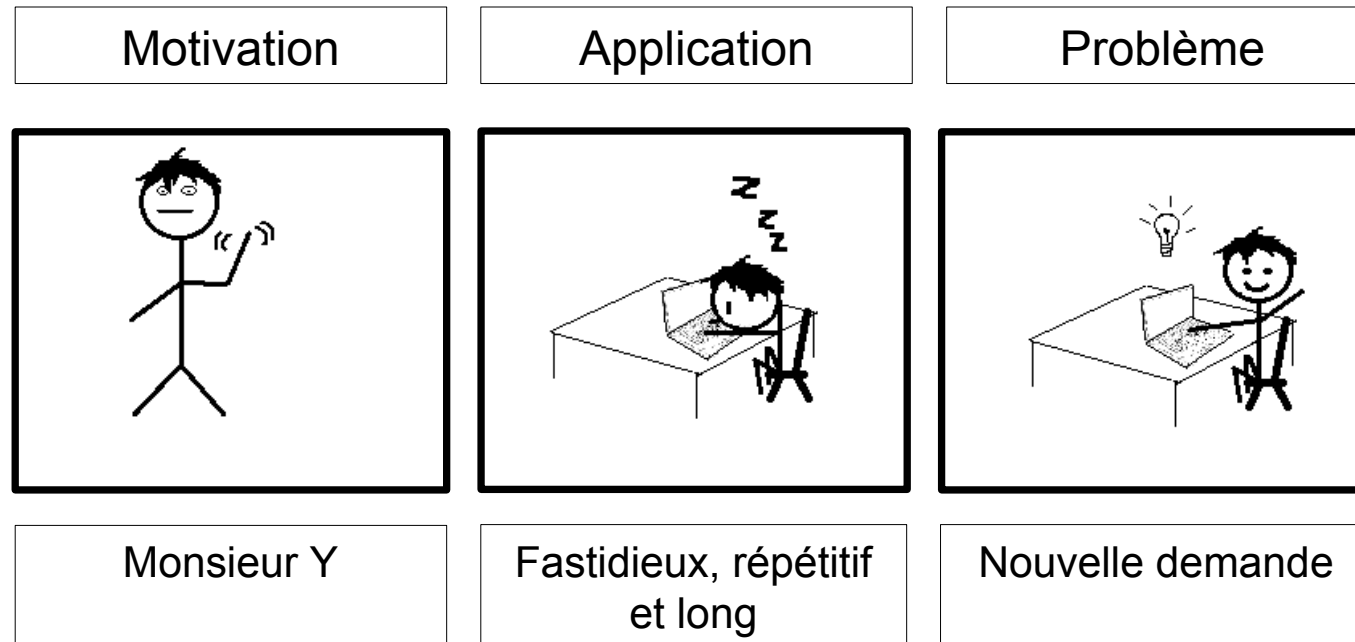


# Introduction au web scraping avec **rvest**

*Stéphanie Terrasse*  
24/01/2018



# Introduction



# Sommaire

- DÉFINITION
- LES OUTILS
- INTRODUCTION À RVEST
- LÉGISLATION

# Définition



## Scraping :

- **Extraction du contenu** d'un ou plusieurs documents web (pages web, contenu renvoyé par un web service)
- Transformation d'une donnée non structurée en une donnée structurée pour analyse

# Les outils

- Un navigateur !



- Les langages :

- R : **rvest**, Rcrawler, XML
- Python : Requests, BeautifulSoup, lxml, Selenium, Scrapy
- Ruby
- ...



- Autres

- Import.io
- Webhose.io
- Scrapex
- Webscraper (extension de Chrome)
- ...



import.io

# Introduction à rvest

*Développé par Hadley Wickham*



## Les étapes majeures :

- Identification des éléments du code source : Ctrl + Maj + I ou F12
- Récupération du contenu avec rvest
  - `read_html( )`
  - `html_nodes( )`
  - `html_text( ), html_attr( ) ...`

# Qu'est-ce qu'un site Web ?

- Un **ensemble de pages** codées en HTML (décrit à la fois le contenu et la forme d'une page Web)
- Structurer le contenu d'une page HTML : les **balises**
  - *Exemples : <p>, <strong>, <head>...*
- Le style de la page Web : fichier(s) **CSS**

# Qu'est-ce qu'un site Web ?

Titre	Année	Durée
Coco	2017	109
Titanic	1997	195



```
<table>
```

```
<tr>
```

```
<th>Titre</th> <th>Année</th> <th>Durée</th>
```

```
</tr> <tr>
```

```
<td>Coco</td> <td>2017</td> <td>109</td>
```

```
</tr> <tr>
```

```
<td>Titanic</td> <td>1997</td> <td>195</td>
```

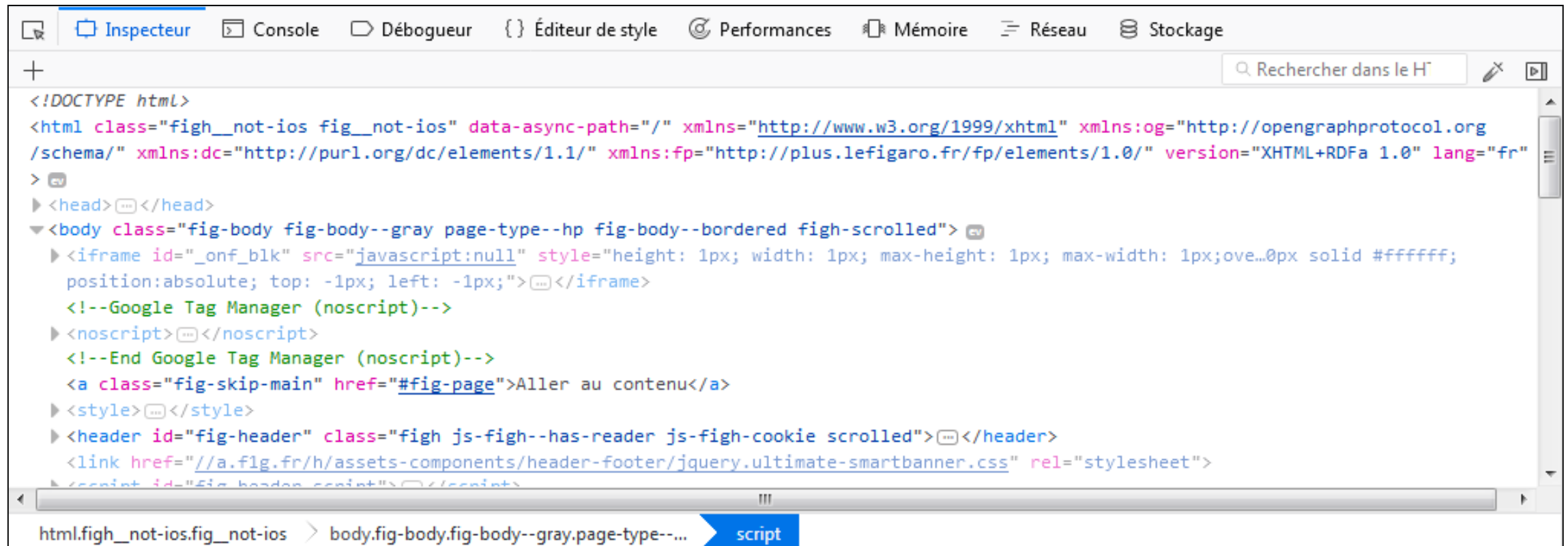
```
</tr>
```

```
</table>
```



# Identification des éléments HTML

- **Ctrl + Maj + I** : affiche le code de la page avec les balises qui vont servir à identifier les éléments que l'on veut récupérer



The screenshot shows a web browser's developer tools interface. The 'Inspecteur' (Inspector) tab is active, displaying the HTML source code of the page. The code is color-coded and includes various attributes like class, data-async-path, xmlns, and href. The browser's toolbar at the top includes icons for Console, Débogueur, Éditeur de style, Performances, Mémoire, Réseau, and Stockage. A search bar is visible in the top right of the developer tools area.

```
<!DOCTYPE html>
<html class="figh_not-ios fig_not-ios" data-async-path="/" xmlns="http://www.w3.org/1999/xhtml" xmlns:og="http://opengraphprotocol.org/schema/" xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:fp="http://plus.lefigaro.fr/fp/elements/1.0/" version="XHTML+RDFa 1.0" lang="fr">
  <head>
    <body class="fig-body fig-body--gray page-type--hp fig-body--bordered figh-scrolled">
      <iframe id="_onf_blk" src="javascript:null" style="height: 1px; width: 1px; max-height: 1px; max-width: 1px; position: absolute; top: -1px; left: -1px;">
      <!--Google Tag Manager (noscript)-->
      <noscript>
      <!--End Google Tag Manager (noscript)-->
      <a class="fig-skip-main" href="#fig-page">Aller au contenu</a>
      <style>
      <header id="fig-header" class="figh js-figh--has-reader js-figh-cookie scrolled">
      <link href="//a.flg.fr/h/assets-components/header-footer/jquery.ultimate-smartbanner.css" rel="stylesheet">
      <script id="fig-header-script">
```

# Récupération du contenu

```
> en_ce_moment
[1] "Aujourd'hui sur Figaro Live"
[2] "Notre-Dame-des-Landes"
[3] "Trump enchaîne les polémiques"
[4] "Affaire Lactalis"
```

Lire une page Web : **read\_html()**

```
library(rvest)

# LE FIGARO

url_le_figaro <- "http://www.lefigaro.fr/"
le_figaro <- read_html(url_le_figaro)
```

```
> le_figaro
{xml_document}
<html lang="fr" data-async-path="/" xmlns=
"http://www.w3.org/1999/xhtml" xmlns:og="h
ttp://opengraphprotocol.org/schema/" xmlns
:dc="http://purl.org/dc/elements/1.1/" xml
ns:fp="http://plus.lefigaro.fr/fp/elements
/1.0/" version="XHTML+RDFa 1.0">
[1] <head>\n<meta http-equiv ...
[2] <body class="fig-body f ...
```

Récupération d'un contenu : **html\_nodes()**, **html\_text()**

```
en_ce_moment <- html_nodes(le_figaro, "div.fig-en-ce-moment.fig-en-ce-moment--actu") %>%
  html_text()
> en_ce_moment
> en_ce_moment
[1] "\n          En ce moment\n          \n          \n          Aujou
rd'hui sur Figaro Live\n          \n          \n          ctu") %>%
\n          Notre-Dame-des-Landes\n          \n
\n          Trump enchaîne les polémiques\n
```

# Récupération du contenu

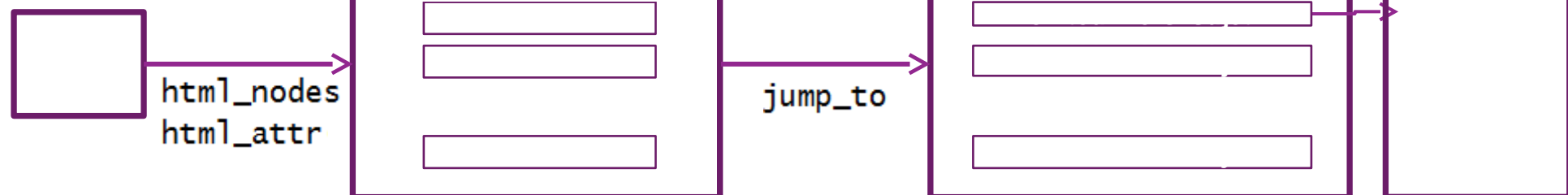
## Récupération des articles

```
## création d'une session virtuelle  
sF < html_session(url_le_figaro)
```

```
> sF  
<session> http://www.lefigaro.fr/ sujets  
Status: 200 _figaro, "div.fig-en-ce-moment.fig-en-ce-moment--actu a") %>%  
Type: text/html; charset=UTF-8 (sf")  
Size: 532653
```

```
> en_ce_moment_links
```

```
[1] "http://www.lefigaro.fr/lefigaromagazine/2018/01/08/01006-20180108ARTFIG00084-aujourd-hui-sur-figaro-live.php"  
[2] "http://www.lefigaro.fr/actualite-france/dossier/projet-d-aeroport-notre-dame-des-landes-zadistes-contre-gouvernement"  
[3] "http://www.lefigaro.fr/international/dossier/donald-trump-actualite-et-polemiques"  
[4] "http://www.lefigaro.fr/societes/dossier/affaire-lactalis-le-scandale-sanitaire-du-lait-contamine"
```



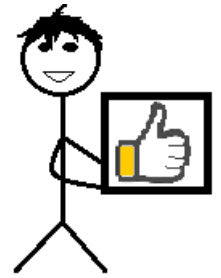
# Récupération du contenu

## Récupération des articles

```
le_figaro_df <- data.frame( "Journal"="", "Date"=Sys.Date(), "Type_article"="", "Theme"="",  
                           "Titre"="", "Contenu"="", stringsAsFactors = FALSE )  
  
for( i in 1:length(en_ce_moment_contents)){  
  top5_links <- as.array( html_nodes(en_ce_moment_contents[[i]], "h2.fig-profile__headline a") %>%  
                        html_attr("href") ) [1:5]  
  top5_contents <- apply( as.array( top5_links ), 1, FUN=jump_to, x=sF )  
  for( j in 0:( length( top5_links ) -1 ) ){  
    le_figaro_df[i+j, "Journal"] <- "Le Figaro"  
    le_figaro_df[i+j, "Date"] <- Sys.Date()  
    le_figaro_df[i+j, "Type_article"] <- "En_ce_moment"  
    le_figaro_df[i+j, "Theme"] <- en_ce_moment[i]  
    le_figaro_df[i+j, "Titre"]<- html_nodes(top5_contents[[1]], "h1.fig-main-title") %>%  
                              html_text()  
    le_figaro_df[, "Contenu"] <- cf. script  
  }  
}
```

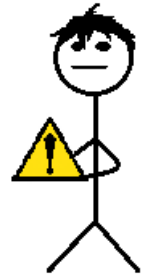
# Résultat

## Filtre sur le terrorisme



Journal	Date	Type_article	Theme	Titre	Contenu
Le Figaro	2018-01-24	En_ce_moment	Notre-Dame-des-Landes	NDDL: après l'abandon du projet, quel avenir pour les zadiste...	Édouard Philippe a annoncé, ce mercredi, l'abandon du p...
Le Figaro	2018-01-24	En_ce_moment	Trump enchaîne les polémiques	«Pays de merde» : la grossièreté de Donald Trump suscite un t...	VIDÉO - Jeudi, le président des États-Unis aurait qualifié
Le Figaro	2018-01-24	En_ce_moment	Affaire Lactalis	Lait infantile contaminé : fin des perquisitions chez Lactalis	VIDÉO - Des investigations ont eu lieu mercredi, dans le
Le Figaro	2018-01-24	Gros_titre	Terrorisme	Procès de Jawad Bendaoud: qui sont les deux autres prévenus?	FOCUS - Du 24 janvier au 14 février, trois hommes seront ju...
Le Figaro	2018-01-24	Gros_titre	Terrorisme	Jawad Bendaoud devant la justice : un procès singulier	VIDÉO - Ce mercredi s'ouvre le très attendu procès du «log...
Le Figaro	2018-01-24	Gros_titre	Terrorisme	Comment Jawad Bendaoud devint la risée des réseaux sociaux	Alors que la France vivait dans la terreur des attentats djiha...
Le Figaro	2018-01-24	Gros_titre	Terrorisme	Comment Jawad Bendaoud devint la risée des réseaux sociaux	Alors que la France vivait dans la terreur des attentats dji...
Le Figaro	2018-01-24	Gros_titre	Politique France	À Bordeaux, la Chambre régionale des comptes épingle la ge...	LE SCAN POLITIQUE - Le maire aurait masqué l'endettem...
Le Figaro	2018-01-24	Gros_titre	Macron	Macron arrive à Davos en terrain conquis et dans un ciel lumi...	Le président de la République s'exprime en fin d'après-m...
Le Figaro	2018-01-24	Gros_titre	Davos	Davos : dans les coulisses du rendez-vous annuel du gotha du...	ENQUÊTE - Le Forum économique mondial s'est ouvert r...
Le Figaro	2018-01-24	Gros_titre	Davos	Qu'est-ce que le forum de Davos ?	Réaction (22)La 48e édition du Forum rassemble dans la
Le Figaro	2018-01-24	Gros_titre	Politique France	Révision constitutionnelle : Larcher confirme ses réserves et fa...	LE SCAN POLITIQUE - Le président du Sénat s'oppose à l...

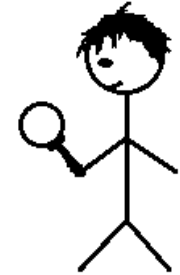
# Législation\*



- **Droit pénal** : condamne cette technique (*cinq ans d'emprisonnement et 150 000 € d'amende*)
- **Droit de la concurrence** :  
l'utilisation des données parasitées est réprimée et non la pratique en elle-même
- **Droit de la propriété intellectuelle**  
l'utilisation, sans modification substantielle, des données scrapées est sanctionnée

=> /robots.txt : fichier pour restreindre les robots d'indexation des moteurs

# Ce qu'il faut retenir



Affichage du code source : Ctrl + Maj + I ou F12

Récupération de la donnée :

- *lecture de la page*: **`read_html()`** ;
- *récupération des parties contenant les données* : **`html_nodes()`** ;
- « mise en forme » / ciblage d'éléments : **`html_text()`**, **`html_attr()`**, **`html_table...`**

*Navigation :*

- session : **`html_session()`**
- suivre un lien : **`jump_to()`** ou **`follow_link()`**
- historique : **`session_history()`**

# MERCI DE VOTRE ATTENTION

