Séminaire RUSS - 3 juin 2022

PROcessus de Publications REproductibles avec R : la démarche PROPRE

La présentation d'aujourd'hui

- La DREAL Pays de la Loire, c'est quoi ?
- La méthode ProPRe : Qu'est-ce que c'est ?
- Comment ça marche pour l'utilisateur ?
  exemple des pages énergie du bilan économique INSEE
- Comment ça marche pour le développeur ?
  exemple des publications RPLS
  - A quoi ça ressemble le produit ?
  - Concevoir un package
  - Mettre à jour les données
- Et pour les chercheurs en SS alors ?

La DREAL Pays de la Loire mais qui est-ce que c'est ?

DREAL Pays de la Loire

PRÉFET
DE LA RÉGION
PAYS DE LA LOIRE

Liberté
Égalité
Fraternité

- direction régionale de l'environnement, de l'aménagement et du logement
- un service déconcentré du Ministère de la la transition écologique et de la cohésion des territoires et du Ministère de la transition énergétique
- un service qui produit des données et qui a dans ses attributions d'apporter de la connaissance, un éclairage objectif, sur les politiques de son ressort

BEAUCOUP de valorisations

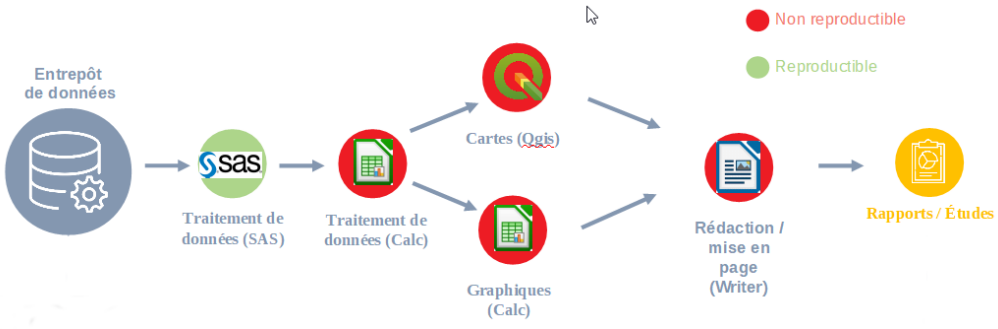
A mettre à jour régulièrement

Autrefois, le problème...

Autrefois, le problème...

Des problèmes d'outillage

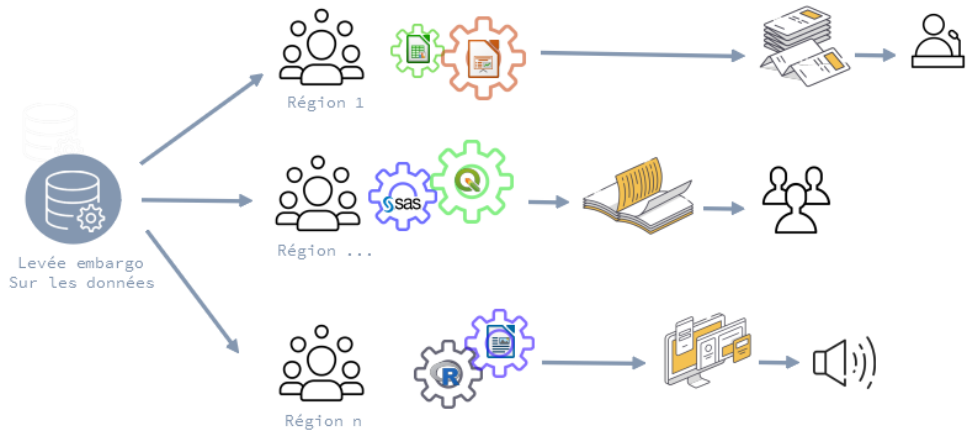
- beaucoup de temps perdu à faire des traitements de données semblables d'une année sur l'autre, voire d'un trimestre à l'autre
- les outils pour analyser les données sont multiples et hétérogènes, les solutions sont peu portables
- les travaux sont à refaire de façon périodique



...communs aux différentes régions

Des disparités organisationnelles

- les formats de valorisation sont hétérogènes et non comparables
- les délais de publication sont longs et inégaux
- un suivi de l'impact des publications sur le lectorat inégal



Et si…

Et s'il existait un outil mutualisé, presque clef en main, qui permettrait aux DREAL/DEAL de publier périodiquement leurs analyses de façon homogène dans des délais raccourcis ?

La méthode ProPRe :

mais qu'est-ce que c'est ?

Les objectifs de la démarche

Une méthode :

- pour réaliser des **publications statistiques** ou applications de **datavisualisation**
- qui cherche à assurer la **reproductibilité** du produit :
  - dans le temps, d'une année ou d'un trimestre à l'autre) ;
  - et dans l'espace (d'un territoire à l'autre).

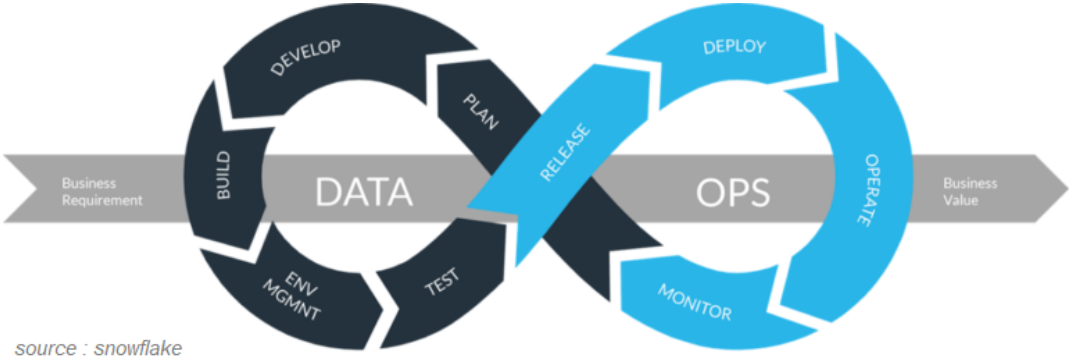
ProPRe = **Pro**cessus de **P**ublication **Re**productible

3 sources d'inspiration

Cette méthode tire parti :

- de la rigueur et des outils du développement logiciel open source,
- de l'accélération des productions, dans le sillon de la mouvance devOps,
- des innovations méthodologiques spécifiques aux données, réalisées dans le cadre du développement de la science ouverte et reproductible.

La méthode ProPRe s'inscrit dans la dynamique **dataOps**.



source : snowflake

Et en pratique ?

La méthode ProPRe est aujourd'hui basée sur le langage de traitement de données R.

Elle consiste à structurer les publications sous la forme d'un **package R** pour :

- diffuser des données,
- le canevas d'une publication relative à ces données,

prête à être personnalisée, voire publiée.

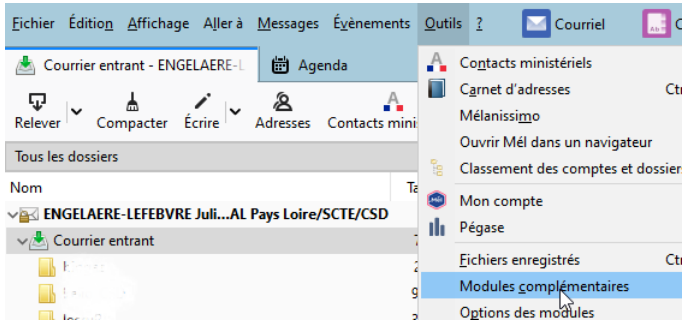
un package R ??

R par le début

R est à la fois un langage de programmation et un logiciel open source.

Comme beaucoup de logiciels open source, on peut lui adjoindre des fonctionnalités supplémentaires, grâce à ce qu'on appelle :

- des extensions (Firefox),
- des plugins (QGIS),
- modules (Thunderbird),
- addin (MS Office),
- addon (jeux vidéo)...



Pour R, ces extensions sont appelées des **packages**.
Elles s'installent facilement et sur toute sorte de machines (linux, windows, MAC).

Des pratiques inspirées du développement logiciel...

La méthode ProPre consiste à dévoyer les procédures du développement informatique des packages pour les adapter aux besoins des publications reproductibles.

Les packages ProPre doivent être développés selon les canons de la rigueur du développement informatique open source, avec des tests unitaires d'intégrité, de bon fonctionnement, la documentation de chaque fonctionnalité.

La structure de chaque package est identique, normalisée, et des dispositifs existent pour automatiser et faciliter son développement (documentation, test, check, déploiement).

On adopte les principes de l'agilité avec la livraison, par intégration continue, d'un produit minimum viable qui s'enrichit au fur et à mesure des itérations. On s'accorde le droit à l'expérimentation...

Des pratiques inspirées du Royaume Uni...

On doit tout aux britanniques et à leur service statistiques gouvernemental, qui ont théorisé l'approche dans le 'RAP' = 'reproductible analytical pipeline'.

- des packages R pour chaque publication,
- basés sur d'autres packages socles (mise en page et autres taches récurrentes...) on sépare les préoccupations !

https://ukgovdatascience.github.io/rap_companion/

...transposées aux besoins des publications

Un package R peut comprendre à la fois des **fonctions** de création d'illustrations, de commentaires automatiques, de lancement d'une application de datavisualisation, mais aussi des **données**.

La spécificité des package méthode ProPre est de contenir un **template** (un canevas de publication) et une fonction d'appel de ce canevas, afin de le personnaliser.

Le dépôt d'un package sur une **forge** de développement logiciel comme github ou gitlab :

- **suffit à sa diffusion**,
- facilite la collaboration entre développeurs,
- permet le déploiement automatique de la documentation (données, fonctions, vignettes).

Les acteurs d'un projet ProPRe

Comment ça marche pour l'utilisateur ?

Exemple des pages énergie du bilan économique régional annuel de l'INSEE

Démonstration

Le dépôt du projet : https://gitlab.com/dreal-datalab/bilan.eco.insee

Lancer un environnement d'exécution R 'vierge' sur le SSP Cloud :
https://datalab.sspcloud.fr/accueil

Les lignes à lancer :

```
remotes::install_gitlab(repo = "dreal-datalab/bilan.eco.insee", build_vignettes = TRUE, depend
# remotes::install_github("pachevalier/tricky")
bilan.eco.insee::edit_rapports(repo = "bes_pdl_2022", changer_rep_travail = TRUE)
```

Comment ça marche pour le développeur ?

Exemple des publications propre.rpls

- Le produit et les travaux
- Concevoir un package R
- Mettre à jour les données

Contexte

- RPLS = Répertoire du parc locatif social, enquête nationale annuelle sur l'état du parc locatif social réalisée par le SDES (Service des données et études statistiques)
- Des valorisation annuelles réalisées dans toutes les régions ou presque
- le premier cas d'école de la méthode (démarrage mi 2020)

Une petite équipe avec une diversité de compétences

- des compétences éditoriales pour constituer un plan problématisé

- des compétences en sémiologie graphique pour proposer des illustrations adaptées

- une connaissance de la source des données

- des compétences d'animation

- des compétences de développement en R, markdown et en GIT, déjà présentes ou acquises pour l'occasion, chez les dev et les éditos

Ce qui a été accompli

- la preuve que ça marche ! des gains de temps de l'ordre de 2 à 3 semaines.agent / région, le recentrage sur des activités à forte valeur ajoutée
- une méthode de travail détaillée dans un guide
- 14 publication déployées le jour de la levée de l'embargo sur les données
- un 4 pages de présentation grand public de la démarche
- une application de consultation des indicateurs RPLS au territoire embarquée dans le package qui peut être déployée pour tout à chacun
- un package documenté, testé, versionné qui a permis d'actualiser ces publications dans de bonnes conditions en 2021 : 1/2 journée nécessaire pour incorporer le nouveau millésime au package.

Un exemple de publication

https://dreal.statistiques.developpement-durable.gouv.fr/parc_social/2021

Comment ça marche pour le développeur ?

Ressources pour concevoir un package R

- utiliser {fusen} de Sébastien Rochette (ThinkR) : pour simplement transformer un rapport Rmd en un package.
- 6 ateliers pour s'initier pas à pas à la création de packages R : ateliers centrés sur les packages avec templates et data type ProPRe.
- la bible sur les packages : R packages par Hadley Wickham et Jenny Bryan (RStudio).
- mettre une application de dataviz RShiny en package avec Golem : découverte (ThinkR) et approfondissement (ThinkR)
- utiliser l'intégration continue de gitlab pour déployer la documentation d'un package et réaliser des tests sur ce dernier : package {gitlabr}

Où est-ce que ça se passe ?

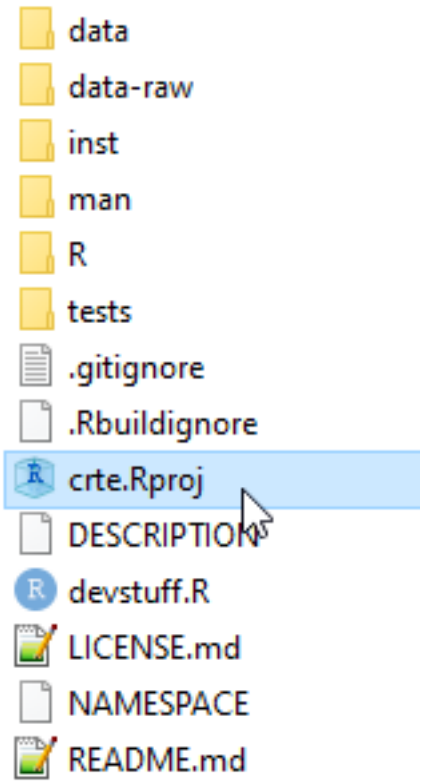
Structuration d'un répertoire de développement d'un package :

`DESCRIPTION` : les métadonnées du packages, dont les dépendances

`inst` : le template

`R` : les fonctions

`data-raw` : les scripts d'incorporation des données



Qu'est-ce qu'il faut faire pour mettre à jour des données ?

Dans data-raw :

- Ouvrir chacun des scripts
- Éventuellement y mettre à jour les connecteurs,
- Exécuter les scripts

Des exemples de projets ProPRe

https://gitlab.com/explore/projects/topics/propre

Et pour les chercheurs en sciences sociales alors ?

Cas d'usages de la méthode propre en SS

- respecter les standards de la science ouverte en partageant les données, pour par exemple :
  - partager les observations et analyses à challenger,
  - partager les observations que des collègues /confrères peuvent analyser selon leurs propres préoccupations ;
- mettre à jour une publication régulière ;
- collaborer facilement avec des confrères formés à la démarche ;
- paramétrer des rapports selon des mots clefs d'analyses des réseaux sociaux ;
- ...

Merci de votre attention