[]|-\_=])
- hyphens (-) & underscores (\_) to separate elements

**Friendly Reminder**  
Comment your code!

**Metadata**

Which information is necessary to interpret, understand, and use a given dataset?

readme.txt files

can be used to describe projects, folders, and files

**References**

- <https://dataverse.harvard.edu/guides/how-to-keep-your-research-project-organized-part-5-folder-structure>
- <https://www.w3.org/Value-Creation-Congress/WoCC/DataManagement/WoCCOrganizingFilesAndFolders.htm>
- <https://www.mssm.edu/research/working-with-data/organizing-data>
- <https://library.bath.ac.uk/research-data/working-with-data/organising-data>
- <https://www.ncbi.nlm.nih.gov/pmc/research/organizing-data-folders-with-folders-method>
- <https://mantra.edina.ac.uk/>
- <https://old.datameo.org/education/modules>

**Important information**

- Who has created the data?
- What is the content of the data?
- Which questions have been answered?
- When were the data created?
- How were the data developed (methods)?
- Why were the data developed?
- With whom can the data be shared?

@Kira Höfler

# Focus - Organisation des dossiers

Accompagné d'un bon README (ici une proposition de recherche.data.gouv)

```
<Les textes d'aide sont écrits entre chevrons et sont destinés à être supprimés avant toute sauvegarde>

<Un README : Pourquoi ?>

La documentation d'un jeu de données doit être suffisante pour permettre à n'importe quel réutilisateur de comprendre et d'évaluer sa qualité. Le README fournit des informations complémentaires et accessibles lorsqu'elles ne sont pas déjà mises à disposition dans les métadonnées du jeu de données, dans les métadonnées des fichiers, et/ou dans des fichiers associés, ou des fichiers accessibles à long terme sur des services d'hébergement (entrept de fichiers ou publication). Dans ce dernier cas, nous vous prions d'inclure les URLs des documents en question ou leurs références

<Privilégier les formats text document (.txt), ou markdown (.md)>

Modèle de Fichier RDG README --- Général --- Version: 0.1 (2022-11-22)

Ce fichier README a été généré le [YYYY-MM-DD] par [NAME].

Dernière mise-à-jour le : [YYY-MM-DD].

INFORMATIONS GENERALES

Titre du jeu de données :

DOI:

Adresse de contact :

<Ci-après suit une liste d'éléments suggérés pour vous aider à enrichir, si nécessaire, votre documentation. La pertinence de certains dépend de la discipline du jeu de données ou du contexte de production>

<***Supprimer toute section non-applicable***>

INFORMATIONS METHODOLOGIQUES

Conditions environnementales / expérimentales :

Description des sources et méthodes utilisées pour collecter et générer les données :
<Si applicable, décrire les standards, les informations de calibration, les instruments utilisés, etc.>

Méthodes de traitement des données :
<Si applicable, décrire le traitement des données et inclure tout détail pouvant être important pour réutiliser ou reproduire les données. Commenter chaque étape.
Par exemple, inclure les méthodes de nettoyage et d'analyse ; les codes et/ou algorithmes ; les procédés d'anonymisation ou de pseudonymisation pour les données sensibles concernant les humains ou des espèces menacées>

Procédures d'assurance qualité appliquées sur les données :

Autres informations contextuelles :
<Toute information que vous considérez importante pour évaluer la qualité du jeu de données ou pour sa réutilisation : par exemple, des informations concernant les logiciels nécessaires pour interpréter les données.
Si applicable et non-inclus préalablement, ajouter les noms complets et les versions de tous les logiciels, de tous les paquets et de toutes les bibliothèques nécessaires pour lire et interpréter les données *e.g.* pour compiler les scripts.>
```

```
APERCU DES DONNEES ET FICHIERS

Convention de nommage des fichiers :

Arborescence/plan de classement des fichiers :

INFORMATIONS SPECIFIQUES AUX DONNEES POUR : {NOM DU FICHIER}

<Le cas échéant, reproduire cette section pour chaque dossier ou fichier.
Les éléments se répétant peuvent être expliqués dans une section initiale commune.>

<Pour les données tabulaires, fournir un dictionnaire des données/manuel de codage contenant les informations suivantes :>

Liste des variables/entêtes de colonne :

Pour chaque nom de variable ou entête de colonne, indiquer :

-- le nom complet de la variable sous forme "lisible par les humains" ;
-- la description de la variable ;
-- unité de mesure, si applicable ;
-- séparateur décimal *i.e.* virgule ou point, si applicable ;
-- valeurs autorisées : liste ou plage de valeurs, ou domaine ;
-- format, si applicable, e.g. date>

Code des valeurs manquantes :
<Définir les codes ou symboles utilisés pour les valeurs manquantes.>

Informations additionnelles :
<Toute information que vous jugez utile pour mieux comprendre le fichier>
```

<https://recherche.data.gouv.fr/fr/categorie/33/quide/modele-de-readme>

*Et maintenant ?*

# *Vous souhaitez vous formez*

- Catalogue des formations IFB (PFs et nationales) :  
<https://www.france-bioinformatique.fr/formations/>
  - Catalogue des formations au niveau ELIXIR/Europe :  
<https://tess.elixir-europe.org/events>
  - Ressources pédagogiques utiles :
    - <https://doranum.fr/>
    - <https://moodle.france-bioinformatique.fr/course/index.php?categoryid=2>
  - Liste des ateliers de la donnée (il y en a peut-être près de chez vous!) :  
<https://recherche.data.gouv.fr/fr/ateliers-de-la-donnee>
- Pour un conseil plus personnalisé, demander gentiment à Lucie (en la payant en chocolat ou en kouignette 😊)

Merci

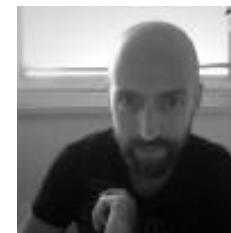
# Une belle équipe !



H. Chiapello



T. Denecker



J-F Dufayard



F. de Lamotte



P. Lieby



Y. Mahmah



G Sarah



J. Seiler

*Si la reproductibilité vous intéresse ...*

Permet d'avoir une vue d'ensemble de la gestion de la reproductibilité dans les autres domaines scientifiques

Revoir les JRFRR 2025 : [Conférence | Canal U](#)

*Comment FAIRe ? Et si on essayait !*

## Création d'un PGD



### Titre du projet

EBAII N2

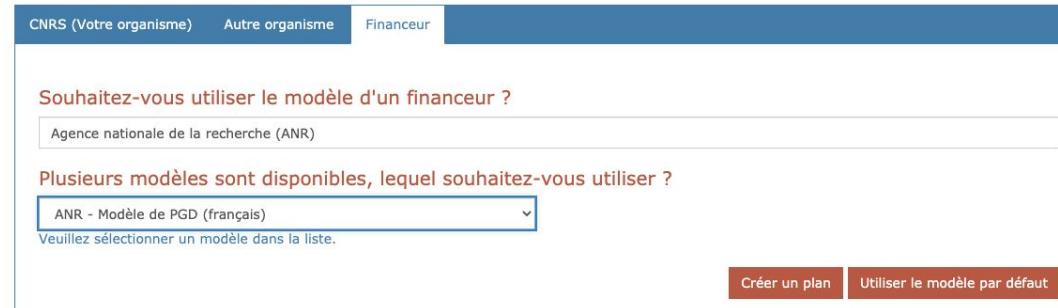
projet de test, d'entraînement ou à des fins de formation

<https://dmp.opidor.fr/plans/9346>

### Choisissez un modèle

Vous pouvez choisir soit un modèle fourni par votre organisme soit par un autre organisme, ou un modèle financeur. Le modèle par défaut est **Science Europe - DMP template (english)**.

Retrouvez la liste des modèles disponibles



Souhaitez-vous utiliser le modèle d'un financeur ?

Agence nationale de la recherche (ANR)

Plusieurs modèles sont disponibles, lequel souhaitez-vous utiliser ?

ANR - Modèle de PGD (français)

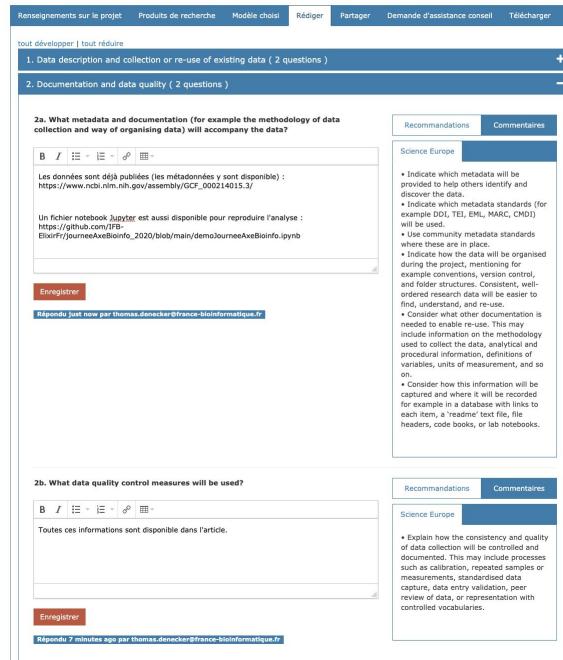
Veuillez sélectionner un modèle dans la liste.

Créer un plan   Utiliser le modèle par défaut

## Comment reproduire l'analyse ?



EBAII N2



2a. What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data?

Les données sont déjà publiées (les métadonnées y sont disponibles) : [https://www.ncbi.nlm.nih.gov/assembly/GCF\\_000214015.3/](https://www.ncbi.nlm.nih.gov/assembly/GCF_000214015.3/)

Un fichier notebook Jupyter est aussi disponible pour reproduire l'analyse : [https://github.com/IFB-ElixirFr/journeeAxelBioInfo\\_2020/blob/main/demoJourneeAxelBioInfo.ipynb](https://github.com/IFB-ElixirFr/journeeAxelBioInfo_2020/blob/main/demoJourneeAxelBioInfo.ipynb)

Enregistrer

Répondre Just now par thomas.denecker@france-bioinformatique.fr

2b. What data quality control measures will be used?

Toutes ces informations sont disponibles dans l'article.

Enregistrer

Réponse 7 minutes ago par thomas.denecker@france-bioinformatique.fr

[https://github.com/IFB-ElixirFr/FAIR\\_EBAII\\_n2/blob/master/demoFAIR\\_EBAII\\_n2.ipynb](https://github.com/IFB-ElixirFr/FAIR_EBAII_n2/blob/master/demoFAIR_EBAII_n2.ipynb)

## Reproduire l'analyse sur JupyterLab

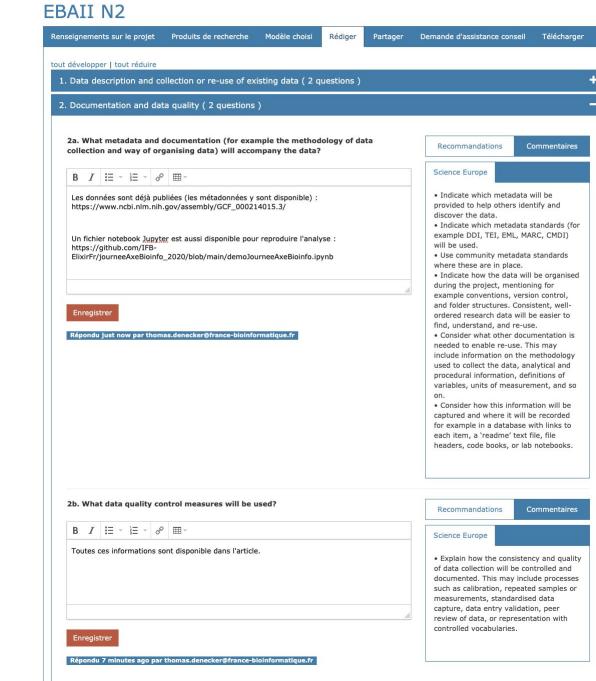


The screenshot shows the DataOPDSR platform interface. At the top, there are navigation links: Tableau de bord, Créer des plans, DMPs publics, Modèles de DMP, Aide, Plus, Français, and Thomas Denecker. Below this is the CNRS logo. The main area is titled "EBAII N2" and contains a form with sections for "Renseignements sur le projet", "Produits de recherche", "Modèle choisi", "Rédiger", "Partager", "Demande d'assistance conseil", and "Télécharger". The "Rédiger" tab is active. The form includes sections for "tout développer" and "tout réduire", and questions 1 and 2 about data description and documentation. It also includes sections for "2a. What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data?" and "2b. What data quality control measures will be used?". Each section has a "Recommendations" tab and a "Commentaires" tab. There are "Enregistrer" and "Répondre" buttons at the bottom of each section.



<https://ondemand.cluster.france-bioinformatique.fr>

Attention, prendre 12 CPUs et 10G de mémoire



This is a screenshot of the EBAII N2 form on the DataOPDSR platform. The "Rédiger" tab is selected. The form consists of two main sections: "1. Data description and collection or re-use of existing data ( 2 questions )" and "2. Documentation and data quality ( 2 questions )".

**Section 1: 1. Data description and collection or re-use of existing data ( 2 questions )**

Question 1a: "What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data?"

Question 1b: "What data quality control measures will be used?"

**Section 2: 2. Documentation and data quality ( 2 questions )**

Question 2a: "What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data?"

Question 2b: "What data quality control measures will be used?"

Both sections include "Recommendations" and "Commentaires" tabs. Buttons for "Enregistrer" and "Répondre" are visible at the bottom of each section.

## Reproduire l'analyse en un click !

### Au programme

- Notebook Jupyter (kernel bash)
- Téléchargement des données
- Import des scripts et des fichiers de paramétrage
- Réalisation des l'analyse
  - En parallèle
  - Sur le cluster
  - De façon reproductible
- Exploration préliminaire des résultats

Les principes FAIR appliqués à la bioinformatique  
Outils et environnements FAIR pour la bioanalyse  
Thomas Denecker - Data brokering - IFB core

Rejouer son article en un click

Cas d'étude

Contexte

L'objectif de l'étude est d'étudier la réponse à une privation en fer chez l'algue verte *Ostreococcus tauri* Lelandais et al, 2016. Il s'agit d'un organisme de 13.0328 Mb séparé en 20 chromosomes.

L'expérience est composée de 16 échantillons de données RNAseq (triplicat, single-end de 100bp). Le plan d'expérience est illustré ci-dessous. Pour accompagner cette démonstration, nous avons choisi les échantillons S11 et S12 (réponse adaptative à long terme),

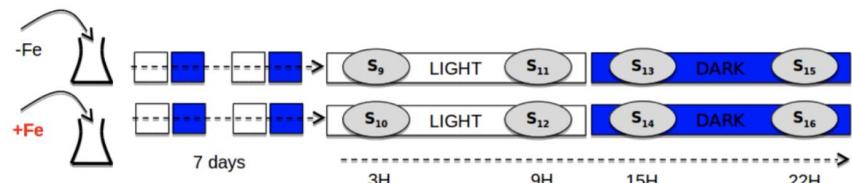


Figure 1 | Plan expérimental

## Partager ses résultats avec un notebook interactif

### Au programme

- Notebook Jupyter (kernel R)
- Import des résultats
- Réalisation d'un tableau explorable avec datatable
- Réalisation d'un volcano plot avec plotly. Lors du survol d'un point, des informations sont retournées

Les principes FAIR appliqués à la bioinformatique  
Notebook R interactif pour le partage des résultats  
Thomas Denecker - Data brokering - IFB core

Un notebook R interactif

Chargement des packages

```
options(warn=-1)
library(plotly)
library(dplyr)
library(gridExtra)
```

Lectures du tableau de résultats de SARtools

```
dataSarTools <- read.csv("backup/tables/DEPLETEDv5STANDARD.complete.txt", sep = "\t", header = T, stringsAsFactors = F)
dataSarTools <- dataSarTools %>%
 filter(log2FoldChange != "") %>%
 arrange(desc(log2FoldChange)) %>%
 mutate(log2FoldChange = as.numeric(log2FoldChange),
 pvalue = as.numeric(pvalue))
```

datatable

```
Documentation : https://rstudio.github.io/DT/

datatable(dataSarTools, {" ", "log2FoldChange", "pvalue"}), extensions = 'Buttons', filter = 'top', options = list(dom = 'Bfrtip', buttons = c('copy', 'csv', 'excel', 'pdf', 'print'), pageLength = 5, autoWidth = TRUE)) |
```

Volcano plot avec Plotly

```
Documentation : https://plotly.com/r/

inter = cbindx = dataSarTools[,c("log2FoldChange",
 y = -log10(as.numeric(pvalue)),
 pvalue = dataSarTools$pvalue,
 feature = dataSarTools$Id)]
inter = as.data.frame(inter)
inter[,1] = as.numeric(as.character(inter[,1]))
inter[,2] = as.numeric(as.character(inter[,2]))
inter[,3] = as.numeric(as.character(inter[,3]))
color = rep("black", nrow(inter))
color[pos] = "red"
color[pos] = "red"

plot_ly(inter, x = ~x, y = ~y, type = "scatter", mode = 'markers',
 text = ~paste("feature:", feature, "\n", "-log10(pValue):", pvalue),
 marker = list(color = color)) %>%
 layout(xaxis = list(type = "log", title = "log2(Fold change)", ticktext = c("1", "10", "100", "1000")),
 yaxis = list(type = "log", title = "-log10(adjusted pvalue)", ticktext = c("1", "10", "100", "1000")),
 xaxis = list(zeroline = FALSE, title = "log2(Fold change)"),
 yaxis = list(zeroline = FALSE, title = "-log10(adjusted pvalue)"))
```

sessionInfo()

# *Snakemake ou Nextflow ?*

Sources :

- <https://academic.oup.com/bib/article/18/3/530/2562749>
- <https://labs.epi2me.io/snakefile-vs-nextflow/>

Même si les deux langages sont assez éloignés, leur fonctionnement et leur intérêt sont très proches.

	Snakemake	Nextflow
Langage	Python	Groovy
Parallélisation automatique des tâches	Oui	Oui
Gestion des noms input/output	Définition explicite	Dénomination automatique
Test du pipeline	Dryrun	Sur un petit jeu de données
Installation	Simple, possible avec conda	Simple, possible avec conda
Communauté	Importante	Importante + Nf-core
Portabilité	Docker et Singularity (Podman en cours d'ajout)	Docker, Singularity et Podman

Il existe de nombreux autres créateur de pipeline, notamment Galaxy qui reste parmi les alternatives les plus utilisées.

Sa plus grande limite est que seul les admins system sont autorisés à y ajouter des outils.

# Podman ou Docker ?

Sources :

- <https://www.ionos.fr/digitalguide/serveur/know-how/podman-vs-docker/>
- <https://phoenixnap.com/kb/podman-vs-docker>

## Apparition de Podman

Le support officiel de Docker a été abandonné il y a quelques temps, petit à petit il deviendra inutilisable sur divers système (+ manque de maintenance)

Podman a donc été créé pour contrer cette perte avant tout.

## Avantages et inconvénients

## Podman, c'est quoi ?

Comme Docker, c'est une solution de conteneurisation légère, pratique et fiable. Elle ne nécessite pas les priviléges root, augmentant sa sécurité. Les commandes de base de Docker sont les mêmes dans Podman (run, build, pull, push, ps, images etc)

	Docker	Podman
Présence d'un daemon	Oui	Non
Démarrage	Classique	Plus rapide
Priviléges	Root	Aucun
Sécurité	Aucune	Plus forte
Compatibilité	Aucune	Avec Docker et autres systèmes de conteneurisation
Kubernetes	Non compatible	Compatible