



Dépôt d'un package R sur Software Heritage et  
référencement sur HAL en vue d'une citation dans  
une publication scientifique

Florent Chuffart

EpiMed, IAB, UGA, Inserm (U1209), CNRS (UMR5309)

Journées du Réseau National de la Recherche Reproductible  
27 mars 2024, Grenoble.

Pourquoi référencer les codes de  
recherche ?

# Pourquoi référencer les codes de recherche ?

## Pourquoi archiver ?

- Le code source est fragile.
- Obsolescence des formats, problème matériel, disparition des forges
- Perte des codes ayant été utilisés pour de la production scientifique

## Pourquoi signaler ?

- Assurer la description
- Faciliter la recherche (e.g. : par domaine scientifique)
- Permettre la citation
- Valoriser les logiciels

1 R et la programmation statistique

2 l'archive ouverte HAL

3 SWH : Software Heritage

4 Démo : R + SWH + HAL

# Plan

1 R et la programmation statistique

2 l'archive ouverte HAL

3 SWH : Software Heritage

4 Démo : R + SWH + HAL

R (1993) est **un logiciel** libre sous licence GPL. Son développement est assuré par sa communauté d'utilisateur en général et par le CRAN (1997) en particulier.

## R et la programmation statistique

R (1993) est **un logiciel** libre sous licence GPL. Son développement est assuré par sa communauté d'utilisateur en général et par le CRAN (1997) en particulier.

R est aussi **un langage de programmation** dédié aux statistiques. A ce titre, peut-on dire que R est un DSL (*Domain-Specific Language*) ?

$a < -1$



```
if (a<-1) {  
  print("OMG!")  
}
```

1->a

```
> df = data.frame(Y=rnorm(100), X1=runif(100),
                  X2=as.factor(c("A", "B")[sample(1:2, 100, replace=TRUE)]))
> head(df)
      Y      X1 X2
1 -1.9907800 0.6948525 B
2 -0.5411307 0.4709891 A
3  0.3035830 0.8641598 A
4 -0.6278876 0.3378647 B
> dim(df)
[1] 100  3
> class(df)
[1] "data.frame"
> typeof(df)
[1] "list"
> class(df$X2)
[1] "factor"
> typeof(df$X1)
[1] "integer"
```

```
> m = lm(Y~X1+X2, df)
> typeof(Y~X1+X2)
[1] "language"
> class(Y~X1+X2)
[1] "formula"
```

```
mat = matrix(rnorm(100), 20)
```

```
res = c()
```

```
for (i in 1:nrow(mat)) {
```

```
  x = mean(mat[i,])
```

```
  res = c(res, x)
```

```
}
```

```
res = apply(mat, 1, function(x) {
```

```
  mean(x)
```

```
})
```

```
res = apply(mat, 1, mean)
```

## R et la programmation statistique

R (1993) est **un logiciel** libre sous licence GPL. Son développement est assuré par sa communauté d'utilisateur en général et par le CRAN (1997) en particulier.

R est aussi **un langage de programmation** dédié aux statistiques. A ce titre, peut-on dire que R est un DSL (*Domain-Specific Language*) ?

Pragmatiquement, R est une calculatrice plutôt complète et dont le succès repose sur sa **communauté** et son mécanisme de **packages**.

## Le package R : définition

*Les packages R sont des extensions du langage de programmation statistique R. Les packages R contiennent du code, des données et de la documentation dans un format **standardisé**. Les packages R peuvent être installés par les utilisateurs depuis des dépôts centralisés tels que CRAN et Bioconductor, ou décentralisé via les nombreux dépôts git disponibles sur le web. Le grand nombre de packages disponibles pour R, ainsi que la facilité de leur installation et de leur utilisation, sont des facteurs clefs de l'essor du langage et la communauté.*

Wikipedia

## Le package R : avantages

- il embarque des métadonnées (fichier DESCRIPTION incluant titre, auteur, licence, ...)
- il organise l'espace de travail (R, vignettes, data).
- il permet de vérifier la qualité du code et de la documentation (`check()/build()` roxygen rmarkdown).
- il propose un mécanisme de test (`testthat`).
- il permet de partager facilement son code (CRAN, bioconductor, github, ...).



## Le package R : cadre formel

Ainsi le package R peut servir de **cadre formel** pour l'analyse ou le développement d'un modèle ou d'une étude statistique.

Valérie Orozco, *Comment améliorer nos pratiques pour aller vers une recherche (plus) reproductible ?*, 2021, **Assemblée générale du RIS** [Vidéos]

Juliette Engelaere-Lefebvre, *PROcessus de Publications REproductibles avec R : la démarche PROPRE*, 2022, **Séminaire RUSS** [Vidéos]

## Le package R : e.g. protopackage

*La programmation orientée prototype est une forme de programmation orientée objet sans classe, fondée sur la notion de prototype. Un prototype est un objet à partir duquel on crée de nouveaux objets.*

Wikipédia

e.g. : JavaScript, groovy

# Plan

- 1 R et la programmation statistique
- 2 l'archive ouverte HAL
- 3 SWH : Software Heritage
- 4 Démo : R + SWH + HAL

# l'archive ouverte HAL

- Archive ouverte pluridisciplinaire
- Initiée en 2000 par le CNRS et exploitée par le CCSD – Centre pour la Communication Scientifique Directe
- Fournit des outils pour l'archivage et la diffusion ouverte des résultats scientifiques.
- Où les scientifiques peuvent déposer leurs résultats académiques dans le respect de leurs droits d'auteur
- Supporte différents types de dépôt (Publications, préprints, rapport, thèses et aussi **logiciel**.)
- Pour une **recherche accessible et ouverte**

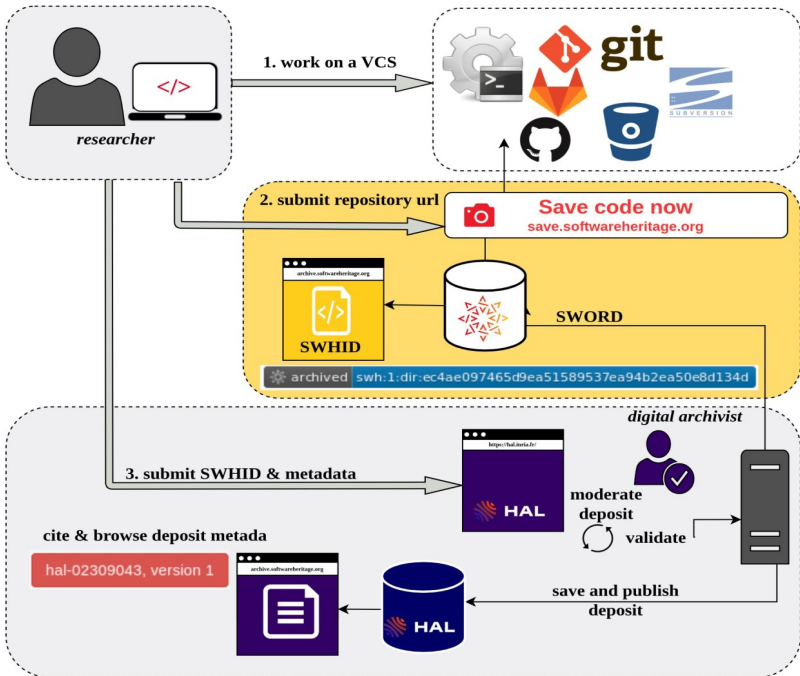
# Plan

- 1 R et la programmation statistique
- 2 l'archive ouverte HAL
- 3 SWH : Software Heritage**
- 4 Démo : R + SWH + HAL

- Initiative dont l'ambition est de construire une **archive universelle des codes sources**
- En les collectant, les préservant et les partageant sur le **long terme**
- Lancée en 2016 par INRIA et soutenue par l'UNESCO
- Collecte de l'intégralité des logiciels disponibles publiquement sous forme de code source.
- Depuis des plateformes d'hébergement de code, comme GitHub, GitLab.com ou Bitbucket, et des archives de paquets.

# Plan

- 1 R et la programmation statistique
- 2 l'archive ouverte HAL
- 3 SWH : Software Heritage
- 4 **Démo : R + SWH + HAL**





<https://github.com/ropensci/codemetar>

<https://github.com/proycon/codemetapy>

Dépôt d'un package R sur Software Heritage et référencement sur HAL en vue d'une citation dans une publication scientifique

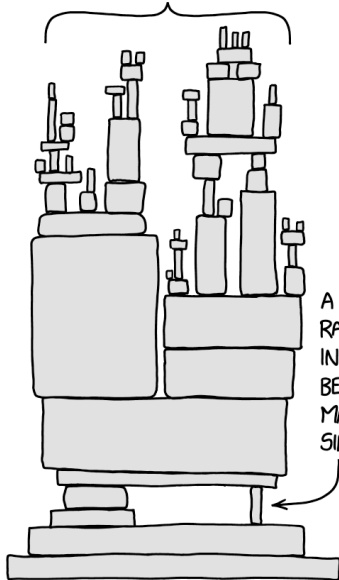
Dépôt d'un package R sur Software Heritage et référencement sur HAL en vue d'une citation dans une publication scientifique

Permet l'archivage et le référencement **sur le très long terme** d'un **version** d'un code de recherche.



Merci à vous...  
et à Lucie Albaret, Alexis Arnaud et Violaine Louvet de la Cellule Data de l'UGA.

ALL MODERN DIGITAL  
INFRASTRUCTURE



A PROJECT SOME  
RANDOM PERSON  
IN NEBRASKA HAS  
BEEN THANKLESSLY  
MAINTAINING  
SINCE 2003

## 2e Plan National pour la Science Ouverte (6 juillet 2021)

Ce nouveau Plan poursuit la trajectoire ambitieuse ...

Il s'appuie sur la politique nationale **des données, des algorithmes** et **des codes sources** impulsée ...

Il s'organise autour de 4 axes :

- Généraliser l'accès ouvert aux publications
- Structurer, partager et ouvrir les données de la recherche
- Ouvrir et promouvoir les **codes sources** produits par la recherche
- Transformer les pratiques pour faire de la science ouverte le principe par défaut