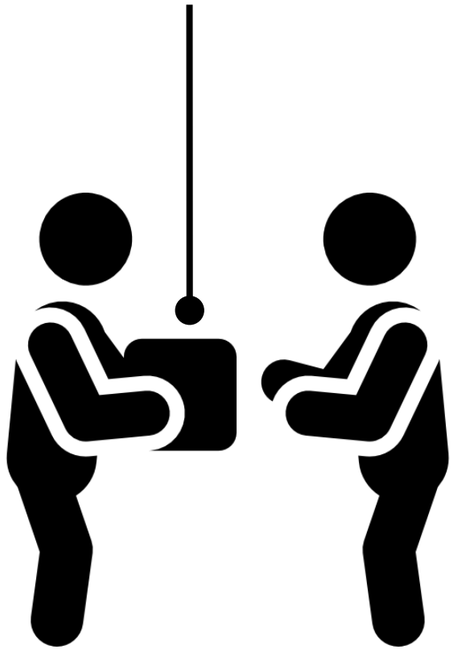
Ateliers R^3

Session 9 - Bonnes pratiques - préparer son code pour une publication

“Any results should be documented by making all data and code available in such a way that the computations can be executed again with identical results. ” (Wikipedia)



Ça devient *obligatoire*

"So, it seemed like a good time to assess the availability and quality of recent data made available with American Naturalist papers. "

"The downside is that only 7 of those 56 of those papers with datasets on DRYAD have provided data in such a way that I, and I assume most users, would find convenient or even comprehensible."

"It is now our Editorial policy that [...] in extreme cases, we reserve the right to retract papers that are not supported by appropriately archived data, or to hold up an author's future submissions until past deficiencies are amended. "

(October 2020)

http://comments.amnat.org/2020/10/data-archiving.html

Les *données* de bases

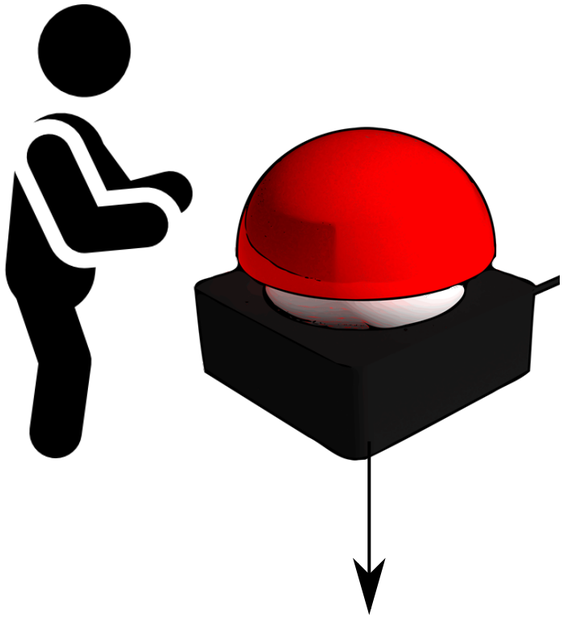
- Données brutes
- Données « propres »

Les *instructions* à réaliser

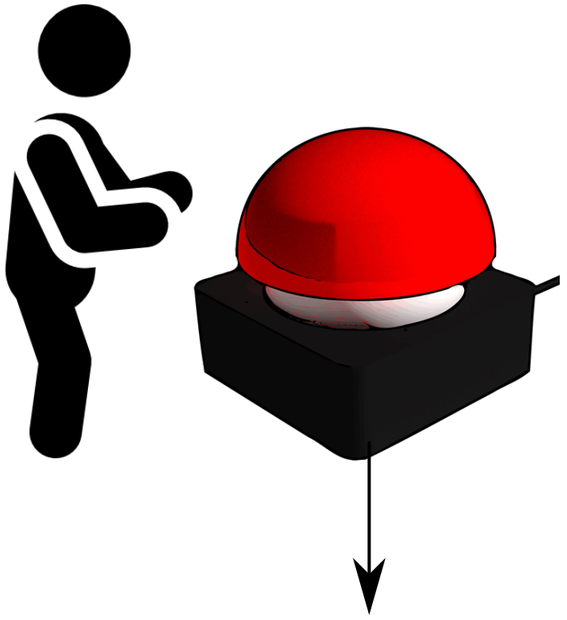
- Instructions écrites
- Scripts

Les *moyens* pour le faire

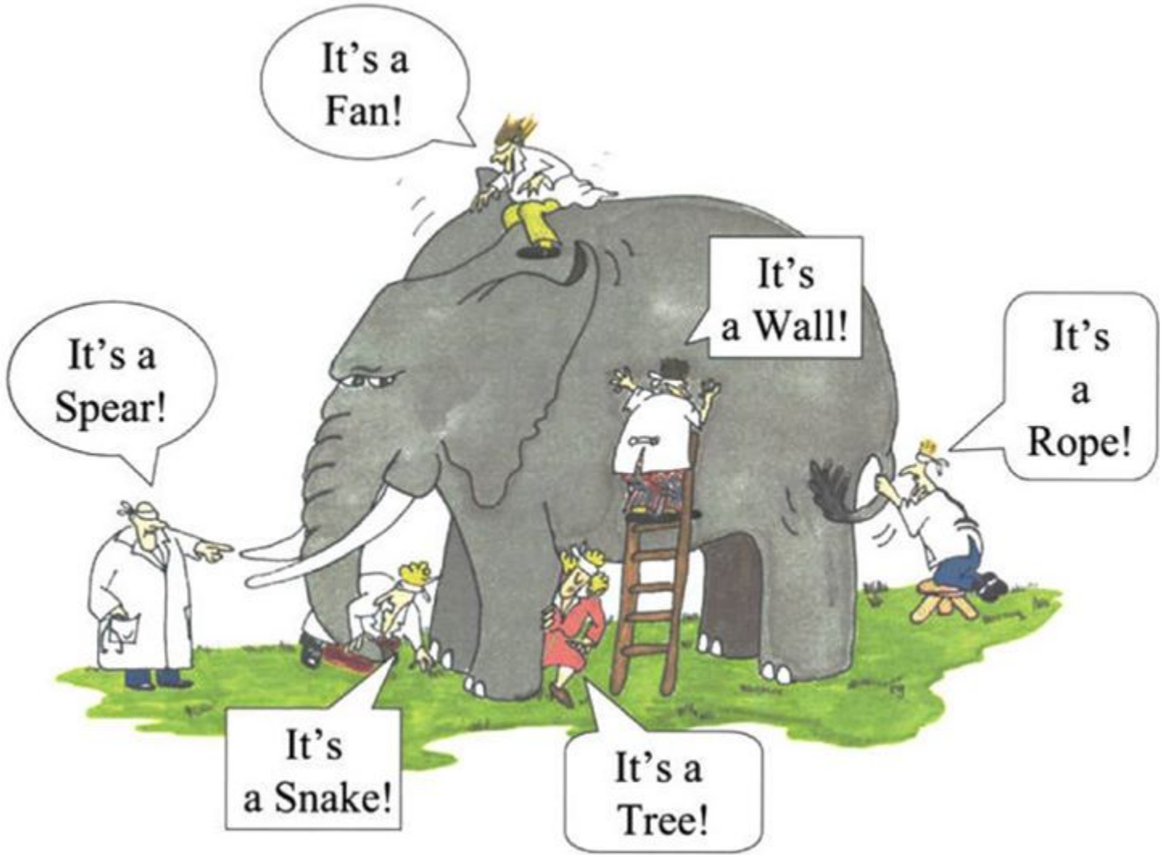
- Logiciels
- Formats lisibles
- Contexte



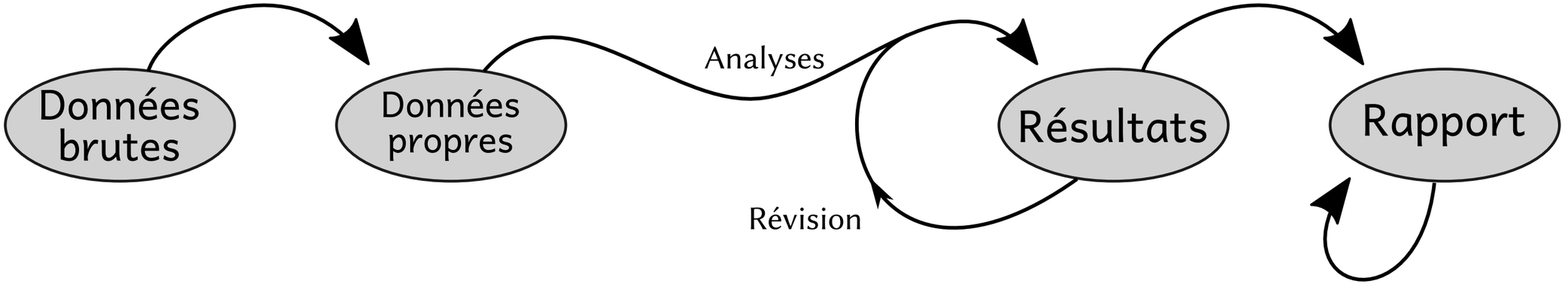
Résultat final



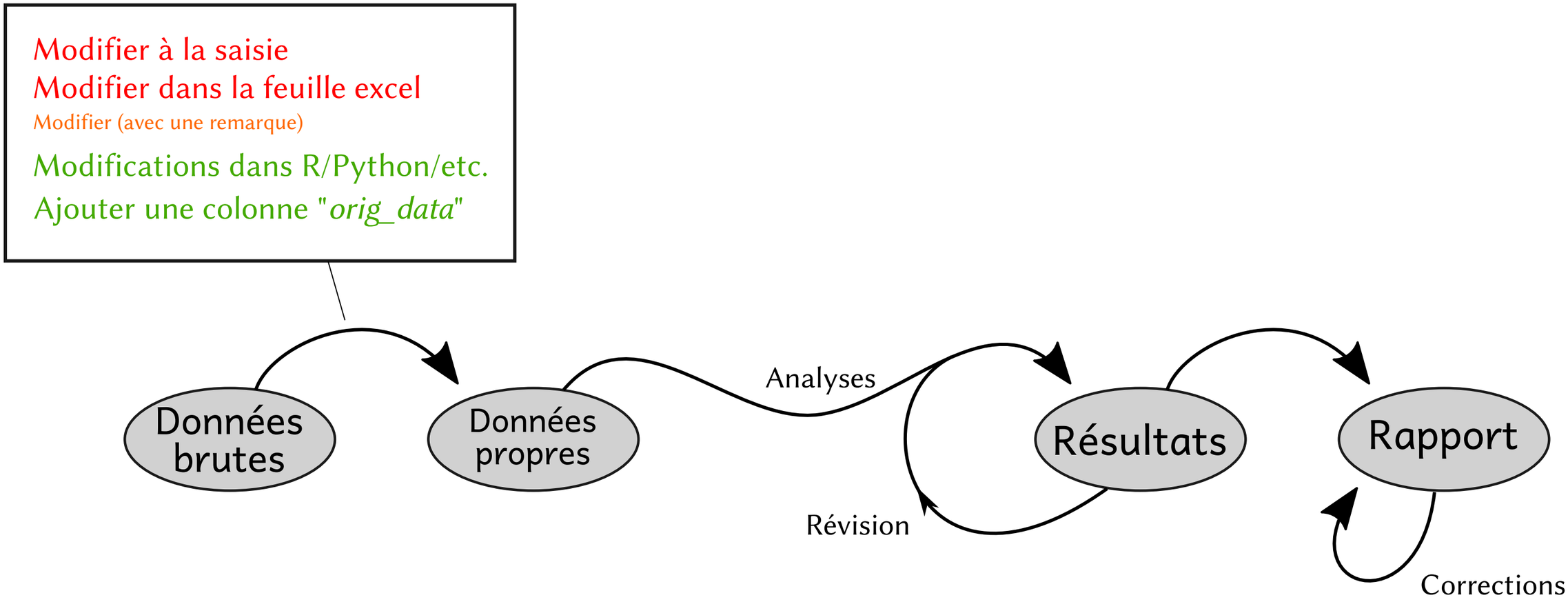
Résultat final



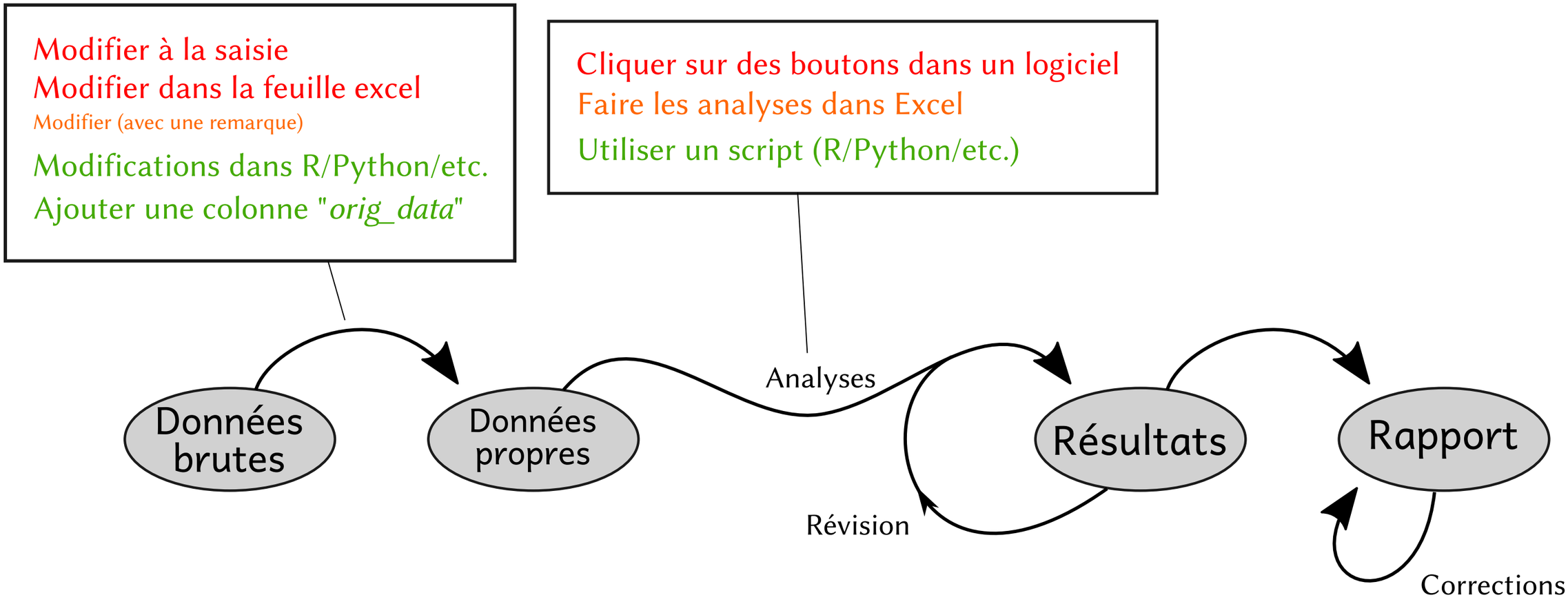
https://www.patheos.com/blogs/driventoabstraction/2018/07/blind-men-elephant-folklore-knowledge/



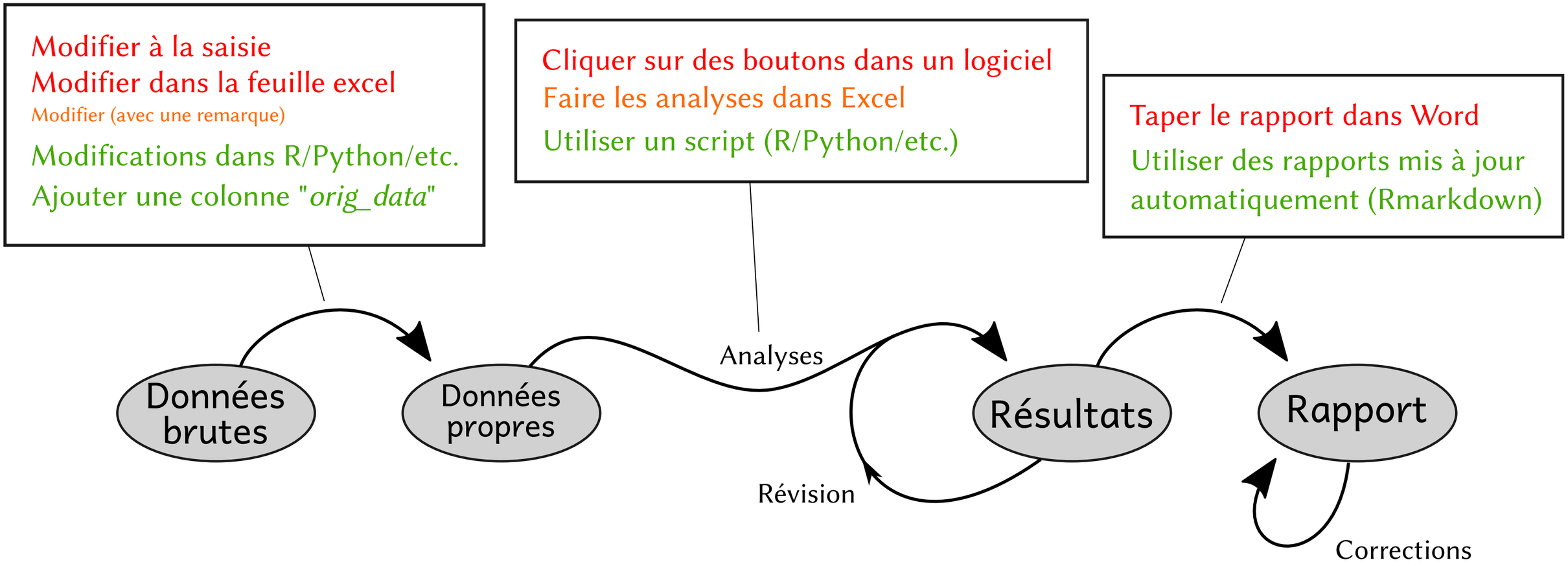
Un *workflow* reproductible



Un *workflow* reproductible

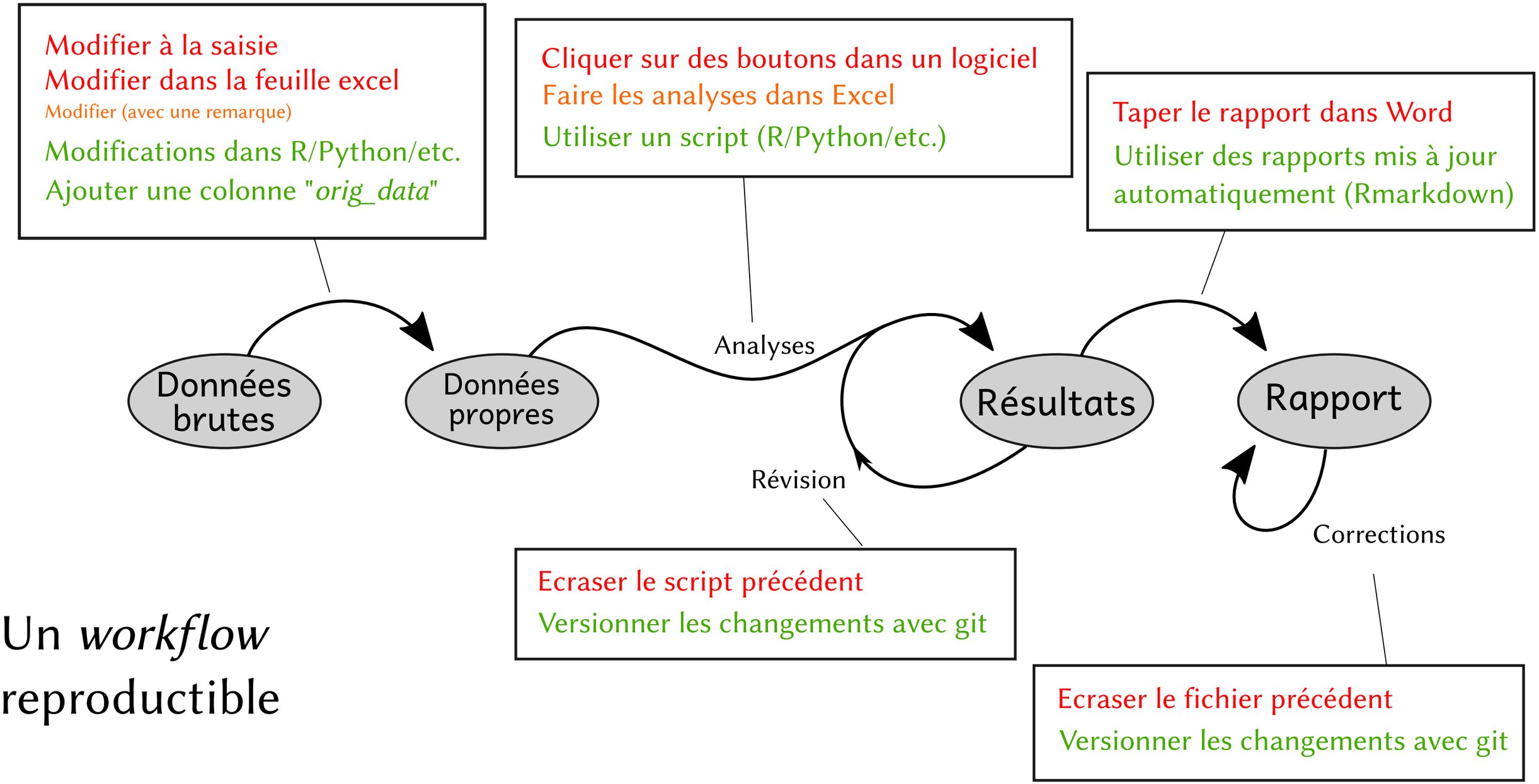


Un *workflow* reproductible



Un *workflow* reproductible

Un *workflow* reproductible



Quelle organisation ?

myproject 

PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS

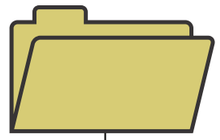
PERSPECTIVE

Good enough practices in scientific computing

Greg Wilson, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, Tracy K. Teal

Published: June 22, 2017 • https://doi.org/10.1371/journal.pcbi.1005510

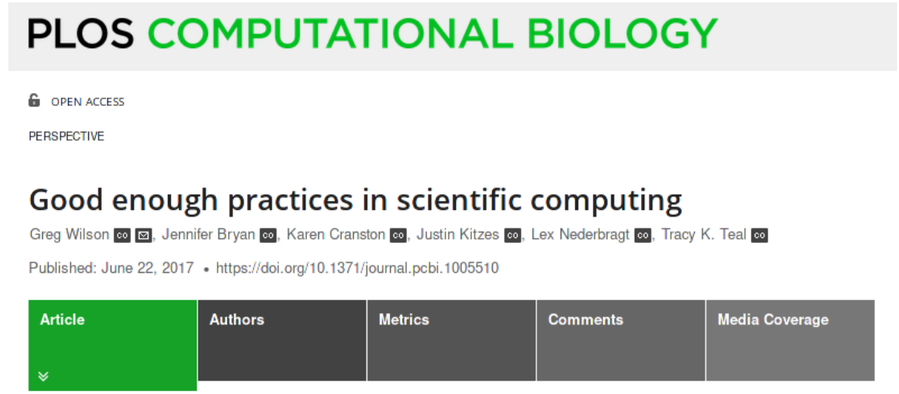
Quelle organisation ?



myproject

Données brutes

- data-raw
  - 2020-10-08
    - fieldwork.csv
  - 2020-10-11
    - soil_measurements.csv
  - 2020-10-11
    - external_db.xls



PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS

PERSPECTIVE

Good enough practices in scientific computing

Greg Wilson, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, Tracy K. Teal

Published: June 22, 2017 • https://doi.org/10.1371/journal.pcbi.1005510

Article | Authors | Metrics | Comments | Media Coverage

Quelle organisation ?

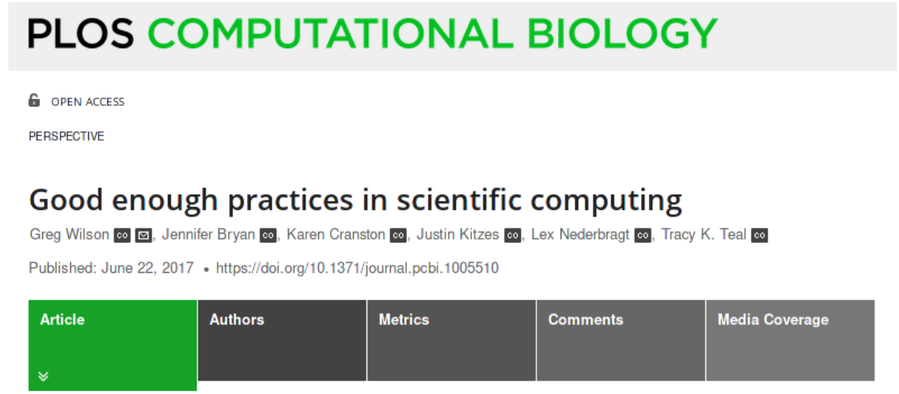
myproject

Données brutes

- data-raw
  - 2020-10-08
    - fieldwork.csv
  - 2020-10-11
    - soil_measurements.csv
  - 2020-10-11
    - external_db.xls

Données propres

- data
  - soil_measurements.rda
  - plant_data.rda
  - convert_raw_data.R



PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS

PERSPECTIVE

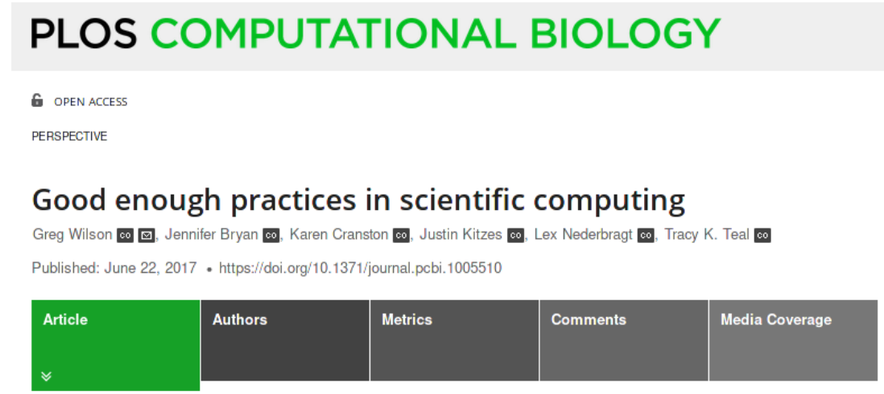
Good enough practices in scientific computing

Greg Wilson, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, Tracy K. Teal

Published: June 22, 2017 • https://doi.org/10.1371/journal.pcbi.1005510

Article | Authors | Metrics | Comments | Media Coverage

Quelle organisation ?


PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS

PERSPECTIVE

Good enough practices in scientific computing

Greg Wilson, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, Tracy K. Teal

Published: June 22, 2017 • https://doi.org/10.1371/journal.pcbi.1005510

Article | Authors | Metrics | Comments | Media Coverage

myproject

- Données brutes
  - data-raw
    - 2020-10-08
      - fieldwork.csv
    - 2020-10-11
      - soil_measurements.csv
    - 2020-10-11
      - external_db.xls
- Données propres
  - data
    - soil_measurements.rda
    - plant_data.rda
    - convert_raw_data.R
- Analyses des données
  - report
    - first_look
      - first_look.Rmd
    - paper_soil_effect
      - analyses.Rmd
      - suppmat.Rmd
      - manuscript.docx
      - data_export
        - soil_measurements.csv
        - plant_data.csv
        - README.txt
      - make_all.sh
      - alldata.rda

Quelle organisation ?

myproject

Données brutes

- data-raw
  - 2020-10-08
    - fieldwork.csv
  - 2020-10-11
    - soil_measurements.csv
  - 2020-10-11
    - external_db.xls

Données propres

- data
  - soil_measurements.rda
  - plant_data.rda
  - convert_raw_data.R

Analyses des données

- report
  - first_look
    - first_look.Rmd
  - paper_soil_effect
    - analyses.Rmd
    - suppmat.Rmd
    - manuscript.docx
    - data_export
      - soil_measurements.csv
      - plant_data.csv
      - README.txt
    - make_all.sh
    - alldata.rda

Versionné et archivé dans un dépot en ligne avec la publication

PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS

PERSPECTIVE

Good enough practices in scientific computing

Greg Wilson, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, Tracy K. Teal

Published: June 22, 2017 • https://doi.org/10.1371/journal.pcbi.1005510

Article | Authors | Metrics | Comments | Media Coverage

Des outils externes

workflowr

```
myproject/
├── .gitignore
├── .Rprofile
├── _workflowr.yml
├── analysis/
│   ├── about.Rmd
│   ├── index.Rmd
│   ├── license.Rmd
│   └── _site.yml
├── code/
│   ├── README.md
├── data/
│   └── README.md
├── docs/
├── myproject.Rproj
├── output/
│   └── README.md
└── README.md
```

workflowr 1.6.2 | Home | Getting started | FAQ | Vignettes | Functions | News

workflowr: organized + reproducible + shareable data science in R

The workflowr R package helps researchers organize their analyses in a way that promotes effective project management, reproducibility, collaboration, and sharing of results. Workflowr combines literate programming (knitr and rmarkdown) and version control (Git, via git2r) to generate a website containing time-stamped, versioned, and documented results. Any R user can quickly and easily adopt workflowr.

Des outils externes

drake

The drake R package

Data analysis can be slow. A round of scientific computation can take several minutes, hours, or even days to complete. After it finishes, if you update your code or data, your hard-earned results may no longer be valid. How much of that valuable output can you keep, and how much do you need to update? How much runtime must you endure all over again?

For projects in R, the `drake` package can help. It analyzes your workflow, skips steps with up-to-date results, and orchestrates the rest with optional distributed computing. At the end, `drake` provides evidence that your results match the underlying code and data, which increases your ability to trust your research.

```
plan <- drake_plan(
  raw_data = readxl::read_excel(file_in("raw_data.xlsx")),
  data = raw_data %>%
    mutate(Ozone = replace_na(Ozone, mean(Ozone, na.rm = TRUE))),
  hist = create_plot(data),
  fit = lm(Ozone ~ Wind + Temp, data),
  report = rmarkdown::render(
    knitr_in("report.Rmd"),
    output_file = file_out("report.html"),
    quiet = TRUE
  )
)

make(plan) # See also r_make().
```

Des outils spécialisés

Outils de traitement de données génétiques

Outils de virtualisation de l'environnement logiciel

Des outils spécialisés

Outils de traitement de données génétiques

Outils de gestion de l'environnement logiciel

```
17
18  ```{r cars}
19  summary(cars)
20  ```
```

Script R

paquets R — vegan v2.0 | ade4 v1.7 | mgcv v9.2

R — R v4.0.1

Libraires systèmes — libBLAS vX.Y | libLAPACK vX.Y

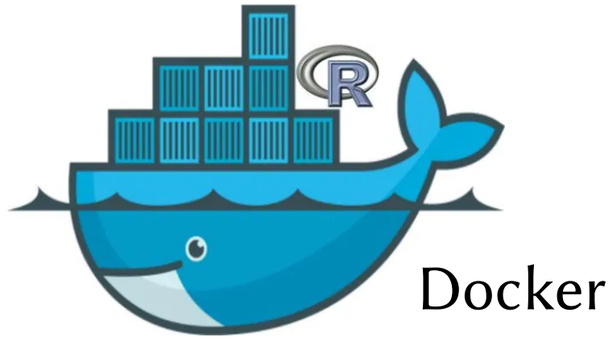
Système d'exploitation — Windows/Linux/MacOS/Solaris

Type d'architecture — x86_64 / x386 / arm

Des outils spécialisés

Outils de traitement de données génétiques

Outils de gestion de l'environnement logiciel



Docker

```
17
18  ```{r cars}
19  summary(cars)
20  ```
```

Script R

paquets R

vegan v2.0 | ade4 v1.7 | mgcv v9.2

R

R v4.0.1

Libraires systèmes

libBLAS vX.Y | libLAPACK vX.Y

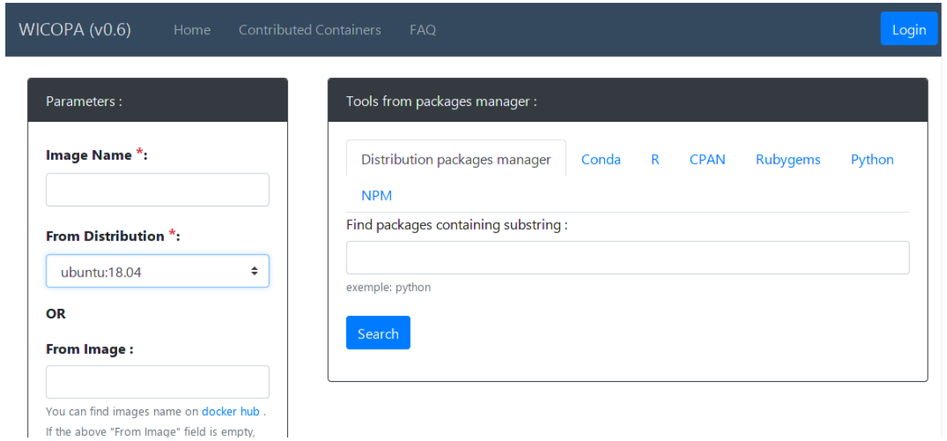
Système d'exploitation

Windows/Linux/MacOS/Solaris

Type d'architecture

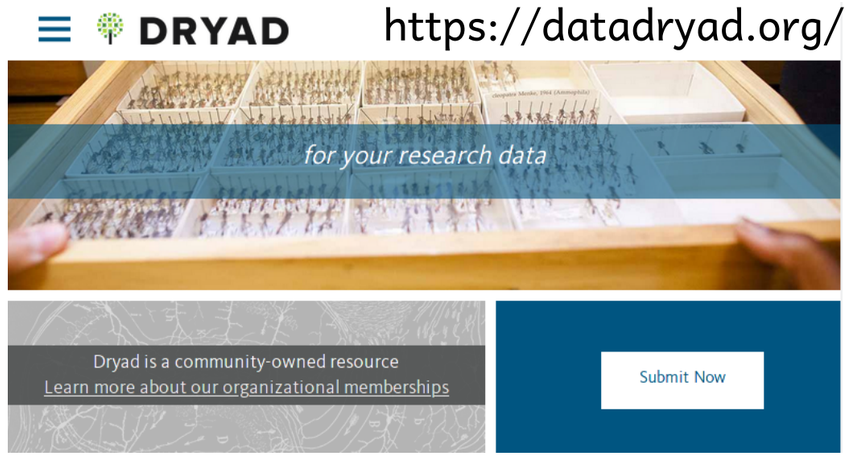
x86_64 / x386 / arm

Outils MBB (e.g. Wicopa)



Archivage dans un dépôt

https://datadryad.org/



https://zenodo.org/

Archivage dans un dépôt

(1) Mettre à jour le code et les résultats

make_all.R

(2) Ajouter un *tag* git

git tag archive_v1.0

(3) Zipper tout les fichiers nécessaires

(4) Mettre en ligne sur le dépôt

https://datadryad.org/

https://zenodo.org/

En pratique

Recap:

- **Une structure minimal pour arriver à un résultat reproductible**
- **Comment archiver son code en ligne**

Atelier: jeudi 17 décembre, 14h !

Tous les exercices et infos sur https://rrr.mbb.cnrs.fr