Rencontre JCJC Développement
Développement logiciel pour la recherche

Science ouverte et principes FAIR dans un projet de bioinformatique

Comment rendre un projet bioinformatique plus reproductible ?

Thomas Denecker

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 — remixer, 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'Œuvre, 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.

Un contenu trouvable simplement, accessible, décrit et réutilisable

figshare

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

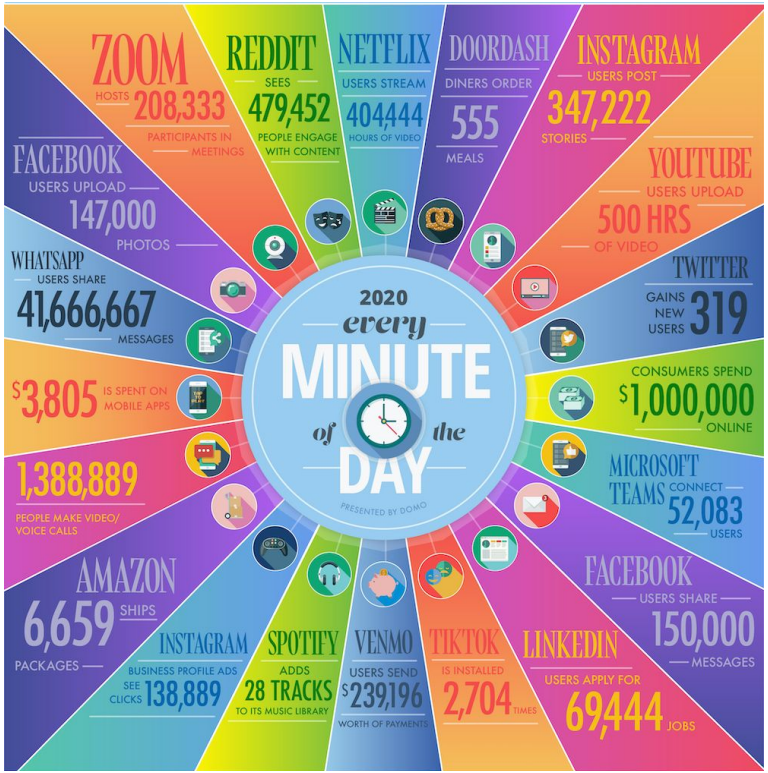
Contexte

De plus en plus de données

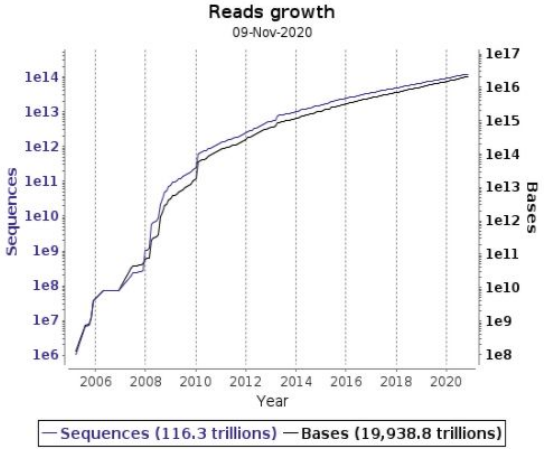
Data is the new oil
Clive Humby

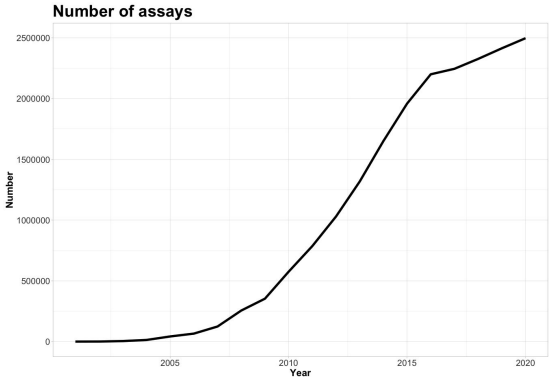
Data is the new oil? No: Data is the new soil.
David Mccandless

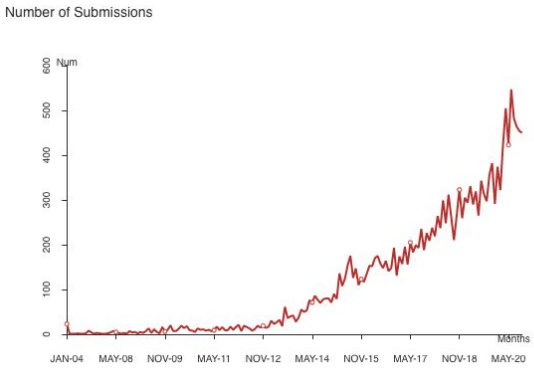
Quelques chiffres au quotidien



Data deluge en biologie







ENA
European Nucleotide Archive

Array express

PRIDE
PRoteomics IDEntifications Database

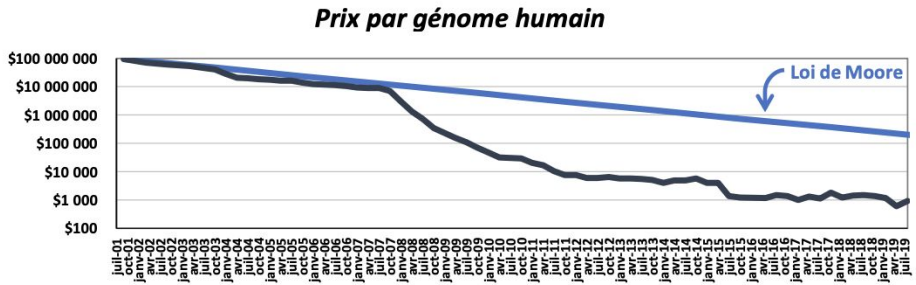
Data deluge en biologie

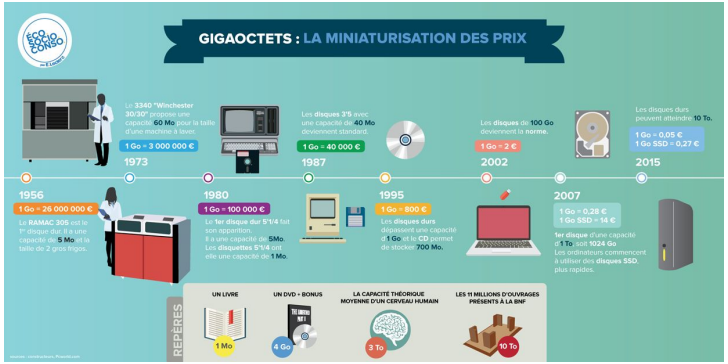
Type de données	Base de données	Volumes de données
Mesures de l'expression des gènes	ArrayExpress	74244 experiments 2520206 assays 60.39 TB of archived data
Mesures de l'expression des gènes	GEO	4 342 761 échantillons
Structure 3D des protéines	PDB	177009 Structures
Séquence nucléotidiques	GenBank	226 241 476 séquences et 776 291 211 106 bases
Données d'identification ou de quantification des protéines	Pride	543 205 140 spectres de masse
Séquence nucléotidiques (COVID-19)	GISAID (EpiCoV)	1 177 402 virus

MAJ : Avril 2021

Comment ?

Cout





Vitesse

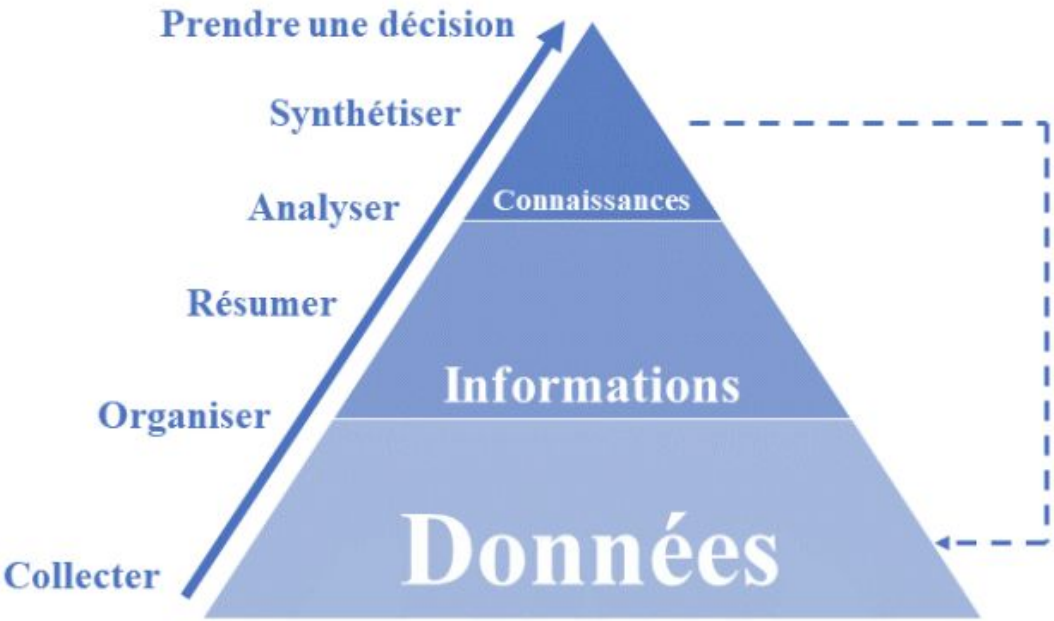
Dans les années 90

4 ans pour le premier milliard de nucléotides du génome humain

Aujourd'hui

Le génome complet en moins de 24h

Pourquoi ? Créer de la connaissance !



Pourquoi générer toujours plus de données ?

Pourquoi ne pas simplement pas 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.

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

Réplicabilité

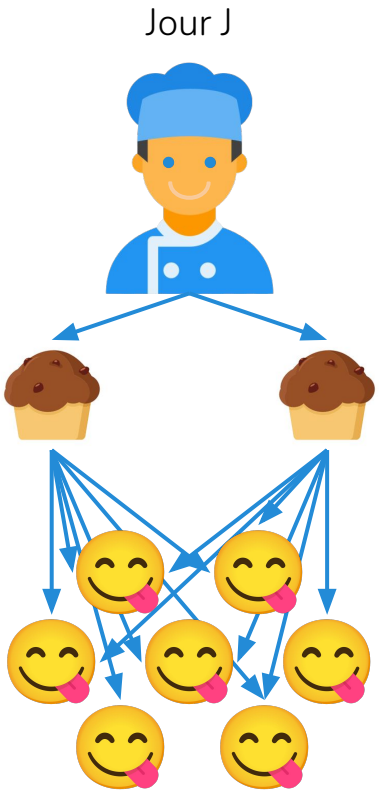
Répétabilité

Reproductibilité

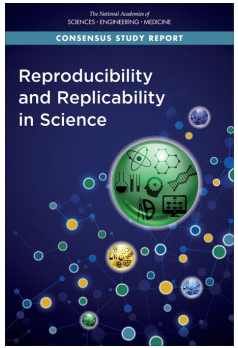
Souvent utilisées mais souvent confondues

et notamment par la langue (Plesser, 2018)

Réplicabilité



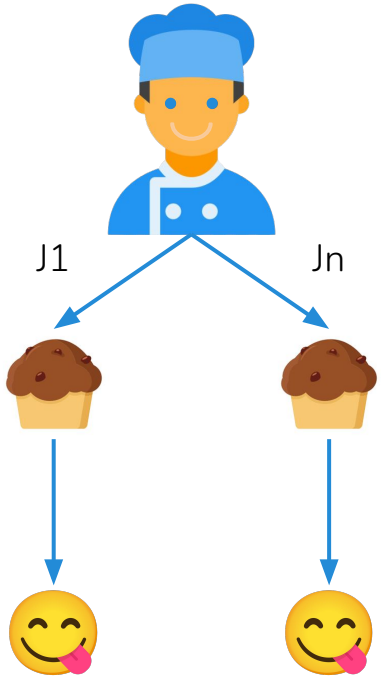
“L’étroitesse de l’accord entre les résultats individuels successifs obtenus sur le même échantillon soumis à l’essai dans le même laboratoire et dans les conditions suivantes : même analyste, même appareil, même jour ”



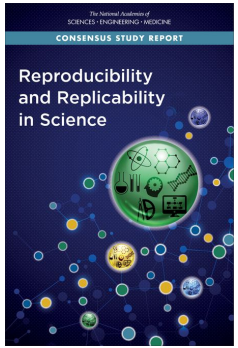
https://doi.org/10.17226/25303

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

Répétabilité



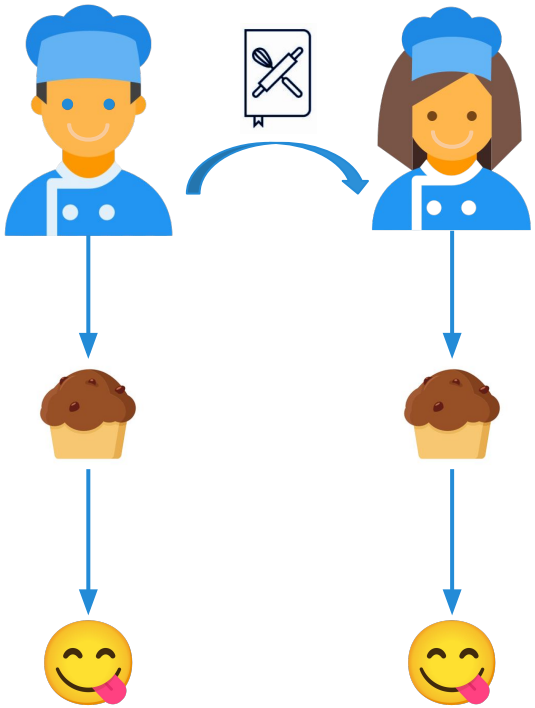
“L’étroitesse de l’accord entre les résultats individuels obtenus sur le même échantillon soumis à l’essai dans le même laboratoire et dont au moins l’un des éléments suivants est différent : l’analyste, l’appareil, le jour”



https://doi.org/10.17226/25303

“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



NORME INTERNATIONALE

ISO 5725-4

Exactitude (justesse et fidélité) des résultats et méthodes de mesure —

Partie 4: Méthodes de base pour la détermination de la justesse d'une méthode de mesure normalisée

“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”

Association for Computing Machinery

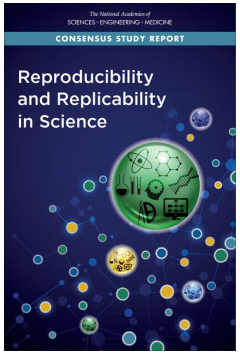
“Obtaining consistent computational results using the same input data, computational steps, methods, code, and conditions of analysis”

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

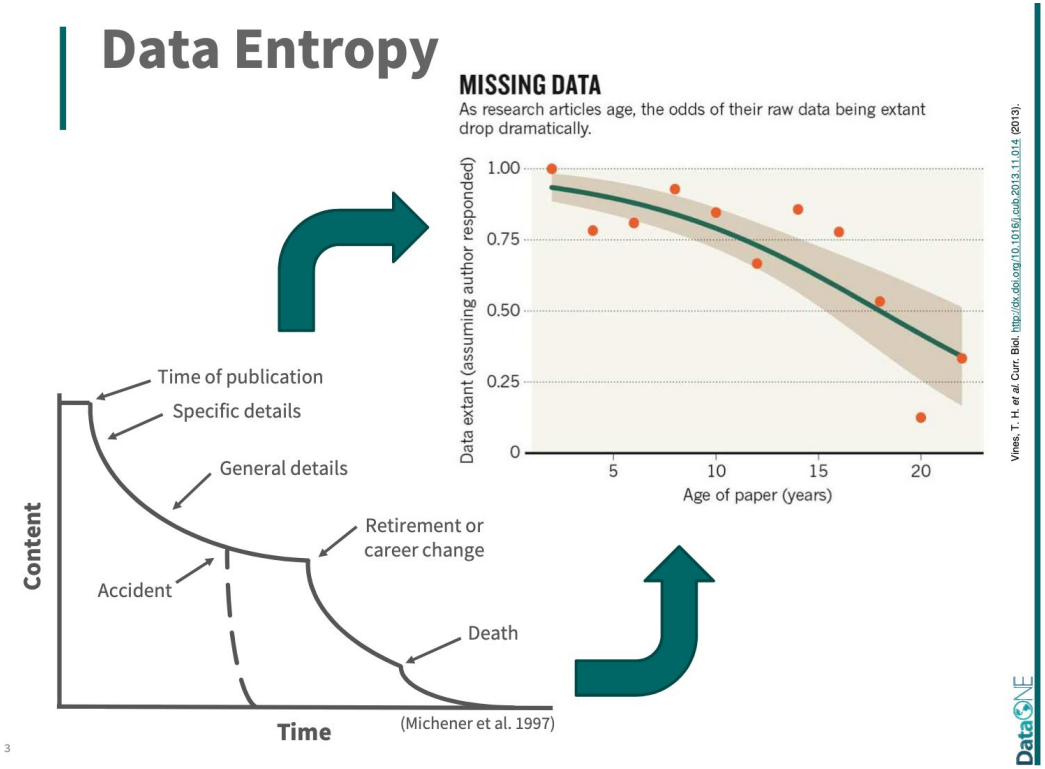
Description de la partie statistique

Décisions analytiques : quand, comment, pourquoi

Discussion des choix et des résultats obtenus

Et dans les faits ?

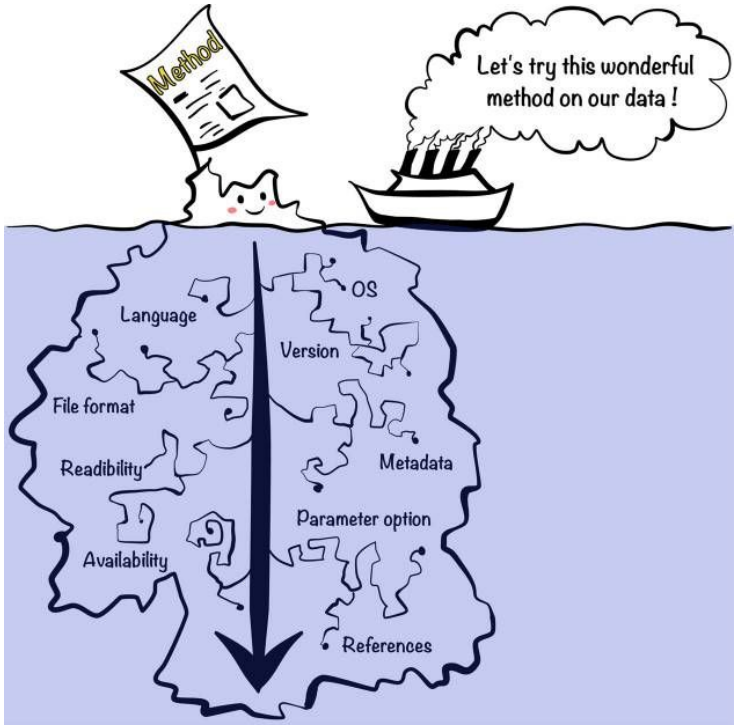
Les données face aux ravages du temps



Récupérer les données

https://youtu.be/66oNv_DJuPc

Reproduire les données



Kim et al, 2018

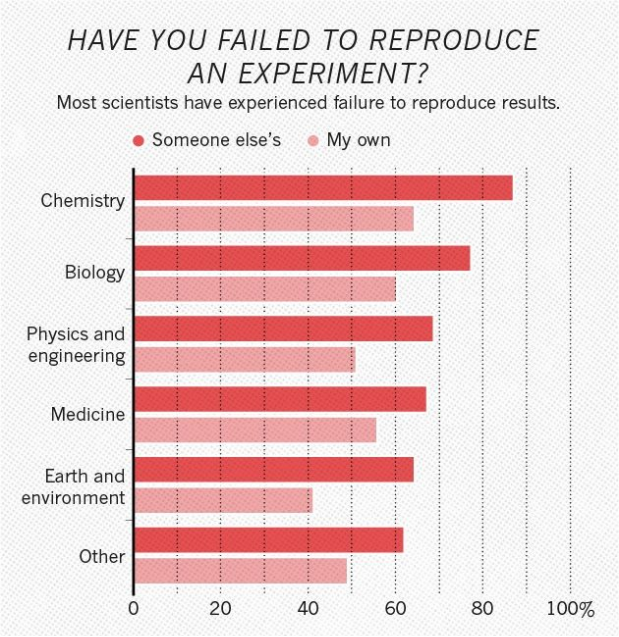
https://dx.doi.org/10.1093%2Fgigascience%2Fgiy077

En biologie

70 %

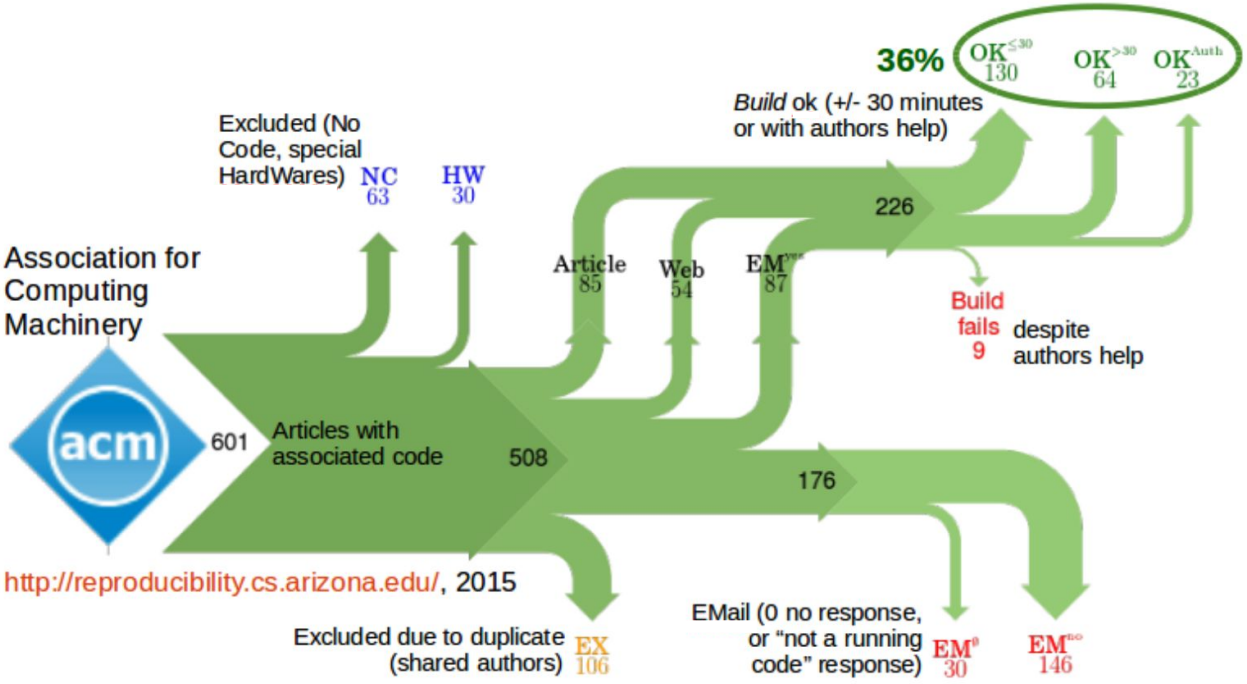
des analyses en biologie expérimentale ne sont

pas reproductibles



Monya Baker, 2016

En informatique



(Collberg et al. 2015)

La bioinformatique n'y échappe pas…

Impossibilité d'installer des outils

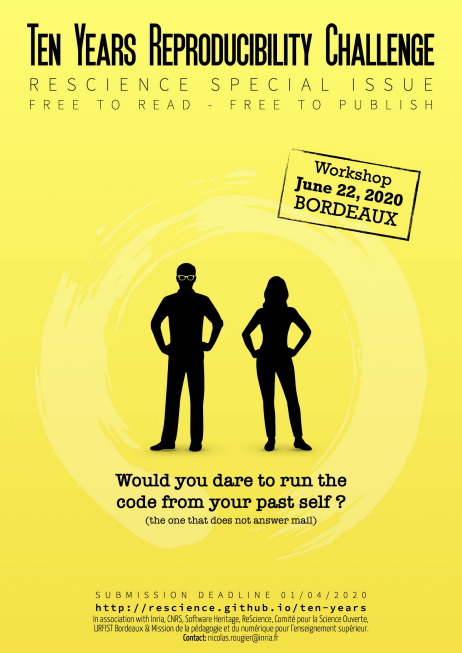
- OS non compatible
- Dépendance plus disponible / plus valide

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

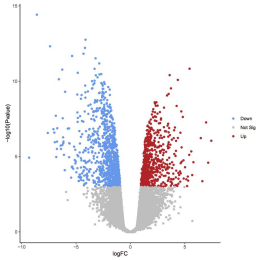


(Perkel, Nature, 2020)

Comment rendre un projet bioinformatique plus reproductible ?

Comment ?

En étant plus FAIR !



Données brutes

Principes FAIR data

&

Plan de gestion de données

Analyse de données

Codes

Algorithmes

Workflow

...

Communications

Articles

Thèse

Poster

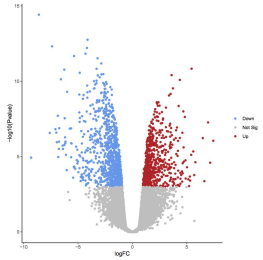
...

PARTIE I

Du point de vue des données

L'objectif ?

Être plus FAIR !



Données brutes

Principes FAIR data

&

Plan de gestion de données

Analyse de données

Codes

Algorithmes

Workflow

...

Communications

Articles

Thèse

Poster

...

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

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*** *(num* ***ific*** ***rese*** *to valid* *tion of th*

Scie *ural phen* ***ata****, whic* *s in phys* *ted through large-scale simulations; and reference data, which are highly curated datasets, such as the human genome.*

Données de recherche
=
Enregistrements factuels utilisés comme sources primaires pour la recherche scientifique et qui sont communément acceptés dans la communauté scientifique.

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**.

Qu'est ce qu'une métadonnée ?

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

- [illegible]
- [illegible] e

Il ex[illegible]

- [illegible]
- [illegible]

Métadonnées
=
Données qui décrivent des données

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**.

En résumé

Les données

Les métadonnées

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*) ;
- **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

Pourquoi discuter des données et des métadonnées ?

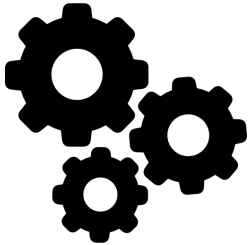
	Problématiques	Solutions
Les données	Organisation des données	Plan de gestion de données
	Partage des données	Question juridique Banque de données Data paper
Métadonnées	Qualité des métadonnées	Utilisation de standard Utilisation d'outil de data brokering

Des ressources sont disponibles à la fin sur ces problématiques

Mais au fait c'est quoi FAIR ?

Findable **A**ccessible **I**nteroperable **R**eusable






Sangya Pundir

Findable -- Faciliter la découverte des données

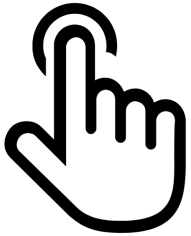
Findable

- 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

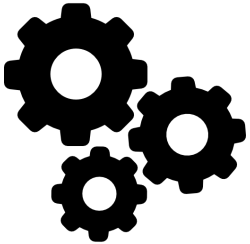


- Les données sont accessibles à travers un **protocole de communication standard**
- Ce protocole est **libre et ouvert**
- Ce protocole permet un accès par **authentification** si besoin
- Les **métadonnées restent accessibles** même si les données ne le sont plus

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

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

Interoperable

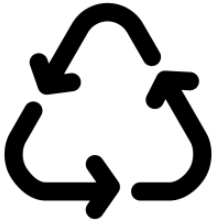


- Les données sont **décrites avec un vocabulaire contrôlé**
- Le vocabulaire utilisé **respecte les principes FAIR**
- Les **métadonnées sont reliées à d'autres données**

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

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

Reusable



- 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 savoir plus

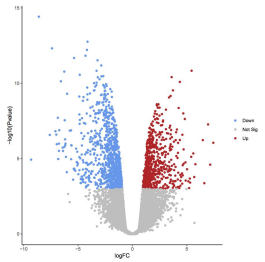
https://doranum.fr/enjeux-benefices/principes-fair/

PARTIE II

Du point de vue de l’analyse des données

L'objectif ?

Être plus FAIR !



Données brutes

Principes FAIR data

&

Plan de gestion de données

Analyse de données

Codes

Algorithmes

Workflow

...

Communications

Articles

Thèse

Poster

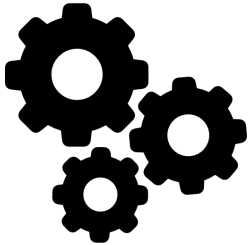
...

Des principes hérités des principes FAIR data

Findable

Accessible

Interoperable



Reusable

Des principes hérités des principes FAIR data

F

A

I

R

Des principes hérités des principes FAIR data

Facile à trouver

- Outils tiers utilisés = références dans leur domaine
- Protocole d'analyses simple à trouver (GitHub pages)

A

I

R

Des principes hérités des principes FAIR data

Facile à trouver

- Outils tiers utilisés = références dans leur domaine
- Protocole d'analyses simple à trouver (GitHub pages)

Accessible

- Ressources disponibles (GitHub, Dockerub)
- Outils tiers *open source* (conda)

I

R

Des principes hérités des principes FAIR data

Facile à trouver

- Outils tiers utilisés = références dans leur domaine
- Protocole d'analyses simple à trouver (GitHub pages)

Accessible

- Ressources disponibles (GitHub, Dockerub)
- Outils tiers *open source* (conda)

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

- Outils tiers utilisés = références dans leur domaine
- Protocole d'analyses simple à trouver (GitHub pages)

Accessible

- Ressources disponibles (GitHub, Dockerub)
- Outils tiers *open source* (conda)

Interopérable

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

Réutilisable

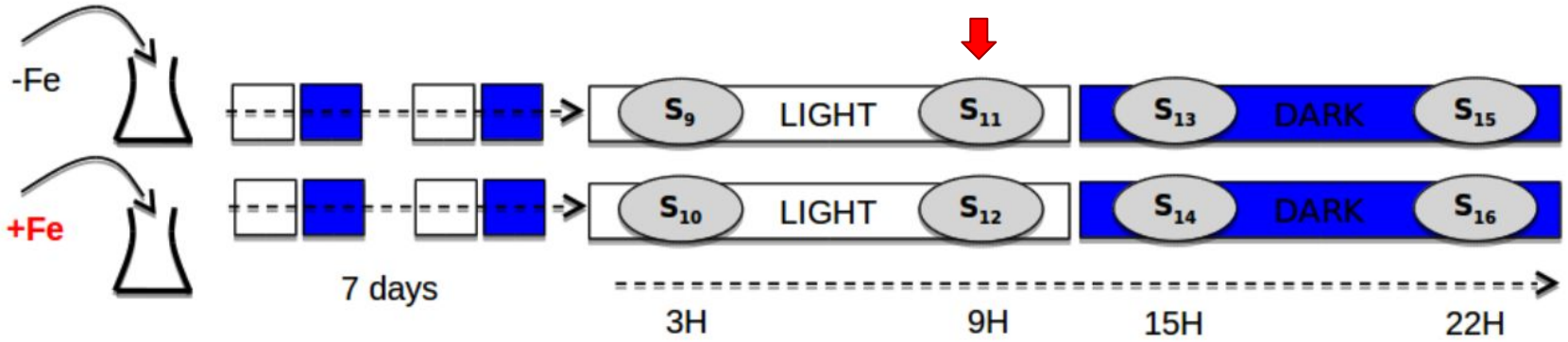
- Protocole rejouable simplement (snakemake) à l'identique (Jupyter) dans un environnement virtuel (docker)

Exemple d’utilisation

Plan expérimental

Étude de la réponse à une privation en fer chez l'algue verte *Ostreococcus tauri*

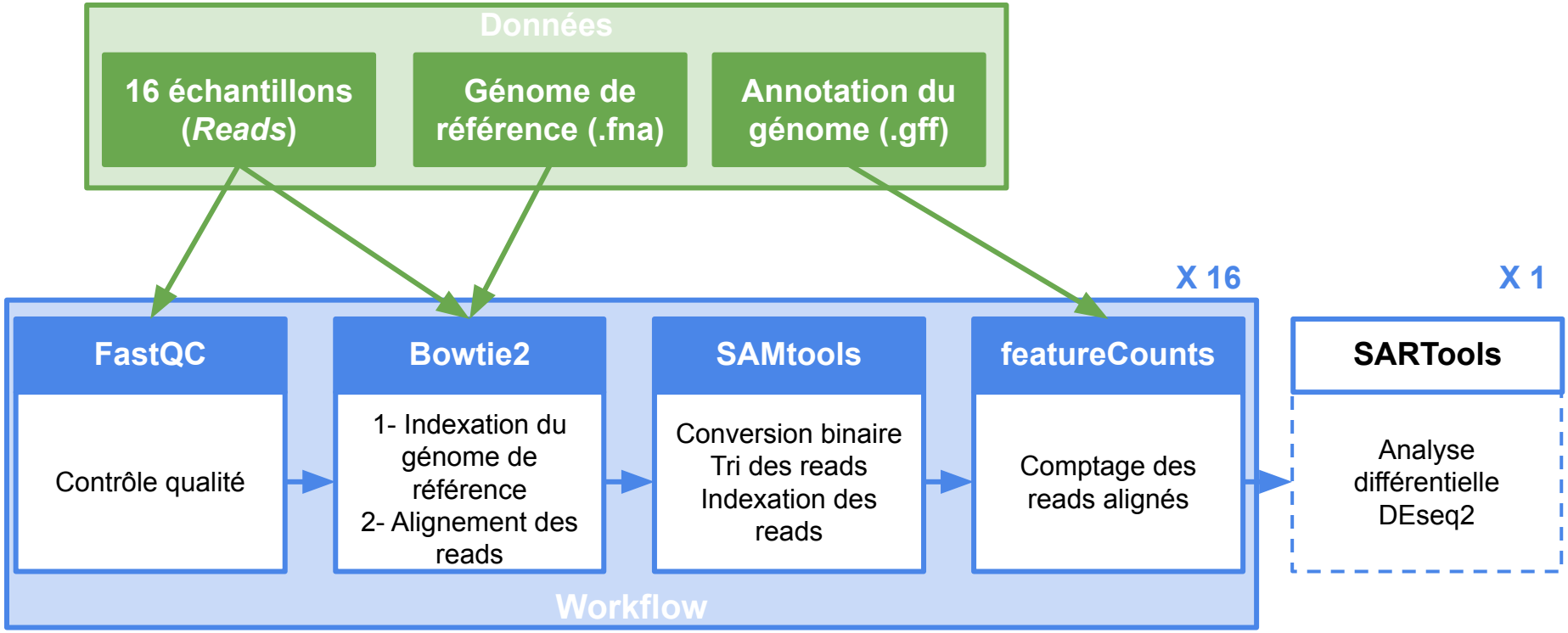
16 échantillons de données RNAseq (triplicat, single-end de 100bp)



(Lelandais *et al.* 2016)

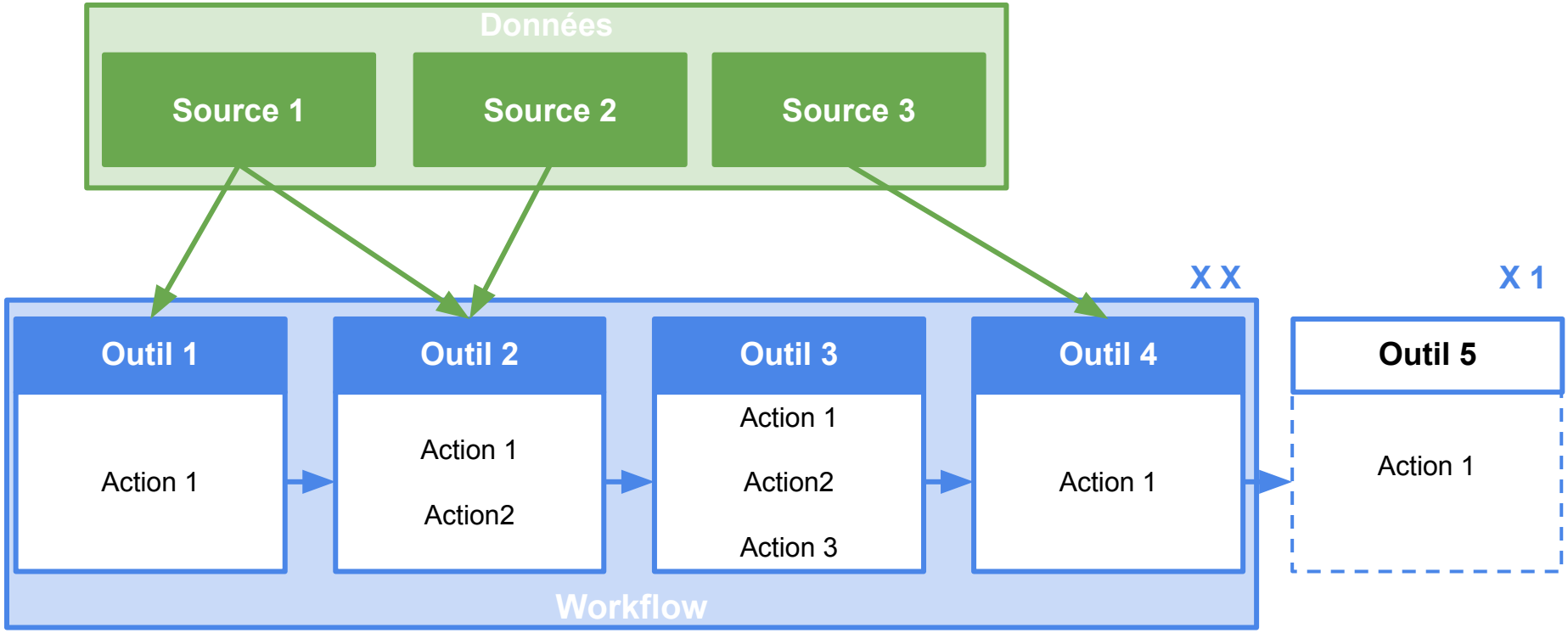
Données réduites pour la démonstration

Analyse de données RNAseq

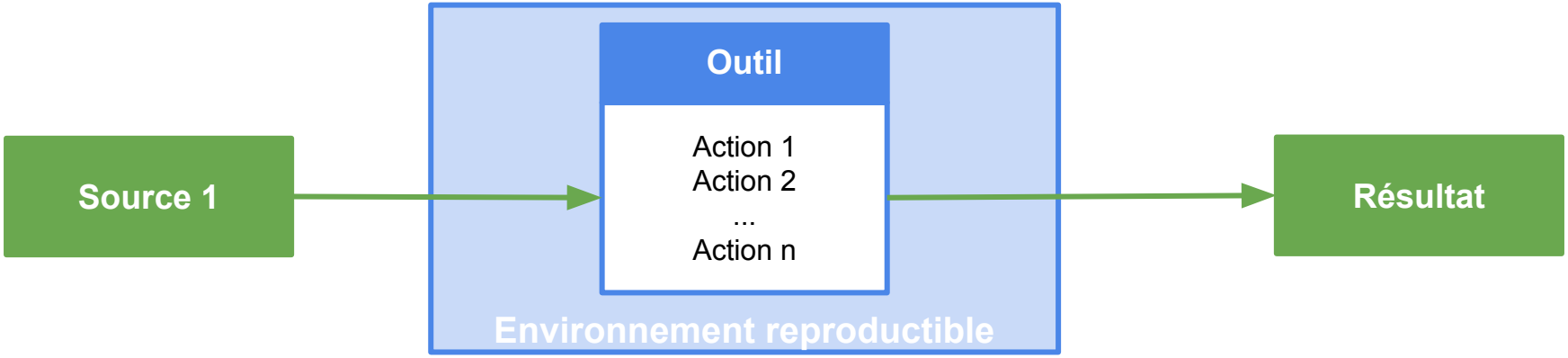


Exemple d'utilisation

Analyse par un pipeline



Développement d'un outil



Notre question à présent ?

Comment réaliser cette analyse de façon reproductible ?

(et pourquoi pas la rejouer en un click)

Exemple d'utilisation

Proposition en 7 étapes

Etape 7 - Editer un rapport d'analyses

Etape 6 - Portabilité de l'exploration des résultats

Etape 5 - Déporter l'analyse sur un serveur

Etape 4 - Création d'un script d'analyses

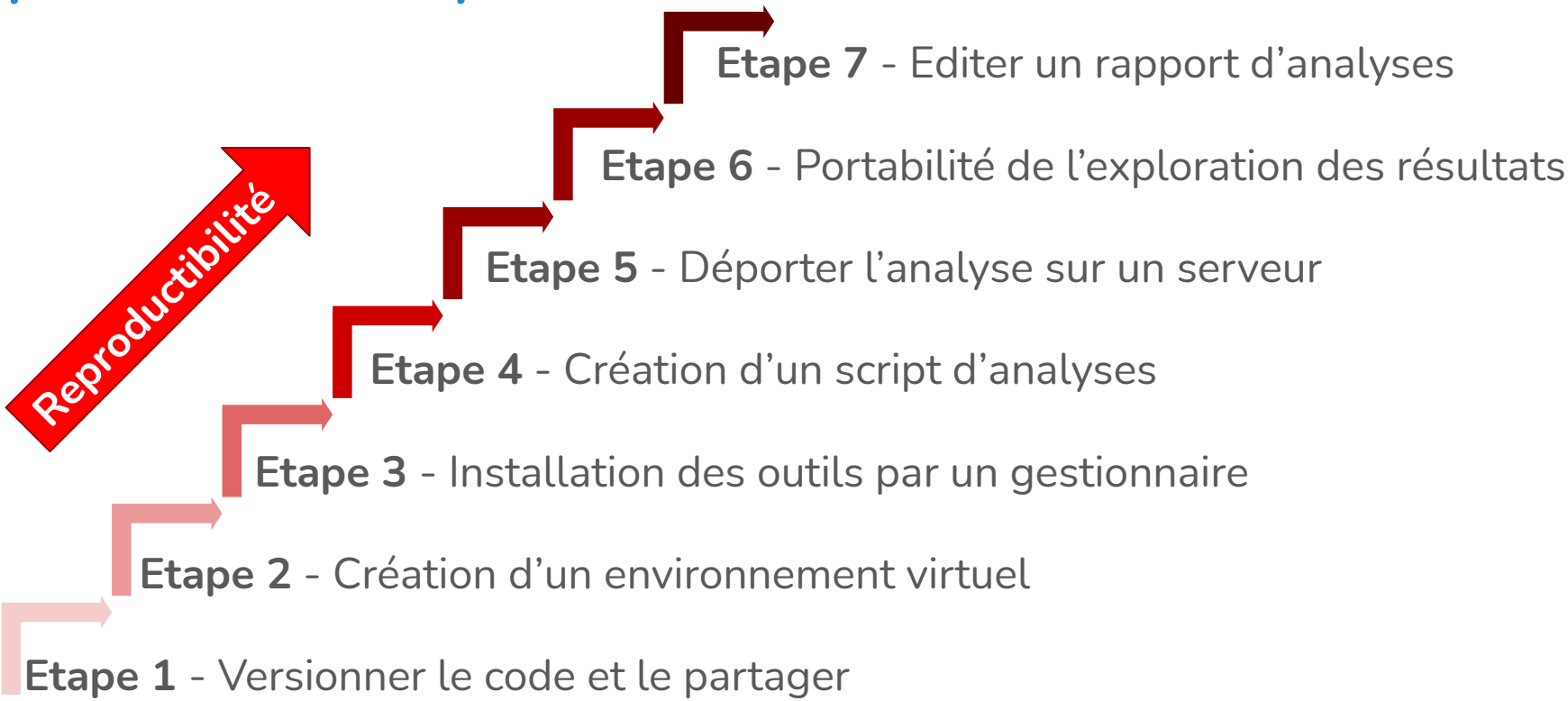
Etape 3 - Installation des outils par un gestionnaire

Etape 2 - Création d'un environnement virtuel

Etape 1 - Versionner le code et le partager

Exemple d'utilisation

Proposition en 7 étapes



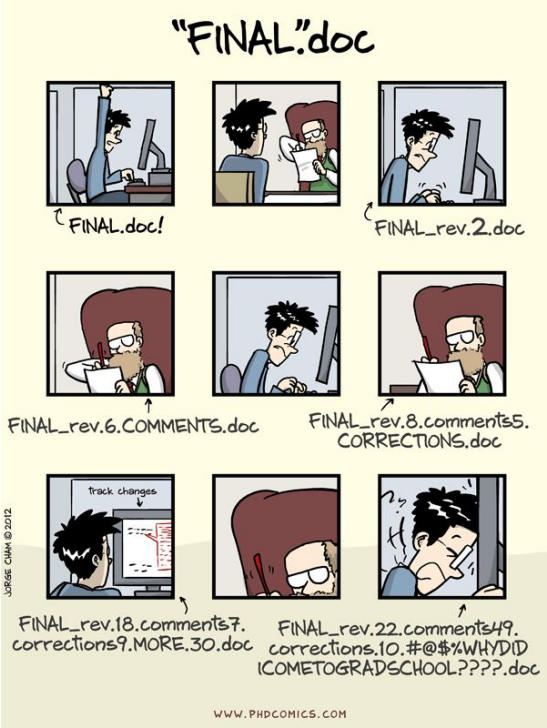
Des choix techniques équivalents

Etape 1 - Versionner le code et le partager

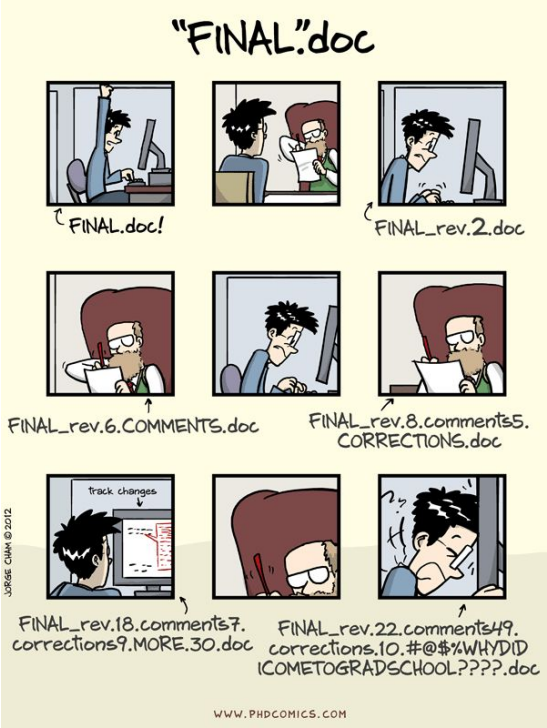
Pourquoi ?

- Avoir la bonne version du code
- Vision dans le temps
- Ouverture à la communauté

Etape 1 - Versionner le code et le partager



Etape 1 - Versionner le code et le partager



git

GitHub

GitLab

Avantages

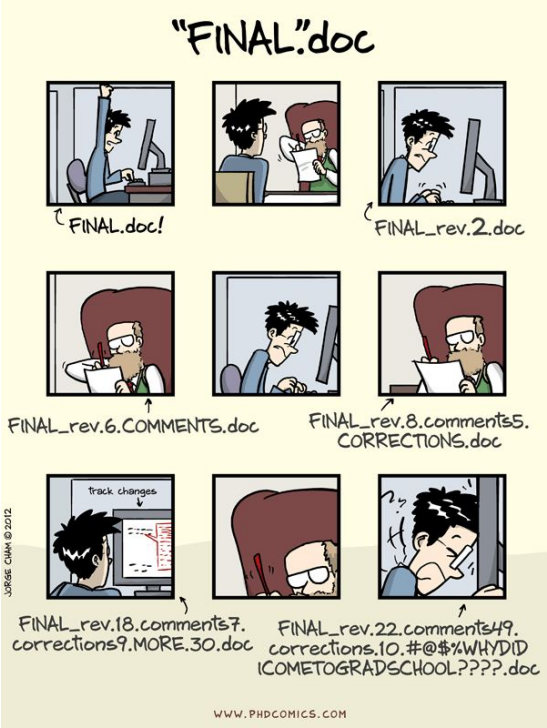
- Sauvegarde du code
- Simple pour partager
- Gestion automatique des versions

Inconvénients

- Pas simple pour les novices

Toujours plus reproductible !

Etape 1 - Versionner le code et le partager

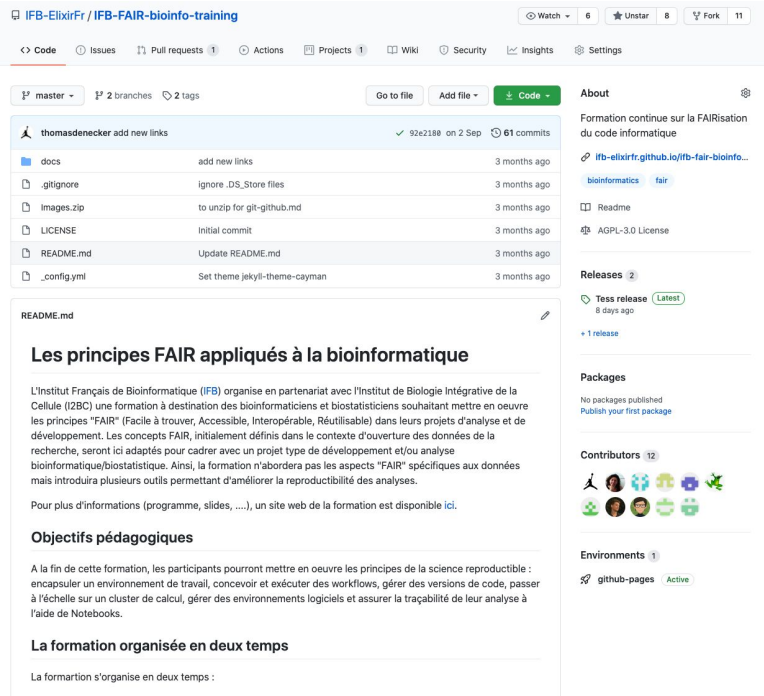


git

GitHub

GitLab

Après



Etape 2 - Création d'un environnement virtuel

Pourquoi ?

- Figer l'environnement
- Partager l'environnement

Etape 2 - Création d'un environnement virtuel

Avant

Etape 2 - Création d'un environnement virtuel

Avant

docker

Avantages

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

Inconvénients

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

Etape 2 - Création 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

CONDA

Après

```
$ conda install -c bioconda -y fastqc=0.11.2
```

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

Installation aussi simple

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 - Création d'un script d'analyse

Avant (script Shell)

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

SNAKEMAKE

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 - Création d'un script d'analyse

Avant (script Shell)

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

Plus court à écrire mais pas à exécuter !

Après (Snakefile)

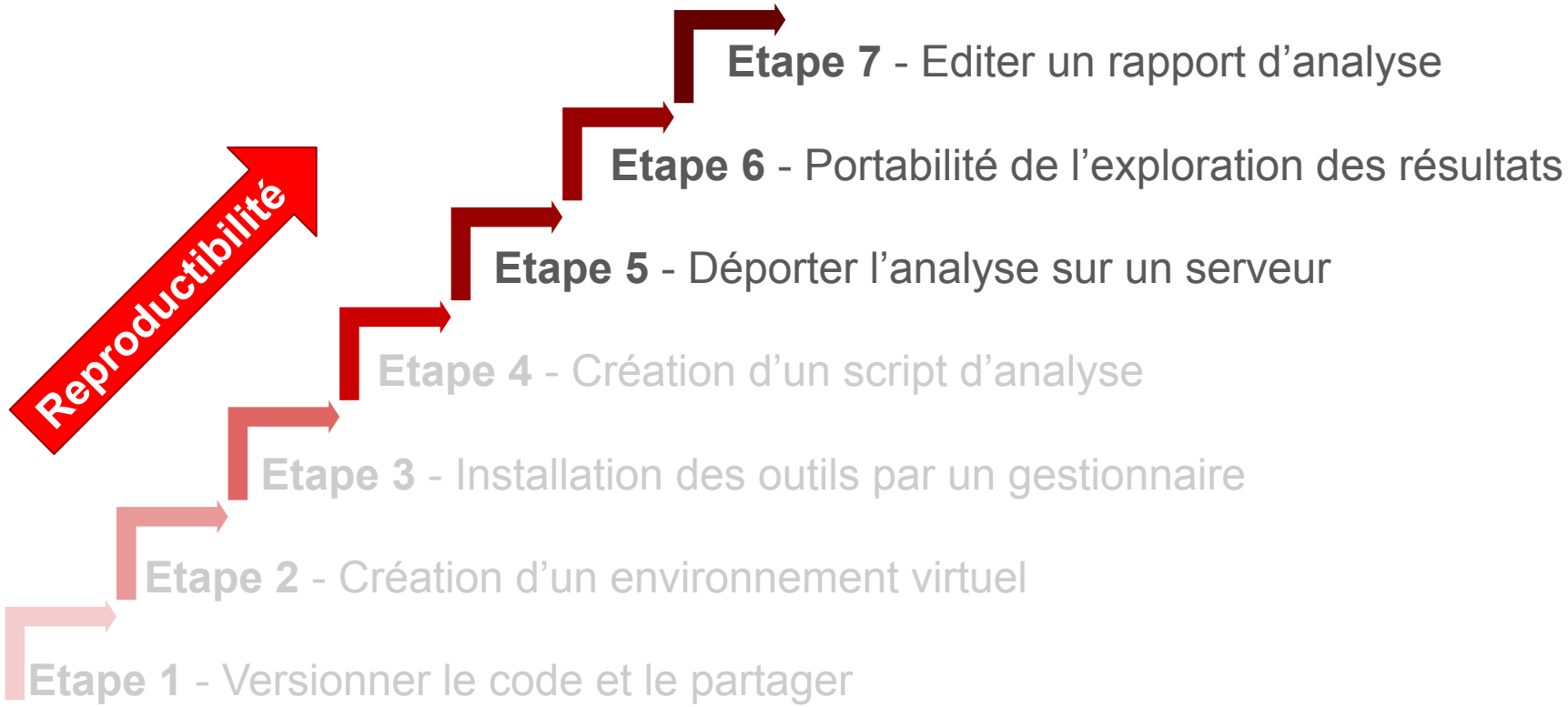
```
$ cat > Snakefile
SAMPLES, =
glob_wildcards("./samples/{smp}.fastq.gz")

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
```

Toujours plus reproductible !

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

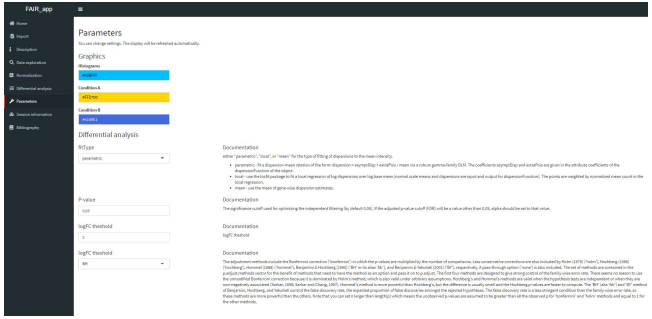
Etape 6 - Portabilité de l'exploration des résultats

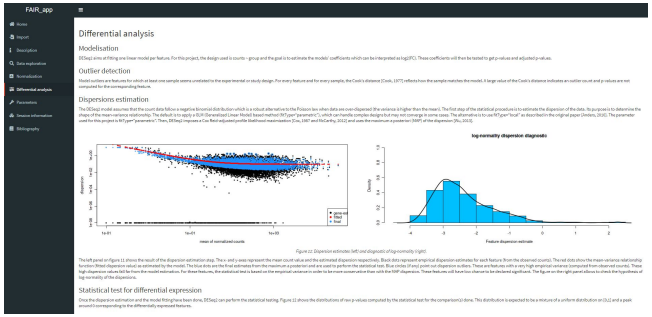
Avant : Terminal R

```
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")
```





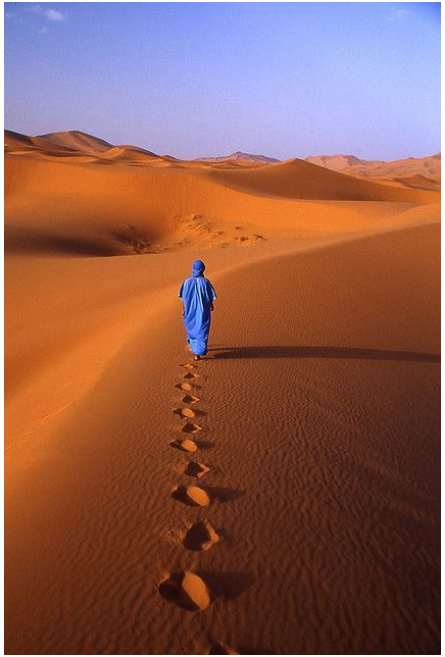
Etape 7 - Editer un rapport d'analyse

Pourquoi ?

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

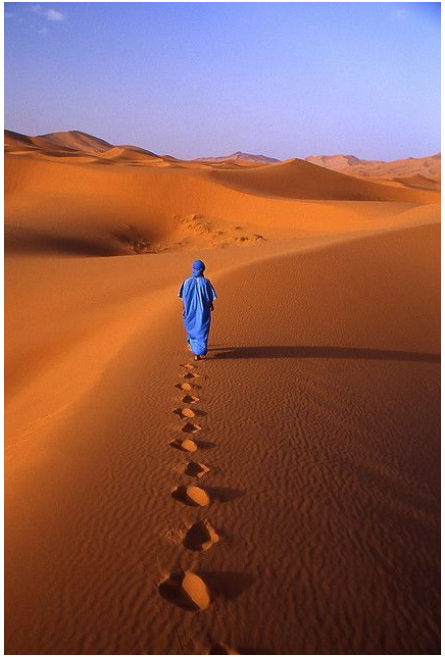
Etape 7 - Editer un rapport d'analyse

Avant



Etape 7 - Editer un rapport d'analyse

Avant



rmarkdown

jupyter

Avantages

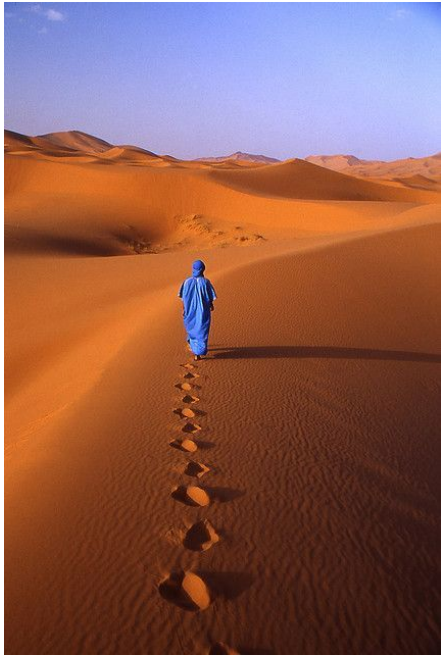
- Syntaxe simple (Markdown)
- Partage (PDF, HTML, ...)

Inconvénients

- Rares problèmes de visualisation en 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 developped at PF2 - Institut Pasteur by M.-A. Dillies and H. Varet (hugo.varet@pasteur.fr). Thanks to cite H. Varet, L. Brillet-Guéguen, 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 analyses 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 PF2 - 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	SRR3099587_chr18_ftc.txt	STANDARD
S2	SRR3099586_chr18_ftc.txt	STANDARD
S3	SRR3099585_chr18_ftc.txt	STANDARD
D1	SRR3105699_chr18_ftc.txt	DEPLETED
D2	SRR3105698_chr18_ftc.txt	DEPLETED
D3	SRR3105697_chr18_ftc.txt	DEPLETED

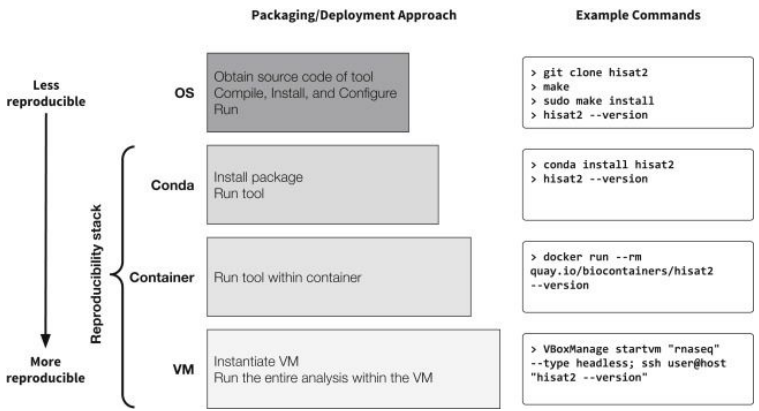
After loading the data we first have a look at the raw data table itself. The data table contains one row per annotated feature and one column per sequenced sample. Row names of this table are feature IDs (unique identifiers). The table contains raw count values representing the number of reads that 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
ostta01g00010	11	12	11	7	5	1

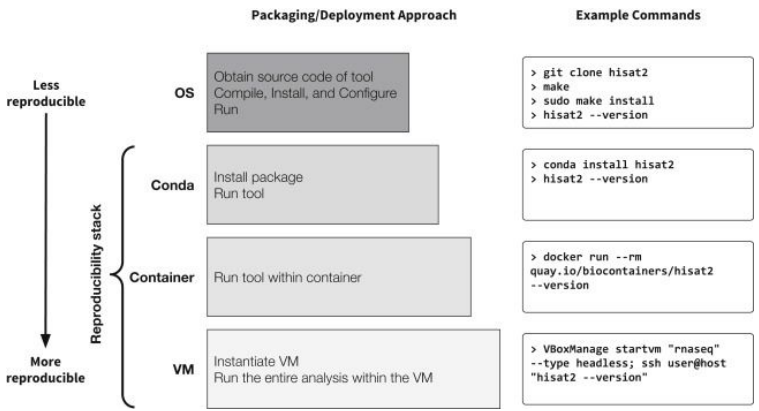
Conclusion

Quel est notre niveau de reproductibilité?



Practical Computational Reproducibility in the Life Sciences, Björn Grüning *et al*, 2018

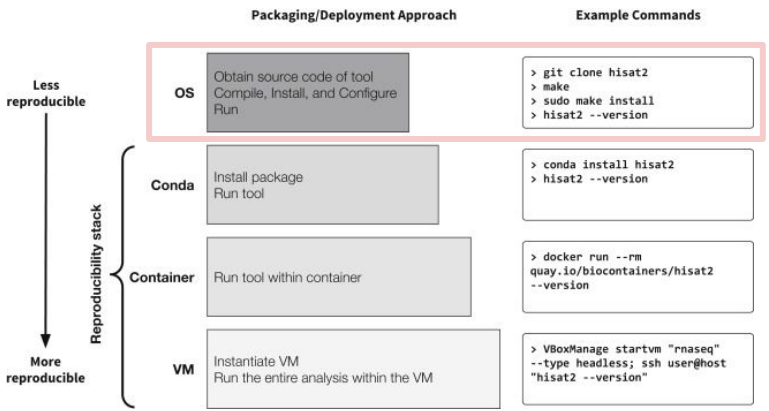
Quel est notre niveau de reproductibilité?



Practical Computational Reproducibility in the Life Sciences, Björn Grüning *et al*, 2018

FAIR bioinfo

Quel est notre niveau de reproductibilité?

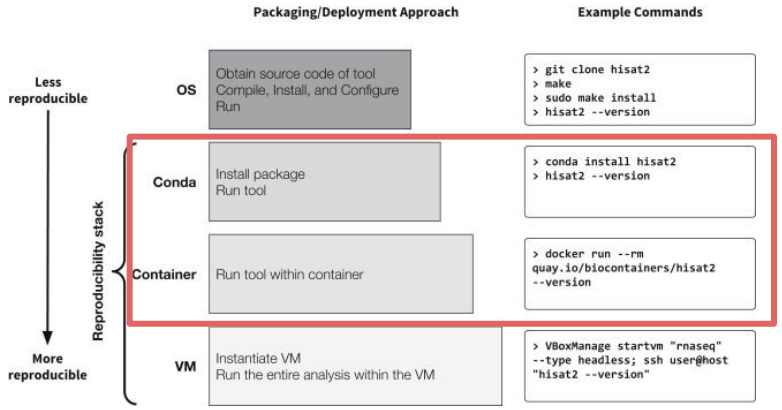


Practical Computational Reproducibility in the Life Sciences, Björn Grüning *et al*, 2018

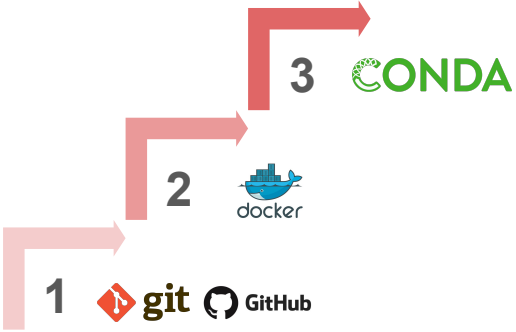
1 git GitHub

FAIR bioinfo

Quel est notre niveau de reproductibilité?

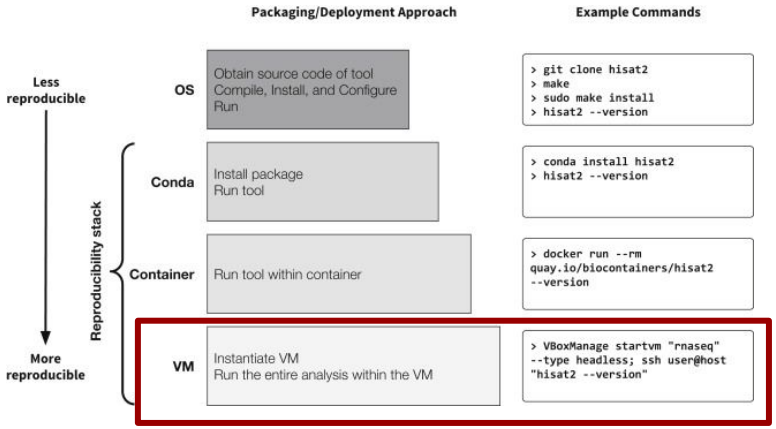


Practical Computational Reproducibility in the Life Sciences, Björn Grüning *et al*, 2018

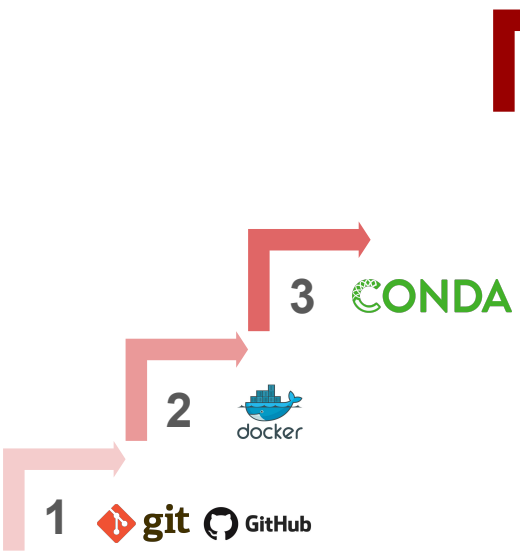


FAIR bioinfo

Quel est notre niveau de reproductibilité?

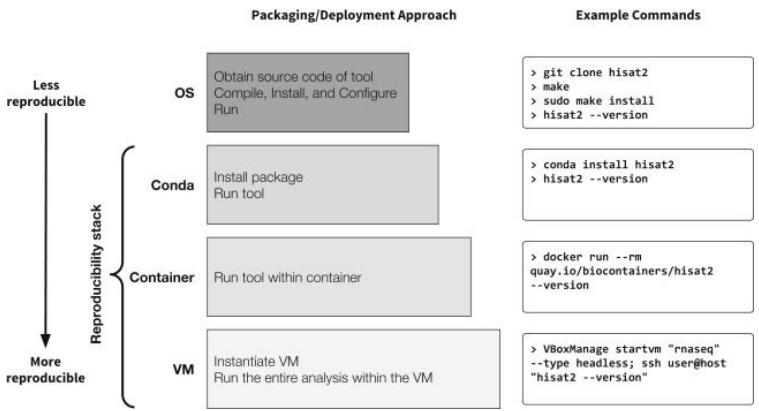


Practical Computational Reproducibility in the Life Sciences, Björn Grüning *et al*, 2018

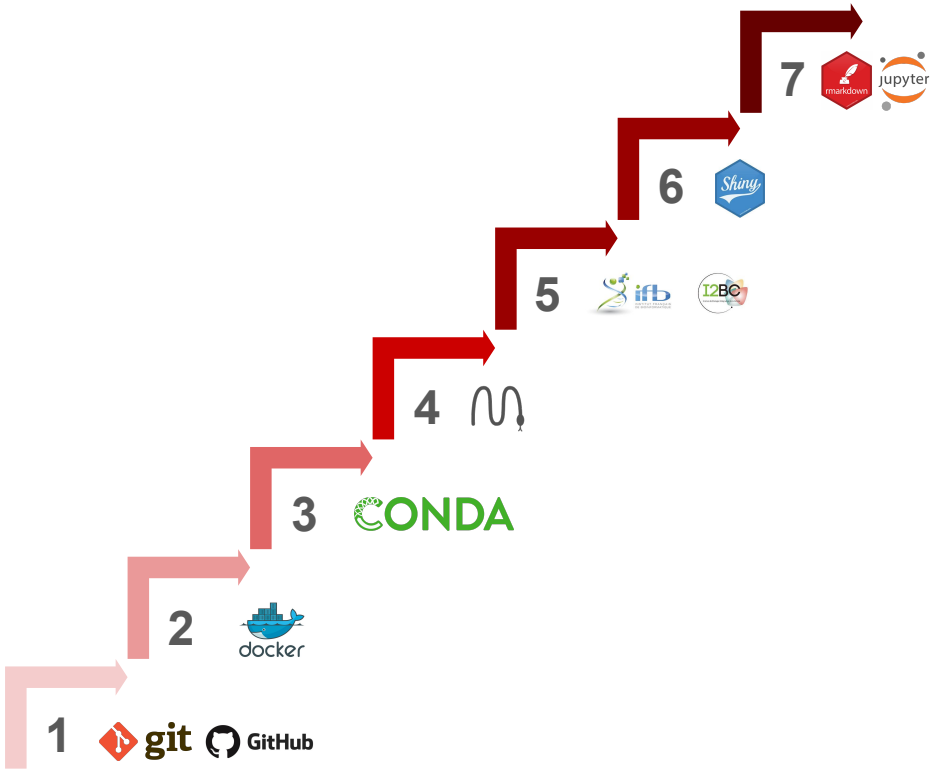


FAIR bioinfo

Quel est notre niveau de reproductibilité?



Practical Computational Reproducibility in the Life Sciences, Björn Grüning *et al*, 2018



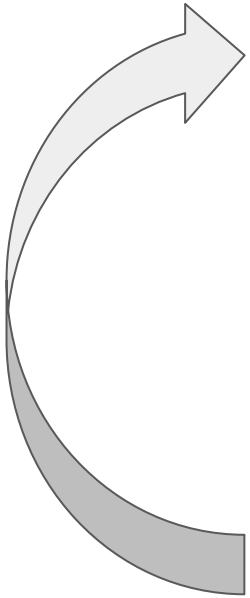
Take home messages

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

Formations

Vous souhaitez

- Savoir comment remplir un plan de gestion de données ?
- Comprendre la différence entre PGD Structure et PGD Projet ?
- En savoir plus sur les métadonnées et ses standards ?
- Le cadre juridique des données ?

L'IFB propose une formation !

(le contenu des éditions précédentes est en ligne)

https://ifb-elixirfr.github.io/IFB-FAIR-data-training/index.html

Merci

Une belle équipe !

H. Chiapello

T. Denecker

J-F Dufayard

F. de Lamotte

P. Lieby

Y. Mahmah

G Sarah

J. Seiler

Mise en application

Notre objectif

(Re)Découvrir des outils complémentaires pour gagner en reproductibilité

Notre crédo

FAIR raw data + FAIR bioinfo = FAIR processed data

Notre méthodologie

Rendre une analyse de données reproductible à partir de données publiées

https://ifb-elixirfr.github.io/IFB-FAIR-bioinfo-training/index.html#home

Conclusion

Une belle équipe !

2019

T. Denecker

C. Toffano-Nioche

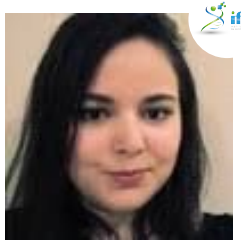
2020-2021

C. Hernandez

H. Chiapello

G. Le Corguillé

P. Lieby



Y. Mahmah

J. Seiler

J. van Helden

Ressources

En savoir plus

scientific data

The FAIR Guiding Principles for scientific data management and stewardship

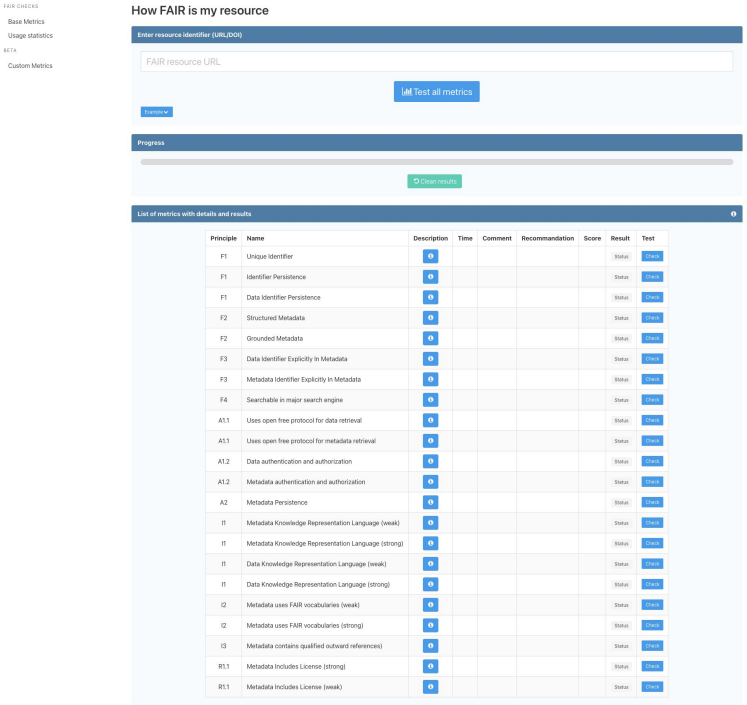
Wilkinson et al., 2016

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

GO FAIR

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

Exemple d'évaluation automatique par FAIR CHECKER



https://fair-checker.france-bioinformatique.fr/base_metrics

Thomas Rosnet

Ressources IFB

FAIR_bioinfo

- 2019 : https://github.com/thomasdenecker/FAIR_Bioinfo
- 2020 : https://ifb-elixirfr.github.io/IFB-FAIR-bioinfo-training/session2020.html
- 2021 : https://ifb-elixirfr.github.io/IFB-FAIR-bioinfo-training/session2021.html

FAIR & le cluster de l'IFB

- Slurm : 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
- Docker/Singularity : 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

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)

Les outils

Il existe plusieurs outils dont

DMP OPIDoR
Solution nationale

DSW - Data Stewardship Wizard
Solution européenne (ELIXIR)

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, …

Exemple d'entrepôts de données

Thématique

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

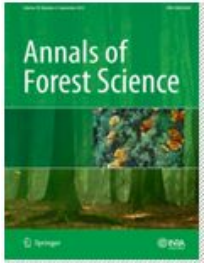
Toujours chercher à valoriser les données

Publier un **datapaper**

Publier un **article de recherche**

Rédiger une brève pour un **magazine** spécialisé

Contribuer à un **blog**, ...



Journal of open health data

BMC Research Notes

A peer-reviewed open-access journal
Biodiversity Data Journal
Making your data count! ISSN 1314-2828 (online)

Open Data Journal for Agricultural Research

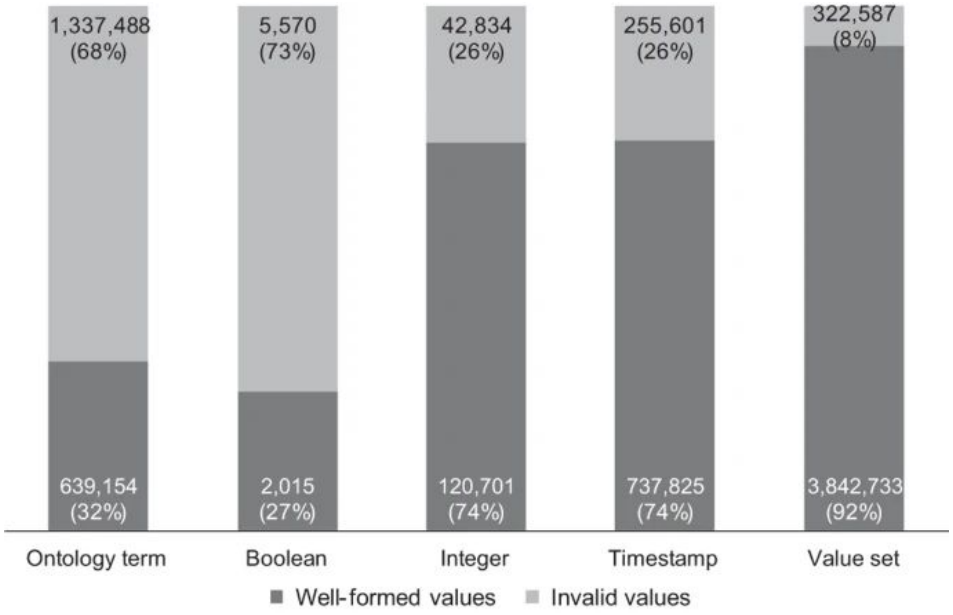
Comment bien décrire les données ?

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

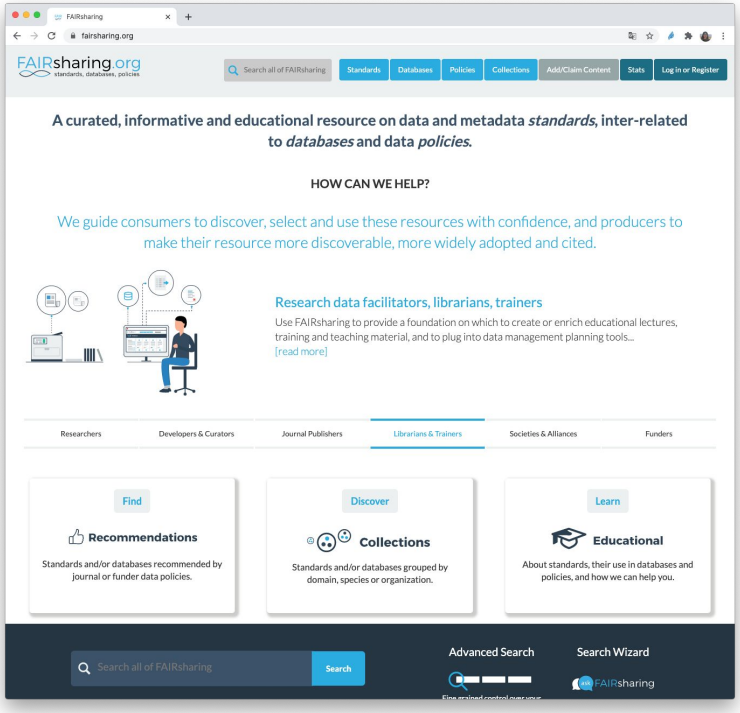
A standard provides the **requirements, specifications, guidelines or characteristics** that can be used for the **description, interoperability, citation, sharing, publication, or preservation** of all kinds of **digital objects** such as data, code, algorithms, workflows, software, or papers.

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



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

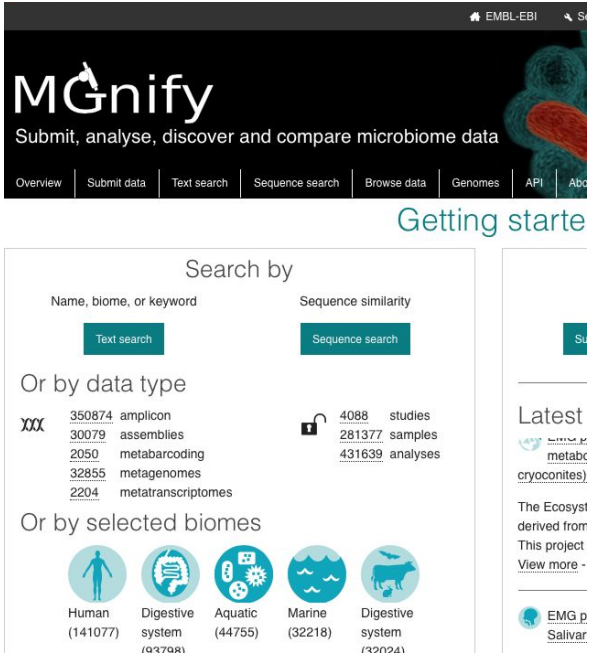
Specification projects	MIGS	MIMS	MIMARKS	New checklists
Checklists	EU, BA, PL, VI, ORG	metagenomes	survey, specimen	e.g., pan-genomes
Shared descriptors	collection date, environmental package, environment (biome), environment (feature), environment (material), geographic location (country and/or sea, 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, Host-associated, Human-associated, Human-oral, Human-gut, Human-skin, Human-vaginal, Microbial mat/biofilm, Miscellaneous natural or artificial environment, Plant-associated, Sediment, Soil, Wastewater/sludge, Water			

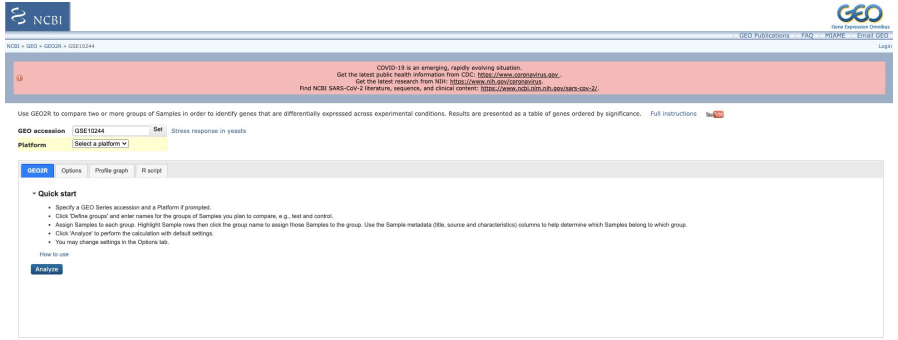
Yilmaz et *al*, 2011

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 MGinfy, GEOtoR, …

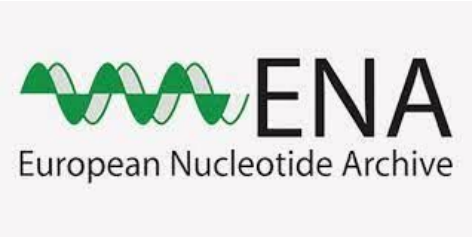




Soumission très hétérogène

Par simple fichier Excel

Un peu plus complexe



MAIS

avec une qualité des métadonnées très 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

gfbio GERMAN FEDERATION FOR BIOLOGICAL DATA

NIH NIAID METAGENOTE

omicsBroker
Annotate and publish easier

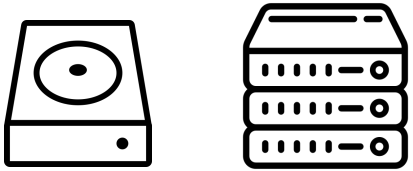
Bonnes pratiques

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

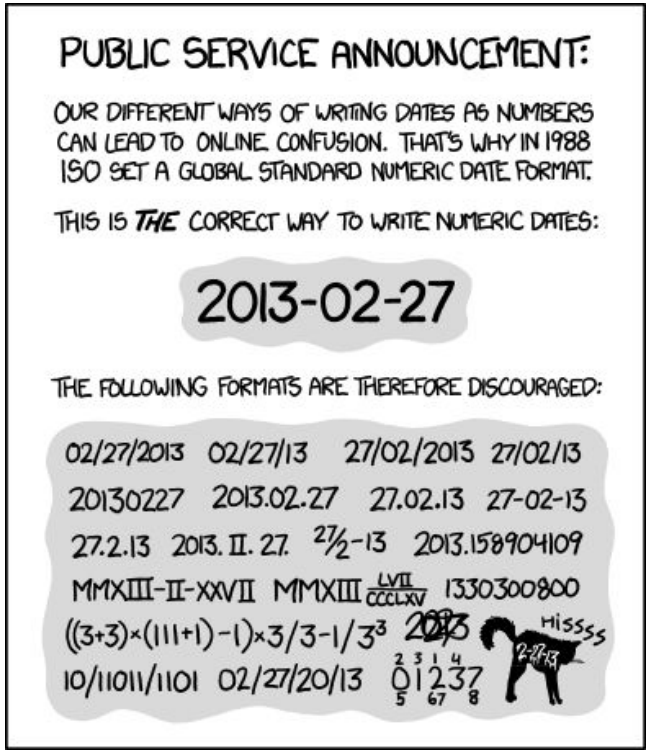
- Bref et explicite
- Sans espace ni caractères spéciaux
- Avec une date au bon format
- Avec l'élément le plus important en premier
- Avec la version du document

Format

Si possible non propriétaire

Les formats qui perdent le moins de données à la conversion

Le format utilisé par la communauté



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)