



CESAB  
CENTRE DE SYNTHÈSE ET D'ANALYSE  
SUR LA BIODIVERSITÉ



# Research Compendium

*ou comment bien ranger sa chambre*



François Guilhaumon

*{{ Chercheur IRD }}*

Mardi 3 Novembre 2020

# Research Compendium

# Research Compenkoi ?

# Research Compendium

"A collection of concise but detailed information about a particular subject"

Un abrégé, un précis, un recueil. Il synthétise de manière exhaustive votre projet d'analyse de données.

# Research Compendium

- Quelques règles simples pour organiser son répertoire de travail
- Et pouvoir s'y retrouver (ou que les autres s'y retrouvent)
- Un jour vous allez partager votre code, vos données, vos résultats. Avec vous-même, avec vos étudiants, votre encadrant, vos collègues ou le reste du monde.
- Autant se préparer depuis le début plutôt que de devoir tout refaire ou risquer la honte internationale !

# Une définition formelle

"The goal of a research compendium is to provide a standard and easily recognisable way for organising the digital materials of a project to enable others to inspect, reproduce, and extend the research."

# Principes

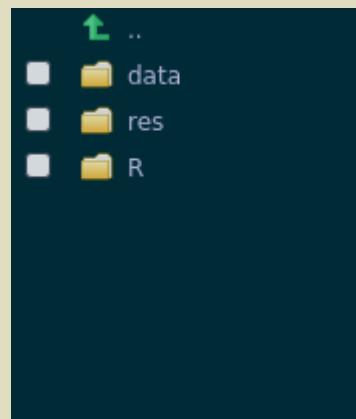
- Should organize its files according to the [...] conventions of (your) community (discipline or a lab). Following these conventions will help other people recognize the structure of the project, and also support tool building which takes advantage of the shared structure.
- Respectez les conventions de votre domaine si elles existent !

# Principles

- Should maintain a clear separation of data, method, and output.

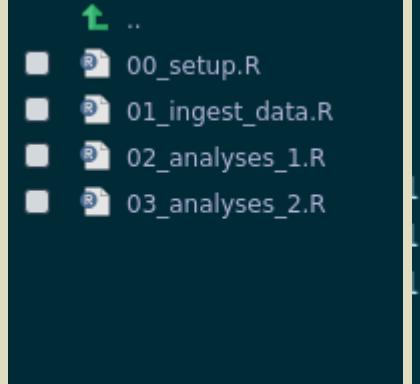
This means data files must be separate from code files. Keeping data and method separate treats the data as “read-only,” so that the original data is untouched and all modifications are transparently documented in the code.

The output files should be considered as disposable, with a mindset that one can always easily regenerate the output using the code and data.



# Principes

- Le flux d'analyses doit être séparé en étapes courtes (ou pas trop longues) et homogènes
- Numérotez les scripts, placez-y des entêtes



```
1 #####
2 #
3 # My Compendium
4 #
5 # 00_setup.R
6 #
7 # libraries and global variables
8 #
9 # francois.guilhaumon@ird.fr
10 #####
11
12
```

# Principes

- The relationship between which code operates on which data in which order to produce which outputs must be specified as well.

Utilisez un script "maître" (**make.R**) qui execute les étapes dans l'ordre (c'est le SEUL script R à la racine !)

```
#####
# My Compendium
#
# make.R
#
# francois.guilhaumon@ird.fr
#####

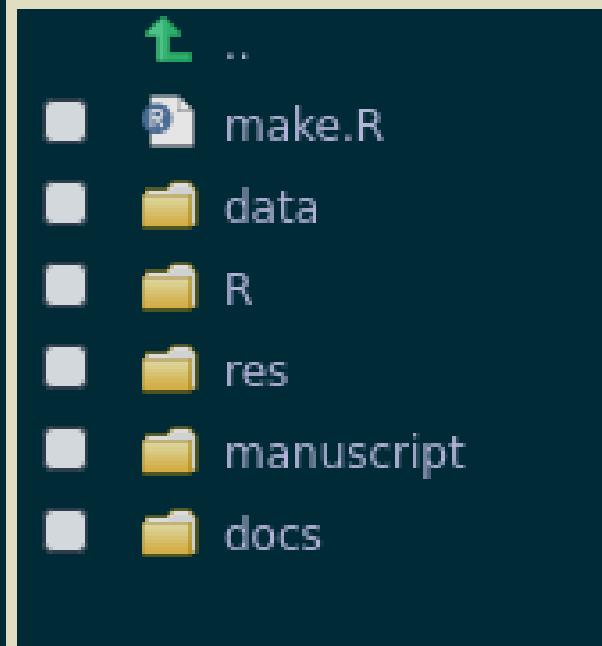
# clean workspace
rm(list = ls())

# 00_setup.R
source("R/00_setup.R")

# 01_ingest_data.R
source("R/01_ingest_data.R")

# 02_analyses_1.R
source("R/02_analyses_1.R")

# 03_analyses_2.R
source("R/03_analyses_2|R")
```



# Principes

Utilisez un script "maître" (`make.R`) qui execute les étapes dans l'ordre (c'est le SEUL script R à la racine !)

Tous les scripts, "tournent" tous à partir de la racine du projet !

Tous les chemins (relatifs) seront définis à partir de la racine du projet !

`res/1_datacleaned.csv` , `data/data_raw.csv`

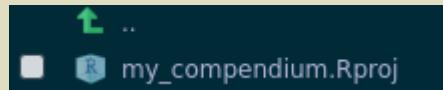
Cela peut poser des problèmes lors de la compilation des .Rmd situés à l'intérieur de votre arborescence. --> Utilisez le package `{here}`



# Principes

- L'utilisation d'un projet Rstudio permet de se retrouver à la racine du projet à son ouverture.

SI VOUS UTILISEZ LA FONCTION SETWD() JE BRÛLE VOTRE ORDINATEUR !



# Principes

- The relationship between which code operates on which data in which order to produce which outputs must be specified as well.

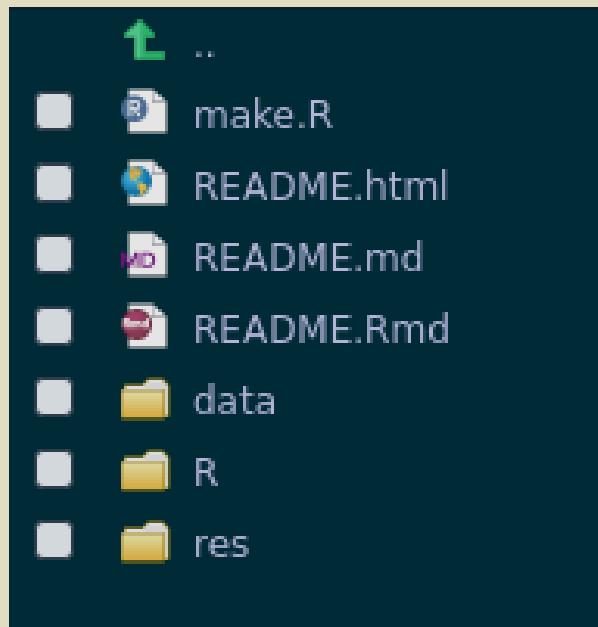
Chaque script écrit des résultats (fichiers) les référençant explicitement dans leur nom.

e.g. le script "R/01\_ingest\_data.R" écrit des résultats du type  
"res/01\_Zzz9Zzz9.truc"

# Principes

- Should specify the computational environment that was used for the original analysis. At its most basic, this could be a plain text file that includes a short list of the names and version numbers of the software and other critical tools used for the analysis. In more complex approaches, described below, the computational environment can be automatically preserved or reproduced as well.

Placez un README à la racine du projet



# Principles

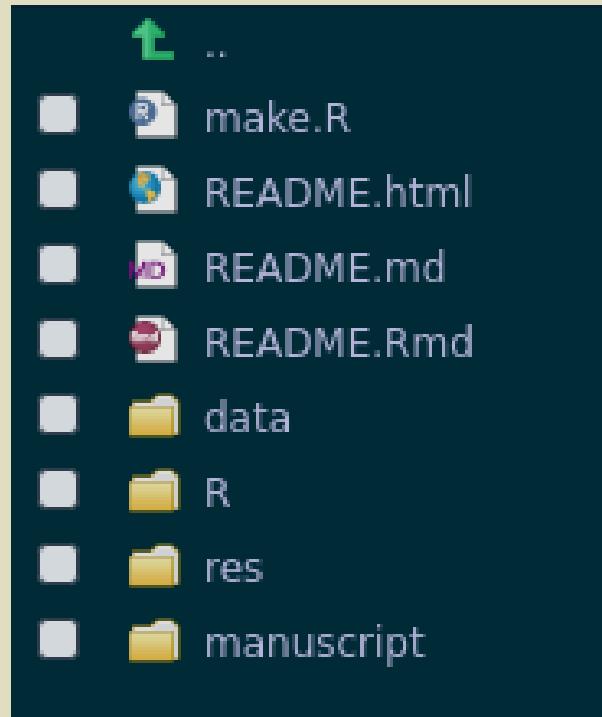
- Should specify the computational environment that was used for the original analysis. At its most basic, this could be a plain text file that includes a short list of the names and version numbers of the software and other critical tools used for the analysis. In more complex approaches, described below, the computational environment can be automatically preserved or reproduced as well.

Utilisez docker !



# Principes

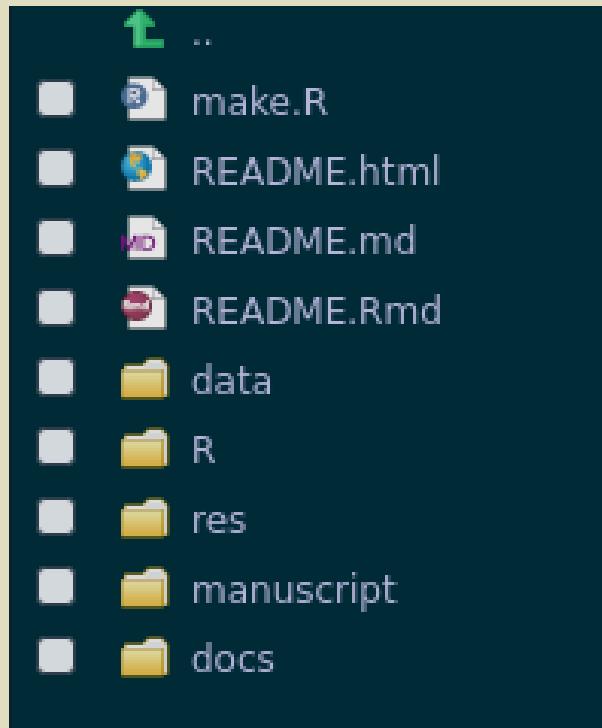
- Séparons aussi le(s) documents de synthèse (papier, présentation, ...)



Utilisez les packages rmarkdown, knitr, et rticles, ils sont puissants.

# Principes

- Séparons aussi le(s) ressources utiles (biblio, etc ...)



# Diffusion / stockage

- Il existe de nombreuses plateformes pour stocker votre supp. mat.
- La plupart sont propriétaires/privées et/ou payantes (e.g. Dryad, <https://datadryad.org/>)
- Il existe Zenodo (<https://zenodo.org/>) : Il a été créé par OpenAIRE et le CERN pour fournir aux chercheurs un lieu pour déposer des ensembles de données. Il a été lancé en 2013, permettant à des chercheurs de télécharger des fichiers jusqu'à 50 GO.



# -> vers le package R

- Finalement cette architecture n'est si différente de celle d'un package R !
- En intégrant les éléments essentiels de la structure d'un package on bénéficie de tous les outils développés dans ce contexte.



# Research Compendium (ressources)

- Marwick B, Boettiger C, Mullen L. 2018. Packaging data analytical work reproducibly using R (and friends). PeerJ Preprints 6:e3192v2  
<https://doi.org/10.7287/peerj.preprints.3192v2>
- <https://research-compendium.science/>
- <https://zenodo.org/communities/research-compendium/about/>
- <https://github.com/cboettig/template> (structure de package)
- <https://github.com/benmarwick/rrtools> (structure de apckage + intégration docker !!)

## Limitation ...

- Il est compliqué, lorsque l'analyse n'est pas linéaire, de se souvenir de toutes les dépendances ...
  - On est souvent ammené à refaire tourner l'ensemble des analyses ...
  - Et si les calculs sont longs, cela devient vraiment contre-productif !

