Module 6 - Publications reproductibles avec RMarkdown

Groupe des référents R

19 janvier 2019

# Introduction

## Le parcours de formation

Le parcours de formation proposé est structuré en modules de 2 jours chacun :

* 1 : Socle de base
* 2 : Préparer ses données avec le tidyverse
* 3 : Statistiques descriptives
* 4 : Analyses multivariées
* 5 : Datavisualisation : Produire des graphiques, des cartes et des tableaux
* 6 : Documents reproductibles avec RMarkdown

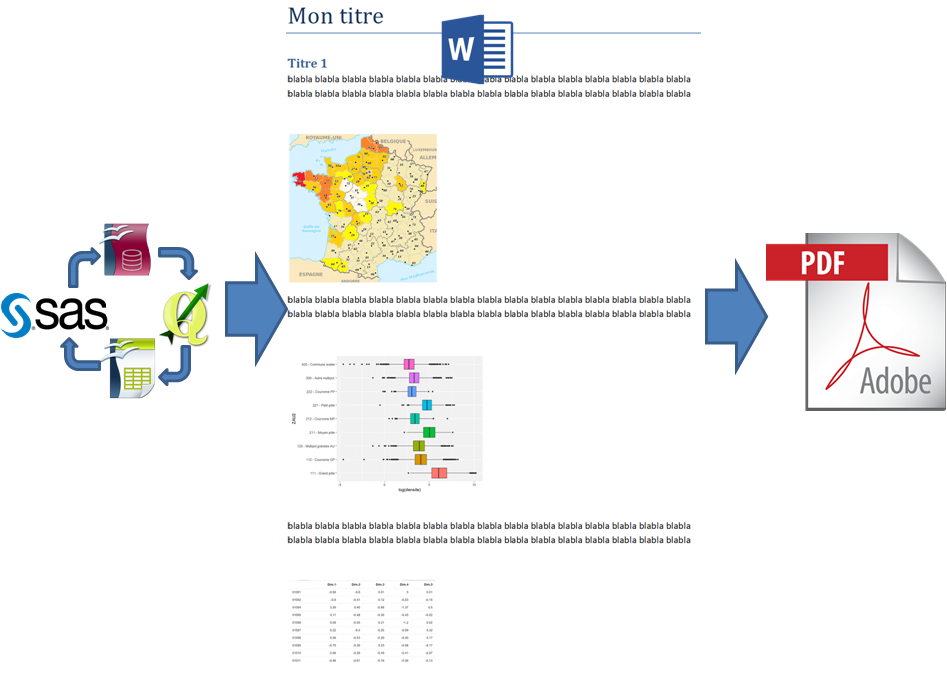
… et en perspective : analyse spatiale, applis interactives avec Shiny, big data et optimisation du code, etc.

Pour vous tenir au courant de l’offre de formation proposée par le réseau des CVRH, [consultez la plateforme OUPS](http://oups-cmvrh.e2.rie.gouv.fr/). Vous pouvez vous y abonner pour recevoir les annonces qui vous intéressent.

Il existe une liste pour diffuser de l’information, échanger autour de R ou lever des points de blocage. Pour s’insrire, envoyer un message vide avec le titre “subscribe labo.communaute-r” à l’adresse [sympa@developpement-durable.gouv.fr](mailto:sympa@developpement-durable.gouv.fr).

## La publication reproductible

Le schéma classique pour produire un rapport ou une publication ressemble à ça :



Les données mobilisées proviennent de bases de données, de tables ou de SIG, et rien que pour les lire, il y a besoin de plusieurs logiciels selon les formats.

Les données sont ensuite nettoyées (valeurs manquantes ou aberrantes, recodages, discrétisation, renommages …), puis traitées (sélection, agrégation, modélisation) et visualisées (tables, graphiques, cartes). Les éléments visuels sont insérés dans un traitement de texte qui sert aussi à la rédaction, et enfin la version diffusable est produite au format pdf.

Plusieurs inconvénients à cette chaîne d’opérations :

* Beaucoup d’opérations manuelles chronophages et sources potentielles d’erreurs.
* Si quelque chose change, il faut tout recommencer (mise à jour, adaptation à une autre zone géographique ou fenêtre temporelle).
* La publication finale n’est pas “reproductible” au sens où la traçabilité est insuffisante pour qu’un autre auteur, avec les mêmes données, arrive à la même publication la recherche d’erreurs est difficile et les passations délicates.
* Les contenus sont statiques, ce qui est peu attractif par rapport aux possibilités d’interactivité offertes par les technologies web.

Le fil conducteur de ce module est la production de A à Z d’un portrait de territoire, à partir de données variées, en travaillant entièrement dans RStudio au format R Markdown.

## Le format “R Markdown”

Ce format de document a été introduit en 2012 à la sortie du package knitr.

Auparavant, il existait Markdown, un “langage de balisage léger”, sorte d’intermédiaire entre du texte brut et du traitement de texte. La syntaxe de Markdown est très simplifiée. On peut le saisir avec un éditeur de texte comme *Notepad* et il est lisible même non formaté.

Pourquoi utiliser un langage de balisage léger ? Des avantages majeurs :

* Séparer le fond de la forme. On peut taper “au kilomètre” en se concentrant sur le contenu, et voir ensuite les questions de forme (y compris le format de sortie, HTML, PDF, traitements de texte, diaporamas, tableaux de bord, voire EPUB).
* Cette simplicité permet le suivi de version par des outils comme [Git](https://fr.wikipedia.org/wiki/Git).

Par contre, la simplicité signifie peu de fonctionnalités donc des sorties assez pauvres.

L’idée, avec R Markdown, était garder les avantages du langage existant Markdown mais en lui ajoutant des fonctionnalités d’interprétation de code R, intégré aux documents sous formes de chunks. En fait, les versions actuelles du package knitr permettent d’interpréter d’autres langages comme du $\LaTeX$ ou du code Python.

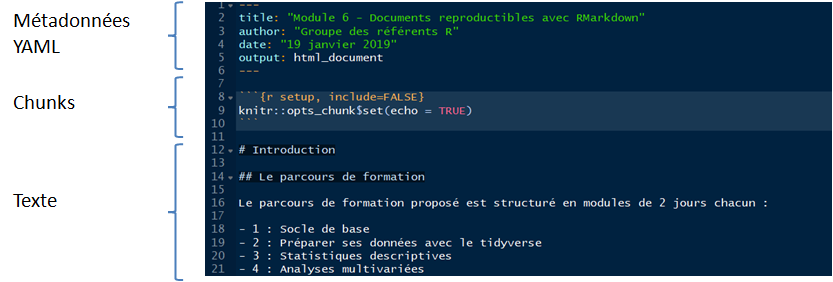
R Markdown est en quelque sorte un cadre pour travailler en data science. Un même document permet de sauvegarder et exécuter du code ainsi que de produire des rapports mis en forme pouvant contenir une grande diversité d’objets statiques ou dynamiques.

# Le fichier R Markdown

## Nouveau fichier

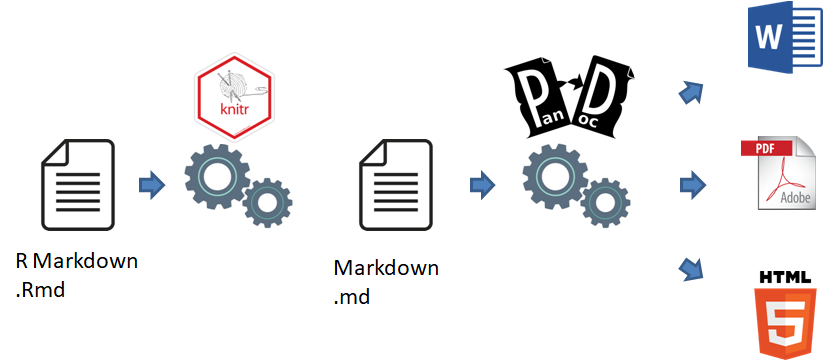
## Structure

Un fichier R Markdown est constitué de 3 éléments principaux.



La partie texte peut contenir, en plus, divers éléments qui peuvent être interprétés comme des balises HTML ou du code $\LaTeX$ ou CSS.

## Du .Rmd au format de sortie



On distingue deux étapes :

* La première, réalisée par le package knitr, interprète le document R Markdown et exécute les chunks. Elle produit un fichier Markdown contenant le texte, le code et le résultat de l’exécution du code.
* Puis, le convertisseur de formats Pandoc se charge de produire le document de sortie au format choisi.

Ca peut paraître compliqué mais en pratique, chacune des fonctions render de knitr exécute la chaîne du début à la fin.

# Les différents éléments d’un fichier .Rmd

## Les titres

## Insérer des images

dimensions Lien. Stockage dans un sous-répertoire.

## Insérer des liens hypertexte

## Insérer du code

dimensions

## Insérer des éléments en LaTeX

$\LaTeX$ permet de mettre en forme toutes les expressions mathématiques.

## 

# Paramétrage général du fichier .Rmd

Dimension, format de sortie, table des matière, css, yml,

# Les chunks

Enchaînement, paramétrage,

# 

# A ne pas oublier

Utilisation backticks Latex Bookdown Balises Table des matières css liens images versioning mise en page Flexdashboard nommer les chunks Les thèmes ggplot pour assurer la présentation homogène des fraphiques + les fonctions custom

# Rappels du jour 1

# Exercice

## Cadrage

### Problématique

### Rendu

Pour ce rendu, quelles données ? Quels packages ? ### Données dispo

## Mise en page

### Quel type ? Barre de navigation ou onglets ?

### Un petit CSS

### Logo

### Métadoonnées

## Importation des données

### jointure des tables

### Données tabulées

### Requêter une API

### Données géo

## Création des objets

### Tables

packages DT, [formattable](https://www.littlemissdata.com/blog/prettytables) Fonction knitr::kable et kableextra cf. exemple [ici] (<https://haozhu233.github.io/kableExtra/awesome_table_in_html.html>)

### Graphiques

Types de graphiques; interactivité plotly et highcharter

### Cartes

leaflet ; fond de carte OSM géolocalisation : corresp adresses-données formats de cartes geojson, shp donnée ponctuelle, polygones

## Assemblage

dimensions formats de sortie. Besoin de Latex ?

règles de semiologie dans m5 hébergement html tenir Thierry au courant