

# A PROPOS DE VOTRE SERVITEUR



**Ettore Rizza**

@Ettore\_Rizza

Researcher [@Belspo](#) & PhD student in Information Sciences & Technologies [@ULBruxelles](#). Former journalist. [@OpenRefine](#) supporter. Parle français et italiano.

📍 Brussels, Belgium

✉️ Inscrit en mars 2009

# **UNE INTRODUCTION AMBITIEUSE AU « DATA JOURNALISME »**

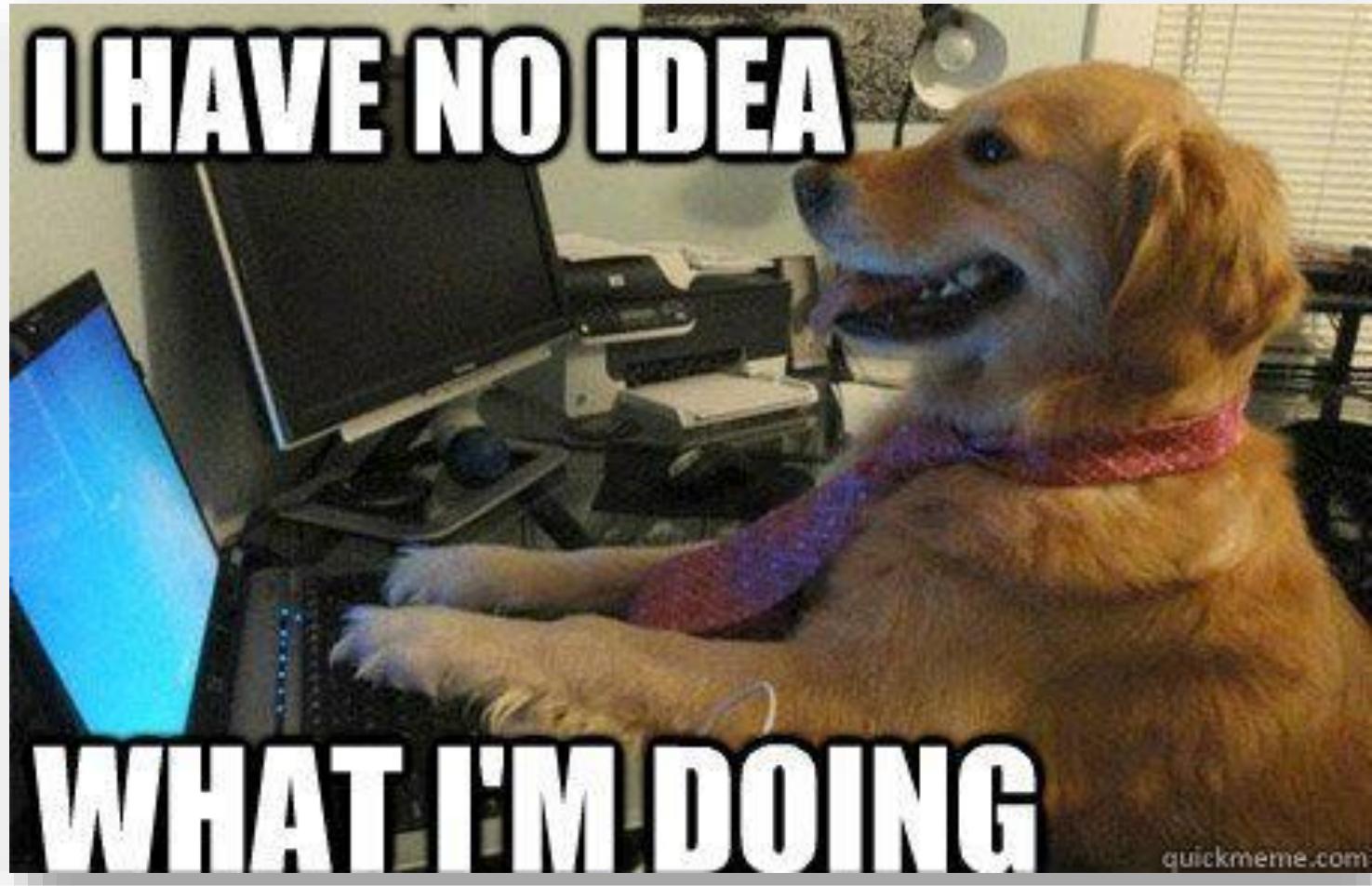
**Ettore Rizza**

**2**

# AMBITIEUSE

- Objectif : tester plusieurs outils **stables** et de **haut niveau**
- Formation « **tapas** »
- Prendre **confiance** en soi (paradoxal)
- Envisager des pistes de **formations**
- **Problème : nous irons (parfois trop) vite**

VOUS À CERTAINS MOMENT AUJOURD'HUI...



quickmeme.com

# « DATA JOURNALISME »

- <http://www.ohmybox.info/datajournalisme.html>

The screenshot displays a web page with two main sections:

- 1952 PRÉ-HISTOIRE**: A green header bar at the top has a navigation menu with items 0, 1950, 1960, 1970, 1980, and 1990. Below it, a circular icon shows a vintage computer terminal. To the right, a text box says "4 novembre 1952." and describes the use of a Remington Univac computer by CBS to predict the 1952 US presidential election. It includes a small video thumbnail titled "Election 1952: Univac helps CBS predict the Winner".
- 1960 LES SCIENCES SOCIALES**: A green header bar at the top has a navigation menu with items 0, 1950, 1960, 1970, 1980, and 1990. Below it, a text box discusses Philip Meyer's work in Detroit in 1967, using IBM 360 mainframe to analyze data from the Detroit Free Press. It includes two circular icons showing a person working at a desk and a group of people. To the right, a text box says "1968" and describes "THE NON-RIOTERS: A HOPEFUL MAJORITY".

# « DATA JOURNALISME » : EXEMPLES

- Pulitzer classique :  
<https://www.flickr.com/photos/juggernautco/sets/72157607210036175/>
- Analyse : <https://www.ft.com/content/62d782d6-31a7-11e7-9555-23ef563ecf9a>
- Pure « dataviz » : <https://www.hindustantimes.com/static/olympics/every-country-fastest-man-in-one-race-100m/>
- Enquête web : <https://www.buzzfeednews.com/article/johntemplon/how-we-used-data-to-investigate-match-fixing-in-tennis#.bjpMp0Rpw>
- Mix de texte-viz : <https://pudding.cool/2018/07/women-in-parliament/>

# « DATA JOURNALISME »

**Stephen « Steve » Doig**



**Objectif :** produire un bon article

**Nicolas Kayser-Bril**



**Objectif :** produire une bonne application



## Steve Doig

ASU's Cronkite School, Journalism professor (US)

[READ FULL BIO](#)



## Simon Rogers

Google, Data journalist (GB)

[READ FULL BIO](#)



## Nicolas Kayser-Bril

Journalism++, Data-driven journalist (FR)

[READ FULL BIO](#)

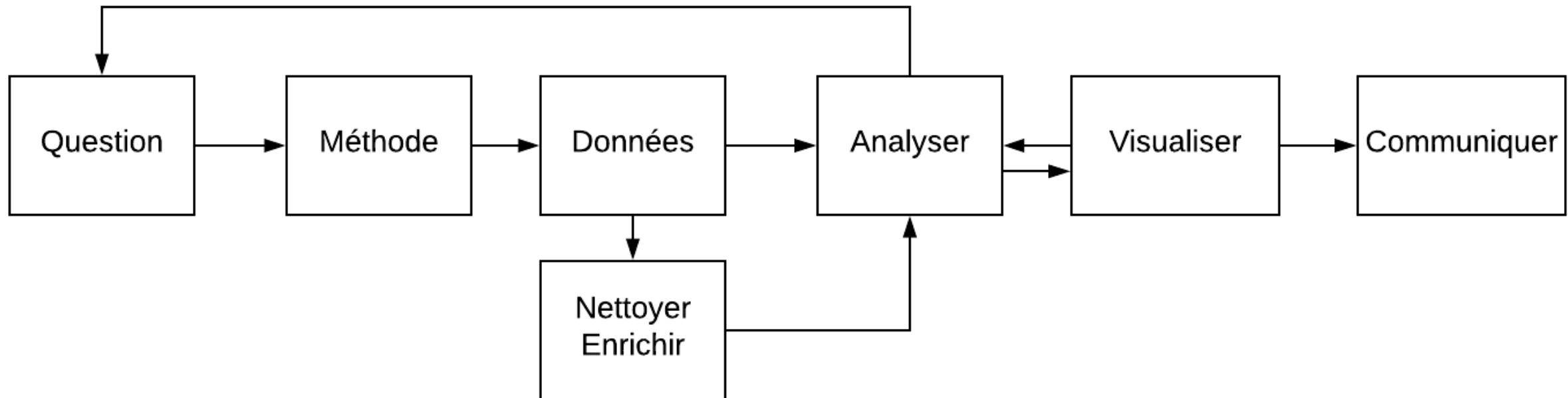
# « DATA JOURNALISME »

- <http://learno.net/courses/doing-journalism-with-data-first-steps-skills-and-tools>

# **MA VISION PERSONNELLE**

**Utiliser des méthodes informatiques pour automatiser des tâches fastidieuses et utiliser des données fiables pour s'approcher le plus rigoureusement possible d'une certaine vérité.**

# MÉTHODOLOGIE STANDARD



# I

# Exposé de la situation

AU LENDEMAIN DU SCRUTIN, VOTRE CHEF  
VIENT VOUS TROUVER...



- [http://bru2012.irisnet.be/fr/com/preferred/preferred\\_CGM21018\\_172.html](http://bru2012.irisnet.be/fr/com/preferred/preferred_CGM21018_172.html)



| COMMUNE DE WOLUWE-SAINT-LAMBERT    |                                     |                      |        |                                  |            |     |          |  |  |
|------------------------------------|-------------------------------------|----------------------|--------|----------------------------------|------------|-----|----------|--|--|
| Choisissez une commune             |                                     |                      |        | Vers les résultats               | Impression |     |          |  |  |
| MR                                 |                                     | Choisissez une liste |        |                                  |            |     |          |  |  |
| Mise à jour à: 2012/10/14 20:34:34 |                                     |                      |        | Statut du dépouillement: complet |            |     |          |  |  |
| Nr                                 | Titulaires                          | Votes                | %      | I                                | II         | III | IV       |  |  |
| 1                                  | <b>VAN GOIDSENHOVEN-BOLLE Julie</b> | 869                  | 12,45% | 2.432                            | 3.301      | 952 | <b>1</b> |  |  |
| 2                                  | <b>VANDERWAEREN Pierre</b>          | 257                  | 3,68%  | 952                              | 1.209      | 0   | <b>2</b> |  |  |
| 3                                  | VERHAEGHE Laurence                  | 211                  | 3,02%  | 0                                | 211        | 0   | 1        |  |  |
| 4                                  | <b>GEELHAND Philippe</b>            | 321                  | 4,60%  | 0                                | 321        | 0   | <b>5</b> |  |  |
| 5                                  | RIABICHEFF Nathalie                 | 145                  | 2,08%  | 0                                | 145        | 0   | 2        |  |  |
| 6                                  | <b>de HARENNE Henry</b>             | 285                  | 4,08%  | 0                                | 285        | 0   | <b>6</b> |  |  |
| 7                                  | PANS Amélie                         | 190                  | 2,72%  | 0                                | 190        | 0   | 5        |  |  |
| 8                                  | HAUBEN Alexandre                    | 95                   | 1,36%  | 0                                | 95         | 0   | 28       |  |  |
| 9                                  | DE MEULENAERE Claudine              | 128                  | 1,83%  | 0                                | 128        | 0   | 13       |  |  |
| 10                                 | <b>DEROUBAIX Emmanuel</b>           | 327                  | 4,68%  | 0                                | 327        | 0   | <b>4</b> |  |  |
| 11                                 | GAUDISSERT Claire                   | 184                  | 2,64%  | 0                                | 184        | 0   | 6        |  |  |
| 12                                 | PEETERS Christian                   | 140                  | 2,01%  | 0                                | 140        | 0   | 11       |  |  |
| 13                                 | de MEEUS d'ARGENTEUIL Chantal       | 198                  | 2,84%  | 0                                | 198        | 0   | 3        |  |  |
| 14                                 | DUMONCEAU Geoffroy                  | 175                  | 2,51%  | 0                                | 175        | 0   | 8        |  |  |
| 15                                 | JACOB Manon                         | 125                  | 1,79%  | 0                                | 125        | 0   | 14       |  |  |
| 16                                 | POLI Marco                          | 89                   | 1,27%  | 0                                | 89         | 0   | 30       |  |  |
| 17                                 | SMTOTS Céline                       | 107                  | 1,53%  | 0                                | 107        | 0   | 22       |  |  |

# VOTRE MISSION

1. Identifier tous les candidats en Wallonie et à Bruxelles élus malgré leur mauvaise place sur la liste (les « outsiders »)
2. Identifier les dix cas les plus emblématiques (selon un critère à définir)
3. Faire un petit graphique (pour le journal papier)
4. Ecrire quelque chose comme 2500 signes comme article d'ensemble



FAISABLE ?

# PAS À LA MAIN EN TOUT CAS...

- Il y a **281 communes** en Fédération Wallonie-Bruxelles (262 en Wallonie et 19 à BXL)
- Et une moyenne de **4 listes par commune**
- Cela fait entre 1100 et 1200 pages web de résultats à examiner manuellement
- A raison d'une minute par page (!), il faudrait environ **20h de travail non stop...** juste pour obtenir les données brutes \o/

# TROIS SOLUTIONS POSSIBLES

1. Avouer que vous n'êtes pas vraiment le roi/la reine de la data... (lol) (^\_^)  
(désolé)
2. PANIQUER ! :o
3. FAIRE N'IMPORTE QUOI 
4. Faire ce que le chef demande ^\\_(ツ)\_/^

# SOLUTION 3

1. Question : on l'a plus ou moins (à affiner)
2. Méthodologie : pas encore clair
3. Données : les résultats officiels sur les deux sites web, qu'il faudra récupérer
4. Nettoyage : OpenRefine
5. Analyse : R
6. Cartographie : QGis



## CONSIGNES

Dès que je dis « **regardez !** »,  
merci de **lever les yeux** pour  
**regarder** ce que je fais. ;)



Quand je dis « **c'est à vous** »,  
vous pouvez commencer à  
**reproduire ce que j'ai montré**



MAKE GIFS AT GFSOUP.COM

# MENU DU JOUR

## Matin

- ❖ 9h30: Préambule et Introduction
- ❖ Vers 10h: Web scraping sur son lit de programmation
- ❖ 10h30: Pause-café
- ❖ 10h45: Suite Web scraping
- ❖ 12h30: Lunch

## Après-midi

- ❖ 14h: Nettoyage de donnée et filer de début d'analyse
- ❖ 15h15: Pause café
- ❖ 15h30: Fines Analyses dans R
- ❖ ~~16h45 Cartographie QGis~~
- ❖ 17h25: Conclusion
- ❖ 17h30: Fin de la formation

# **II**

# **Web Scraping**

## WEB SCRAPING : DÉFINITION

- Processus qui consiste à convertir automatiquement des ressources présentes sur le web en un **format structuré**.
- Un script informatique parcourt une série de pages web et extrait certains éléments.
- Le résultat est le plus souvent renvoyé sous forme de **tableaux**.

# WEB SCRAPING : GO !

- Résultats BXL : [http://bru2012.irisnet.be/fr/com/results/results\\_start.html](http://bru2012.irisnet.be/fr/com/results/results_start.html)
- Résultats Wallonie : [http://electionslocales.wallonie.be/2012/fr/com/results/results\\_start.html](http://electionslocales.wallonie.be/2012/fr/com/results/results_start.html)

(Placez ces deux adresses en favoris, on en aura souvent besoin)

**COMMUNALES**

Vous êtes ici : [Home](#) > [Elections 2012](#) > [Conseil Communal](#)

■ COMMUNALE : PAGE DE DÉPART  
Cette page donne une vue d'ensemble de toutes les entités dans les élections courantes, avec un lien direct à ses résultats. L'icône à côté d'une entité indique le statut de dépouillement :  
■ = Complet ■ = Partiel □ = Sans résultats [En savoir plus.](#)  
[Une vue géographique du statut de dépouillement.](#)

■ COMMUNALES

Mise à jour à 2012/10/14 23:38:39

■ Communes

|                                       |                                      |                                       |
|---------------------------------------|--------------------------------------|---------------------------------------|
| <a href="#">ANDERLECHT</a>            | <a href="#">GANSHOREN</a>            | <a href="#">SAINT-JOSSE-TEN-NODE</a>  |
| <a href="#">AUDERGHEM</a>             | <a href="#">IXELLES</a>              | <a href="#">SCHAERBEEK</a> ★★         |
| <a href="#">BERCHEM-SAINTE-AGATHE</a> | <a href="#">JETTE</a>                | <a href="#">UCCLE</a>                 |
| <a href="#">BRUXELLES</a>             | <a href="#">KOEKELBERG</a>           | <a href="#">WATERMAEL-BOTSFORT</a>    |
| <a href="#">ETTERBEEK</a>             | <a href="#">MOLENBEEK-SAINT-JEAN</a> | <a href="#">WOLUWE-SAINT-LAMBERT</a>  |
| <a href="#">EVERE</a> ★★              | <a href="#">SAINT-GILLES</a>         | <a href="#">WOLUWE-SAINT-PIERRE</a> * |
| <a href="#">FOREST</a>                |                                      |                                       |

[rafrâicher la page](#)

**PROVINCIALES** **COMMUNALES** **CPAS**

Vous êtes ici : [Home](#) > [Elections 2012](#) > [Conseil Communal](#)

■ COMMUNALE : PAGE DE DÉPART  
Cette page donne une vue d'ensemble de toutes les entités dans les élections courantes, avec un lien direct à ses résultats. L'icône à côté d'une entité indique le statut de dépouillement :  
■ = Complet ■ = Partiel □ = Sans résultats [En savoir plus.](#)  
[Une vue géographique du statut de dépouillement.](#)

■ COMMUNALES

Mise à jour à 2013/03/04 14:11:28

■ Communes

|                                |                                      |                             |
|--------------------------------|--------------------------------------|-----------------------------|
| <a href="#">AISEAU-PRESLES</a> | <a href="#">FERRIÈRES</a>            | <a href="#">NAMUR</a>       |
| <a href="#">AMAY</a>           | <a href="#">EXHE-LE-HAUT-CLOCHER</a> | <a href="#">NANDRIN</a>     |
| <a href="#">AMBLÈVE</a>        | <a href="#">FLEURUS</a>              | <a href="#">NASSOGNE</a>    |
| <a href="#">ANDENNE</a>        | <a href="#">FLOBECQ</a>              | <a href="#">NEUFCHÂTEAU</a> |
| <a href="#">ANDERLUES</a>      | <a href="#">FLOREFFE</a>             | <a href="#">NEUPRÉ</a>      |
| <a href="#">ANHÉE</a>          | <a href="#">FLORENNES</a>            | <a href="#">NIVELLES</a>    |
| <a href="#">ANS</a>            | <a href="#">FLORENVILLE</a>          | <a href="#">OHEY</a>        |

# MINI-EXERCICE

1. Examinez les URL vers les pages des communes et des listes.
2. Pourriez-vous les reconstituer sans les connaître ? Sur base de quoi ?

# QUE DEVRAIT-ON FAIRE ?

1. Récupérer les **URLS** des listes (ou naviguer vers elles)
2. Récupérer le **code source** des pages contenant les résultats
3. Extraire un **élément** de ce code (la table en HTML, avec toutes ses balises)
4. Extraire le texte de cette table

# QUESTIONS

- Comment récolter les URL ?
- Comment sélectionner des éléments dans une page Web ?

# COMMENT RÉCUPÉRER DES URIS ?

- Les récupérer dans une page (simple)
- Suivre des liens « next » ou des paginations
- Recréer soi-même les URLs (<https://www.neodownloader.com/tools/link-builder/>)

Exemple : <http://example.webscraping.com/places/default/index/0>

# COMMENT SÉLECTIONNER DES ÉLÉMENTS DANS UNE PAGE WEB ?

- **XPath** (puissant, mais un peu compliqué)
- **Sélecteurs CSS** (ce qu'il nous faut aujourd'hui)
- Au besoin : **expressions régulières** (pour extraire du texte)

# LES SÉLECTEURS CSS

- <https://www.w3schools.com/code/tryit.asp?filename=FTXZ66BI3AEH>
- <https://www.w3schools.com/cssref/trysel.asp>

# WEB SCRAPING : EXERCICES WEB SCRAPER CHROME

- **Exercice 1 :** à l'aide du web scraper chrome, récupérez toutes les listes qui se présentent à Bruxelles (5 minutes + temps de scraping)

# WEB SCRAPING : EXERCICES RVEST

1. **Echauffement 1** : à partir d'une liste au hasard, récupérez le nom des **élus** (5 minutes max)
2. **Echauffement 2** : à partir d'une liste au hasard, récupérez toute la table des résultats (5 minutes max)
3. **Echauffement 3** : récrivez le code précédent avec des %>% (5 minutes max)
4. **Exercice final (choses sérieuses)** : récupérez (dans un seul fichier) tous les résultats pour toutes les listes à Bruxelles

# WEB SCRAPING : TUTOS ET RESSOURCES

## ***Web scraper Chrome***

- <https://www.youtube.com/watch?v=-cxNhoVufEo>
- <https://www.webscraper.io/tutorials>

## ***Rvest***

- <https://www.analyticsvidhya.com/blog/2017/03/beginners-guide-on-web-scraping-in-r-using-rvest-with-hands-on-knowledge/>
- <https://github.com/ropensci-training/user2016-tutorial/blob/master/03-scraping-data-without-an-api.pdf>

# WEB SCRAPING : QUELQUES REMARQUES

- Parfois, pas besoin de scraper (APIs) (voir slides en annexes)
- Attention aux sites web dynamiques bourrés de Javascript (Twitter)
- Attention aux iFrames (vieux sites institutionnels)
- Attention à la légalité (*Terms of use, robots.txt...*)
- Attention aux serveurs (*fair use*)
- Un outil tout fait est souvent indispensable (Outwit Hub, Octoparse, Helium scraper...)

# WEB SCRAPING : APERÇU OUTWIT HUB

OutWit Hub Pro

File Edit View Navigation Tools Help Registration

www.lachambre.be/kvcr/showpage.cfm?section=/depute&language=fr&cfm=/site/ URL C Google

Actuels | 54 (2014-2016) | 53 (2010-2014) | 52 (2007-2010) | 51 (2003-2007) | 50 (1999-2003) | 49 (1995-1999) | 48 (1991-1995)

Par nom | Par groupe politique

|                        |             |  |                             |
|------------------------|-------------|--|-----------------------------|
| <u>Almaci Meyrem</u>   | Ecolo-Groen | <a href="mailto:meyrem.almaci@dekamer.be">meyrem.almaci@dekamer.be</a>       | <a href="#">Site web(*)</a> |
| <u>Becq Sonja</u>      | CD&V        | <a href="mailto:sonja.becq@dekamer.be">sonja.becq@dekamer.be</a>             | <a href="#">Site web(*)</a> |
| <u>Beke Wouter</u>     | CD&V        | <a href="mailto:wouter.beke@dekamer.be">wouter.beke@dekamer.be</a>           | <a href="#">Site web(*)</a> |
| <u>Bellens Rita</u>    | N-VA        | <a href="mailto:rita.bellens@dekamer.be">rita.bellens@dekamer.be</a>         |                             |
| <u>Ben Hamou Nawal</u> | PS          | <a href="mailto:nawal.benhamou@lachambre.be">nawal.benhamou@lachambre.be</a> | <a href="#">Site web(*)</a> |

Found 357 links.  
Found 150 HTML table rows.  
Page loaded: La Chambre des représentants de Belgique - (150 HTML table rows)  
--- Loading URL: http://www.lachambre.be/kvcr/showpage.cfm?section=/depute&language=fr&cfm=/site/  
Updated tables: 29 rows.  
Page loaded: La Chambre des représentants de Belgique - (150 HTML table rows)  
Found 14 HTML table rows.  
Found 79 links.  
Page loaded: La Chambre des représentants de Belgique - (150 HTML table rows)  
--- Loading URL: http://www.lachambre.be/kvcr/showpage.cfm?section=/depute&language=fr&cfm=/site/  
Found 45 links.  
Page loaded: OutWit - Harvest The Web  
--- Loading URL: http://www.outwit.com/?  
Displayed Macro Manager.  
Starting OutWit Hub Pro (Runner version: 2.1.1.1000)

Séances plénaires TOP VIDEO

Local IP: 94.225.17.142 Remote IP: 193.191.129.46

La Chambre des représentants de Belgique - (150 HTML table rows)

| Url  | Column 4           | Column 5    | Column 6                    |
|--|--------------------|-------------|-----------------------------|
| http://www.lachambre.be/kvcr/showpage.cfm?section=/depute&language=fr&cfm=/site/n?section=/depute&languag... | Almaci Meyrem      | Ecolo-Groen | eb.remaked@icamla.meryer... |
| http://www.lachambre.be/kvcr/showpage.cfm?section=/depute&language=fr&cfm=/site/n?section=/depute&languag... | Becq Sonja         | CD&V        | eb.remaked@qceb.ajnos...    |
| http://www.lachambre.be/kvcr/showpage.cfm?section=/depute&language=fr&cfm=/site/n?section=/depute&languag... | Beke Wouter        | CD&V        | eb.remaked@ekeb.retuow...   |
| http://www.lachambre.be/kvcr/showpage.cfm?section=/depute&language=fr&cfm=/site/n?section=/depute&languag... | Bellens Rita       | N-VA        | eb.remaked@snelleb.atir...  |
| http://www.lachambre.be/kvcr/showpage.cfm?section=/depute&language=fr&cfm=/site/n?section=/depute&languag... | Ben Hamou Nawal    | PS          | eb.erbmahcal@uomahneb.l...  |
| http://www.lachambre.be/kvcr/showpage.cfm?section=/depute&language=fr&cfm=/site/n?section=/depute&languag... | Blanchart Philippe | PS          | eb.erbmahcal@trahcnalb.e... |

Select row if any column contains Limit to Options

0  Clean Text  Deduplicate Catch Auto-Empty

EXPORT

34

**III**

# **Nettoyage/Enrichissement**

# EXEMPLE DE DONNÉES « SALES »

| <b>Nom</b>                   | <b>Date_naissance</b> | <b>Date_décès</b> | <b>Pays</b>    |
|------------------------------|-----------------------|-------------------|----------------|
| Jean DUPONT                  | 29/11/03              | 10/8/45           | Belgique       |
| MAILLARD Aline               | 1962-01-01            | 2010-12-05        | Milan (Italie) |
| Dupont Jean                  | 29 novembre 1903      | 10 août 1945      | Belgium        |
| Assia JELLAL;20 septembre 63 | NA                    | Blegique, Maroc   |                |
| Jean Dupont (BEL)            | 29/11/1903            | 10/08/1945        | Belgique       |



# A NE PAS SOUS-ESTIMER

- Article du NY Times : <https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html>

“Data scientists, according to interviews and expert estimates, spend from 50 percent to 80 percent of their time mired in this more mundane labor of collecting and preparing unruly digital data, before it can be explored for useful nuggets.”



# OPEN REFINE



- <http://openrefine.org/download.html>
- Anciennement Google Refine
- Open Source et gratuit
- Une sorte d'Excel spécialisé en **texte**
- Permet de le nettoyer, de le transformer, de l'enrichir...
- Couteau-suisse du nettoyage de données textuelles
- Très utilisé par les data journalistes
- N'est plus mis à jour aussi souvent que durant l'époque Google
- Mais en pleine activité pour le moment



# OPEN REFINE : OPÉRATIONS

(N'ESSAYEZ PAS DE LIRE CE SLIDE)

- Facette sur la colonne « File », renommer en « Bruxelles » et « Wallonie »
- Facette sur la colonne « Commune » : combien y en-a-t 'il ?
- Text length facet sur les noms
- Facette sur « position\_elu » et tri par nombre : rien de bizarre ?
- Pour que ce soit plus clair, créer une colonne « élu » (oui/non) : if(isNonBlank(value), "oui", "non")
- Enlever les colonnes inutiles (voix dévolues, etc.)
- Clustering sur nom de parti
- Créer un Id pour les listes (concaténation commune-parti)
- Ajouter provinces : <https://goo.gl/ZeHzwf>
- Extraire le code INS depuis les URLs (import re;return re.findall("CGM(.+)\_", value)[0])
- **Aperçu de Wikidata et des plugins Vib-Bits ?**
- Question : qui est le ou la candidate le plus intéressant sur la liste PS d'Aiseau-Prêle ? Comment pourrait-on l'identifier par calcul ?  
[http://electionslocales.wallonie.be/2012/fr/com/preferred/preferred\\_CGM52074\\_3.html](http://electionslocales.wallonie.be/2012/fr/com/preferred/preferred_CGM52074_3.html)
- Commencer à réfléchir à la manière de trouver nos « outsiders ». Le nettoyage est une bonne occasion de manipuler les données.
- La simple différence ne suffira pas (dans l' exemple précédent, le 10<sup>e</sup> à une différence supérieure au 14<sup>e</sup>) (note : Didier Demars, 16, aurait pu être élu à 25 voix près. Intéressant pour d'autres articles sur le thème « ils ont raté l'élection d'un cheveu »)
- Transformer place\_liste et position\_elu en nombres (verts), puis soustraire le premier du second
- Tiens, n'y aurait-il pas un papier à faire sur les élus en trois premières place les moins « bien élus » ? (Catherine Mourreau) : facette place\_liste (1,2,3), facette sur place\_positionelu négatifs)



# IV

# Analyse

# MÉTHODE PROPOSÉES

- Exclure les trois derniers de chaque liste (*outliers*)
- Identifier le « dernier du groupe de tête » ?
- Calculer pour chaque élu sa « distance du groupe de tête » ?
- Retenir les candidats élus dont la distance et l'écart entre leur place sur la liste et la place d'élue (déjà calculé) se situe sous un certain seuil (à définir)
- Créer un graphique de leur nombre par commune
- Isoler les dix premiers



- Langage de programmation spécialisé en statistiques (mais pas que)
- Open Source et gratuit
- À utiliser de préférence avec Rstudio





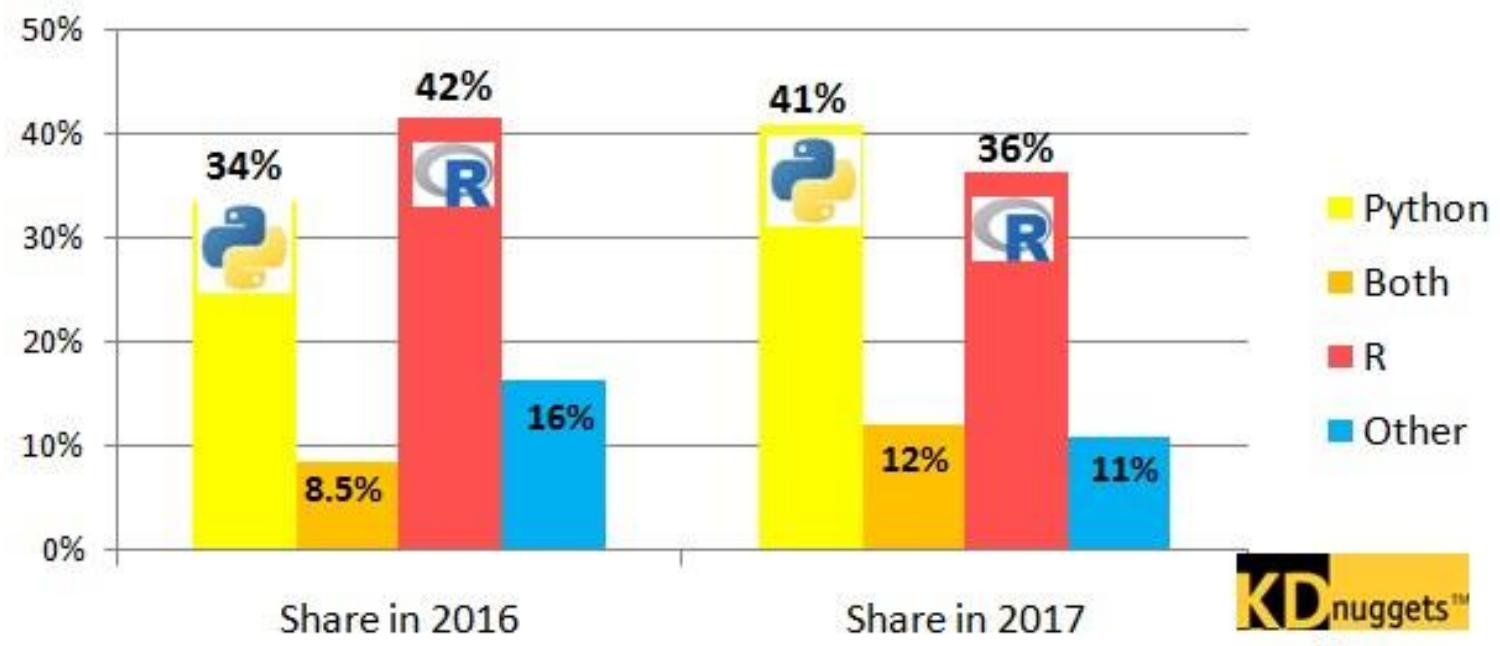
# AVANTAGES SUR EXCEL

- Reproductibilité par d'autres... et par votre futur vous !
- Automatisation
- Débogage plus facile
- Partage de code (mail, StackOverflow, Github... )
- Beaucoup plus puissant et rapide
- Plus de 10 000 packages pour toutes sortes de tâches
- Graphiques nettement plus variés et plus beaux
- Possibilité d'effectuer des analyses infiniment plus poussées
- Parfois plus clair qu'Excel
- **Nécessite toutefois un apprentissage plus long**
- Ne pas hésiter à utiliser des addins et gadgets :  
`devtools::install_github("dkilfoyle/rpivotGadget")`



# R VERSUS PYTHON

Python, R, Both, or Other platforms for  
Analytics, Data Science, Machine Learning





# PRIVILEGIER LE « TIDYVERSE »



- Nous utiliserons le « super package » *Tidyverse* - dont fait partie rvest
- Pour l'analyse, nous utiliserons *Dplyr* (l'un de ses packages)
- Ce package d'analyse de données se fonde sur des verbes :
  - `select()`      sélectionne des colonnes
  - `filter()`      filtre des lignes selon certains critères
  - `arrange()`     trie les lignes dans un certain ordre
  - `mutate()`     crée de nouvelles colonnes
  - `summarise()` résumé d'une colonne (moyenne, somme, etc...)
  - `group_by()`   groupe les valeurs selon certains critères





# PACKAGE GGPILOT2



- Permet de faire des graphiques superbés.
- Mais difficile à maîtriser...
- Certains gadgets RStudio facilitent la tâche (ggplot Theme Assistant...)

```
ggplot(table_partis_province,  
       aes(x=reorder(parti_propre, -nombre), y = nombre, fill=province, label=nombre)) +  
  geom_bar(stat="identity") +  
  theme_minimal() +  
  labs(x = "Parti", y = "Nombre") +  
  geom_text(size = 3, position = position_stack(vjust = 0.5))
```





# RESSOURCES ADDITIONNELLES

- <http://learn.r-journalism.com/en/>
- <https://journalismcourses.org/RC0818.html>
- <https://learno.net/courses/r-for-journalists>
- <https://rddj.info/>
- <https://www.datacamp.com/community/open-courses>
- <https://www.rstudio.com/resources/cheatsheets/>
- <https://cran.r-project.org/web/packages/dplyr/vignettes/dplyr.html>



# Conclusion

# NOUS AVONS BIEN BOSSÉ, MAIS...

- Nous aurions pu voir encore énormément de choses (SQL, lignes de commandes, APIs, expressions régulières, Github, Gephi, QGis...)
- Nous avons juste survolé les outils et techniques (R et Tidyverse, Open Refine, scraping en général, programmation)...

## DONC

- N'hésitez pas à vous former de votre côté
- A demander des formations AJPro, publiques ou pour entreprise
- **A utiliser votre service après-vente : 3 questions par mail chacun(e) ! (sérieux)**
- **Et, bien sûr, à adapter les codes R dans votre dossier de cours !**



*That's all Folks!*

# Annexes

(choses non vues)



- Abréviation de Quantum Geographic Information System
  - GIS Open Source et gratuit (contrairement à ArcGIS)
  - Logiciel professionnel
  - Usine à gaz à la Photoshop
  - Que l'on peut encore enrichir à l'aide de plugins
- 
- **Nous installerons OpenLayers, MMQgis et qgis2web**







# CE QUE NOUS ALLONS FAIRE

1. Trouver une carte des communes belges :  
<https://goo.gl/ZeHzwf>
2. L'éditer et l'enrichir avec nos données dans QGis
3. L'exporter sous forme de carte interactive [Leaflet](#) (Javascript), intégrable dans une page Web, avec le plugin qgis2web



# Les APIs

# LES APIs

- *Application Programming Interface*
- Ensemble de codes et protocoles qui permettent à deux logiciels de **communiquer**
- Les APIs web utilisent des **requêtes HTTP**
- Résultats fournis au format **JSON** ou **XML**

# LES APIS

## Exemples

### Google Maps

- <http://maps.googleapis.com/maps/api/geocode/json?sensor=false&address=avenue%20franklin%20roosevelt%20bruxelles>

### Open Movie Database (OMDb)

- <http://www.omdbapi.com/?apikey=9799b14d&t=Titanic>

### Wikipedia

- <https://fr.wikipedia.org/w/api.php?action=query&format=json&list=search&utf8=1&srsearch=bruxelles>

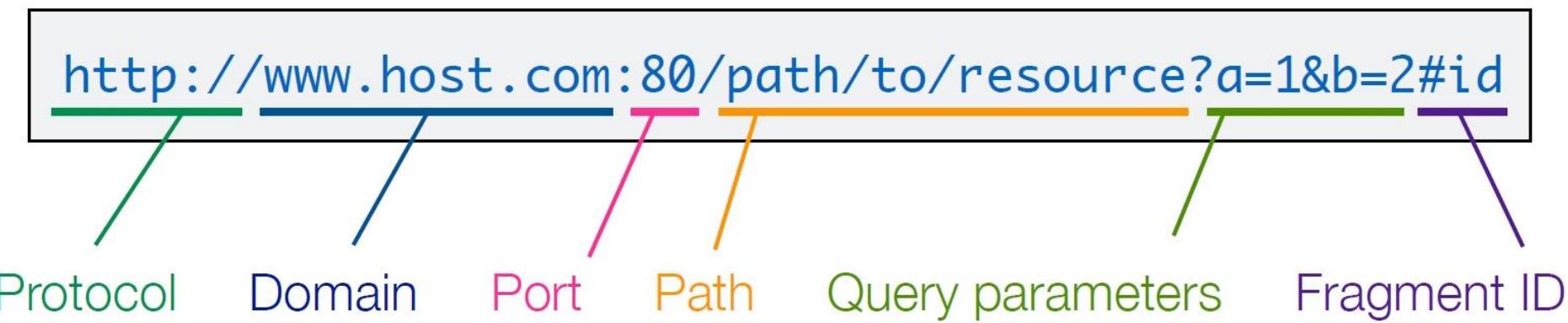
## Exemple (plus long)

### Wikipedia

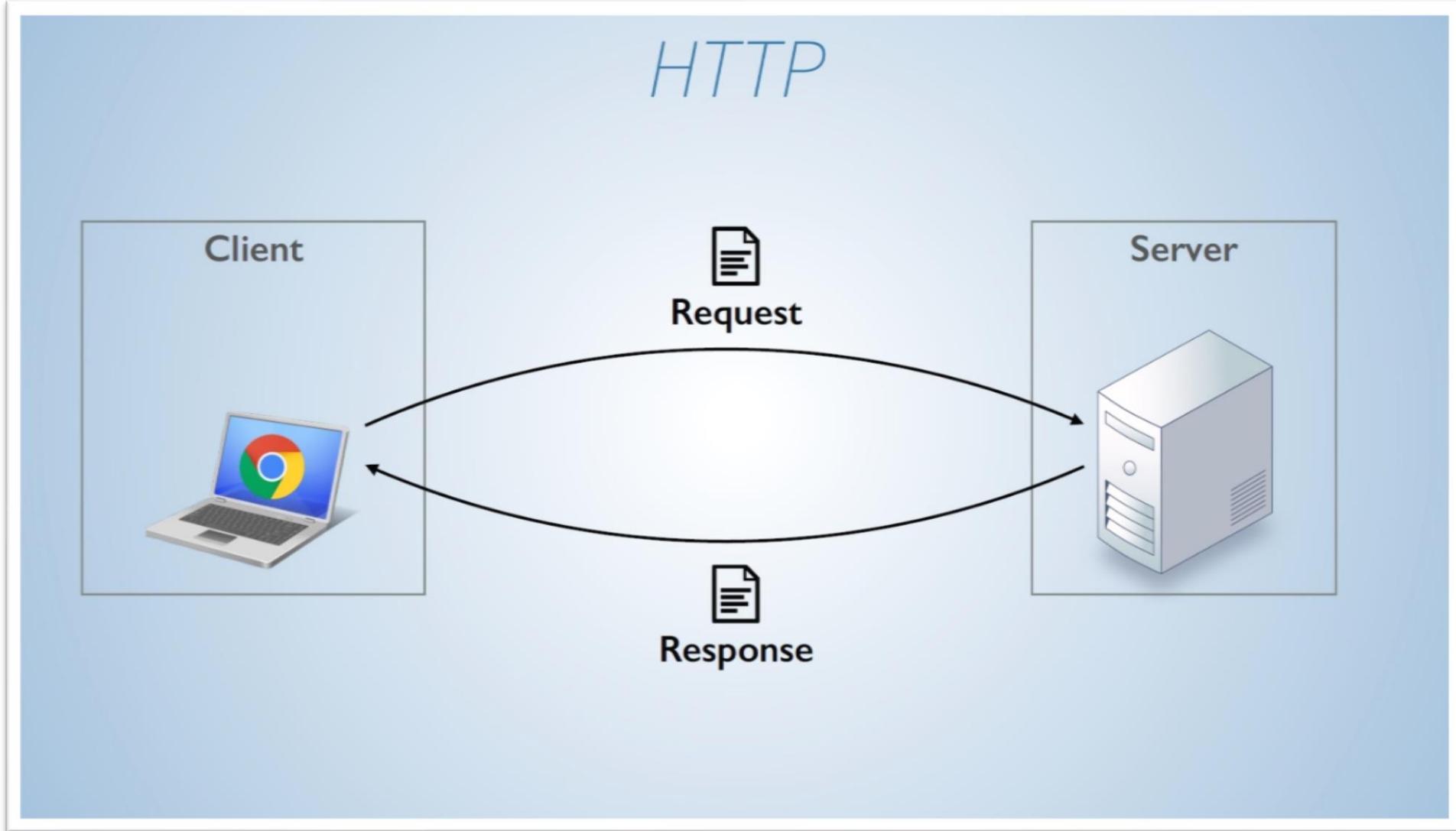
- [https://fr.wikipedia.org/w/api.php?format=json&action=query&generator=search&gsrnamespace=0&gsrsearch=einstein&gsrlimit=20&prop=categories%7Cextracts%7Cinfo%7Crevisions%7Cpageprops&inprop=url&ppprop=wikibase\\_item&rvprop=content&exintro=&explaintext=&exsentences=2&exlimit=max&cllimit=max](https://fr.wikipedia.org/w/api.php?format=json&action=query&generator=search&gsrnamespace=0&gsrsearch=einstein&gsrlimit=20&prop=categories%7Cextracts%7Cinfo%7Crevisions%7Cpageprops&inprop=url&ppprop=wikibase_item&rvprop=content&exintro=&explaintext=&exsentences=2&exlimit=max&cllimit=max)

# LES APIS

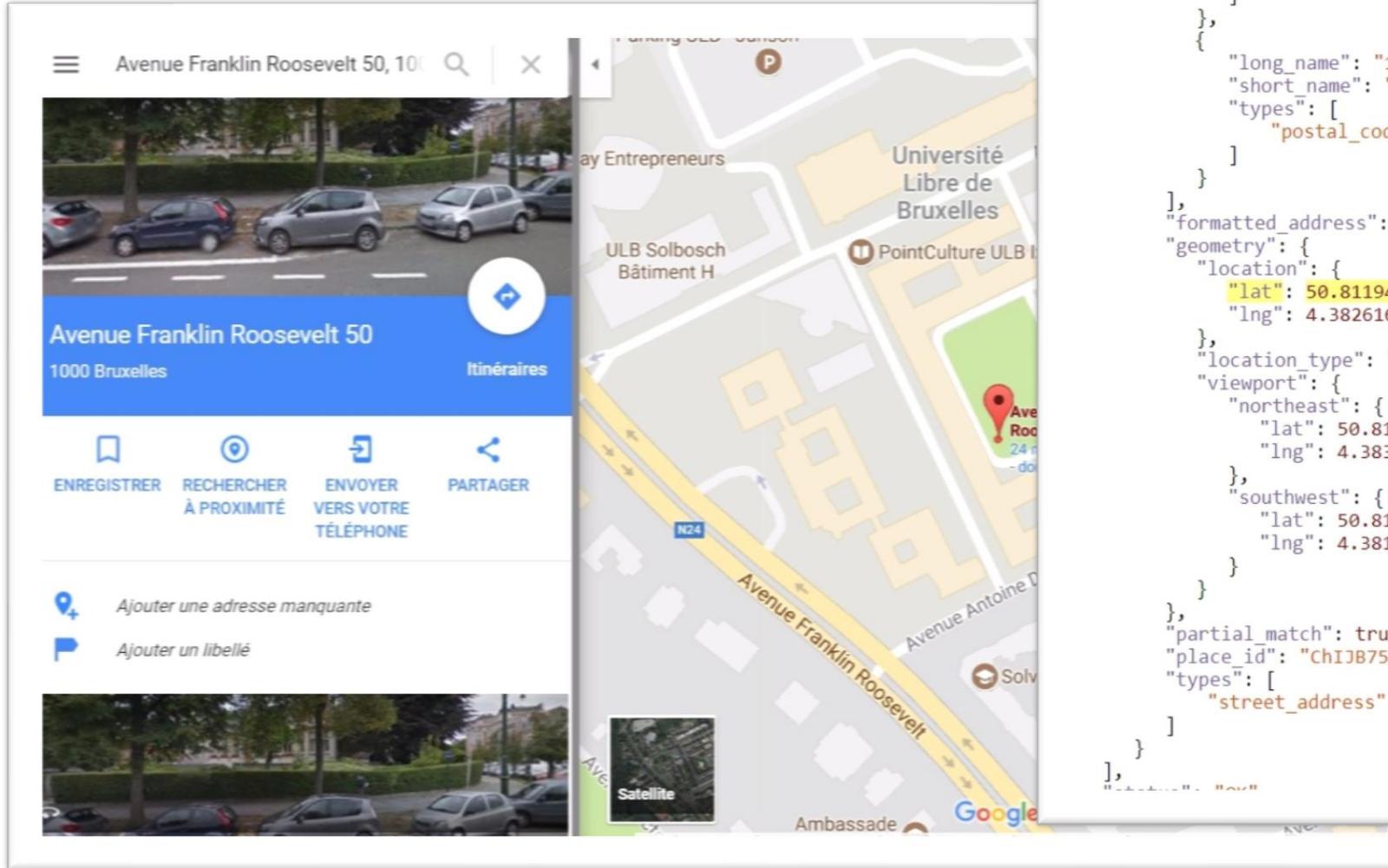
## URL



# LES APIS



# LES APIS



```
results["0"].geometry.location.lat
```

COPY PATH

```
political
],
{
  "long_name": "1000",
  "short_name": "1000",
  "types": [
    "postal_code"
  ]
},
"formatted_address": "Avenue Franklin Roosevelt 50, 1000 Bruxelles, Belgique",
"geometry": {
  "location": {
    "lat": 50.8119483,
    "lng": 4.382616899999999
  },
  "location_type": "ROOFTOP",
  "viewport": {
    "northeast": {
      "lat": 50.8132972802915,
      "lng": 4.383965880291502
    },
    "southwest": {
      "lat": 50.8105993197085,
      "lng": 4.381267919708497
    }
  },
  "partial_match": true,
  "place_id": "ChIJB75Std3Ew0cRYG39Y4jeL7I",
  "types": [
    "street_address"
  ]
},
```

# LES APIS (EXEMPLE DE SCRIPT PYTHON)

```
import requests, pprint

base_url = "http://omdbapi.com/?plot=short&apikey=9799b14d&t="

movies_list = ["Scream",
               "Titanic",
               "The revenant",
               "la Grande vadrouille"]

movies = {}

for movieTitle in movies_list:
    response = requests.get(base_url + movieTitle)
    if response.status_code == 200:
        movies[movieTitle] = response.json()
    else:
        raise ValueError("Bad request!")

pp = pprint.PrettyPrinter(indent=4)

pp.pprint(movies)
```

# LES APIs

## *Parsing des résultats*

### Démonstration avec Open Refine

The screenshot shows the Open Refine interface with a single row of data. The data table has columns: All, adresse, lien, json, long, and lat. The 'json' column displays a complex nested JSON object. A modal dialog titled "Add column based on column json" is open, prompting for a new column name ("lat") and an expression ("value.parseJson().results[0].geometry.location.lat"). The preview pane at the bottom shows the resulting value for the first row.

| All                                   | adresse   | lien | json  | long              | lat       |
|---------------------------------------|---|------|---|-------------------|-----------|
| 1. St.-Jobstraat<br>196 9300<br>Alost | http://maps.googleapis.com/maps/api/geocode/json?sensor=false&address=St.-Jobstraat+196+9300+Alost+ |      | { "results": [ { "address_components": [ { "long_name": "196", "short_name": "196", "types": [ "street_number" ] }, { "long_name": "Sint Jobstraat", "short_name": "Sint Jobstraat", "types": [ "route" ] }, { "long_name": "Aalst", "short_name": "Aalst", "types": [ "locality", "political" ] }, { "long_name": "Oost-Vlaanderen", "short_name": "OV", "types": [ "administrative_area_level_2", "political" ] }, { "long_name": "Vlaanderen" } ], "long_name": "Aalst", "short_name": "Aalst", "types": [ "area_level_2", "place", "political" ] }, { "sort_name": "9300", "location": { "lat": 50.93347798029149, "lon": 4.030133699999999 } }, "place_id": "ChIJLcJyfzgBtjwRqXWQFkIuH0" } ] } | 4.030133699999999 | 50.932129 |

Add column based on column json

New column name: lat

set to blank  store error  copy value from original column

Expression: value.parseJson().results[0].geometry.location.lat

Language: General Refine Expression Language (GREL)

No syntax error.

Preview

| row | value     |
|-----|-----------|
| 1.  | 50.932129 |

## Exercice

**Copiez-collez ces cinq adresses dans Open Refine**

Avenue de l'Héliport 22 b55 1000 Bruxelles

Avenue de Broqueville 2 1150 Woluwe-Saint-Pierre

Auguste Reyerslaan 15 b6 1030 Schaerbeek

Duivenstraat 34 8940 Wervik

Rue du Lombard 43 1000 Bruxelles

**Retrouvez leur latitude et longitude avec l'API suivante :**

<https://loc.geopunt.be/v4/Location?q=50%20avenue%20Franklin%20Roosevelt>

# **RESSOURCES COMPLÉMENTAIRES**

***Bon tuto en anglais :***

**<https://programminghistorian.org/lessons/fetch-and-parse-data-with-openrefine>**

***Quelques vidéos en français:***

**<https://www.youtube.com/watch?v=65Ai44AMROY&list=PL8iMgiAEwz6RKvqDU1AICqviiSKX-IFk1>**

# **Structure de l'information et modèles de données**

# INFORMATION NON STRUCTURÉE (TEXTE BRUT)

## Charles Michel devient le plus jeune Premier ministre belge

07/10/14 à 20:15 - Mise à jour à 21:19  
Source: Le Vif

Charles Michel va devenir le plus jeune Premier ministre de l'histoire de Belgique. A 38 ans, il succèdera à Elio Di Rupo et sera le deuxième libéral francophone à exercer cette fonction, après Paul-Emile Janson en 1937. Charles Michel baigne dans la politique depuis sa tendre enfance.

504  
Fois partagé



Lire plus tard

# INFORMATION STRUCTURÉE 1 (TABLEAU)

| All | Nom                   | début du mandat  | Fin du mandat    | Parti   |
|-----|-----------------------|------------------|------------------|---------|
| 51. | Achille Van Acker     | 23 avril 1954    | 26 juin 1958     | PSB-BSP |
| 52. | Gaston Eyskens        | 26 juin 1958     | 25 avril 1961    | PSC-CVP |
| 53. | Théo Lefèvre          | 25 avril 1961    | 28 juillet 1965  | PSC-CVP |
| 54. | Pierre Harmel         | 28 juillet 1965  | 19 mars 1966     | PSC-CVP |
| 55. | Paul Vanden Boeynants | 19 mars 1966     | 17 juillet 1968  | PSC-CVP |
| 56. | Gaston Eyskens        | 17 juillet 1968  | 26 janvier 1973  | PSC-CVP |
| 57. | Edmond Leburton       | 26 janvier 1973  | 25 avril 1974    | PSB-BSP |
| 58. | Leo Tindemans         | 25 avril 1974    | 20 octobre 1978  | CVP     |
| 59. | Paul Vanden Boeynants | 20 octobre 1978  | 3 mars 1979      | PSC     |
| 60. | Wilfried Martens      | 3 mars 1979      | 31 mars 1981     | CVP     |
| 61. | Mark Eyskens          | 31 mars 1981     | 17 décembre 1981 | CVP     |
| 62. | Wilfried Martens      | 17 décembre 1981 | 7 mars 1992      | CVP     |
| 63. | Jean-Luc Dehaene      | 7 mars 1992      | 12 juillet 1999  | CVP     |
| 64. | Guy Verhofstadt       | 12 juillet 1999  | 20 mars 2008     | VLD     |
| 65. | Yves Leterme          | 20 mars 2008     | 30 décembre 2008 | CD&V    |
| 66. | Herman Van Rompuy     | 30 décembre 2008 | 25 novembre 2009 | CD&V    |
| 67. | Yves Leterme          | 25 novembre 2009 | 6 décembre 2011  | CD&V    |
| 68. | Elio Di Rupo          | 6 décembre 2011  | 11 octobre 2014  | PS      |
| 69. | Charles Michel        | 11 octobre 2014  |                  | MR      |

# INFORMATION STRUCTURÉE 2 (XML)

```
<?xml version="1.0" encoding="UTF-8" ?>
<prime_ministers>
    <nom>Étienne de Gerlache</nom>
    <début>26 février 1831</début>
    <fin>23 mars 1831</fin>
    <parti>Catholique</parti>
</prime_ministers>
<prime_ministers>
    <nom>Joseph Lebeau</nom>
    <début>23 mars 1831</début>
    <fin>21 juillet 1831</fin>
    <parti>Libéral</parti>
</prime_ministers>
<prime_ministers>
    <nom>Félix De Müelenaere</nom>
    <début>26 juillet 1831</début>
    <fin>17 septembre 1832</fin>
    <parti>Catholique</parti>
</prime_ministers>
<prime_ministers>
```

# INFORMATION STRUCTURÉE 3 (JSON)

```
[{"prime_ministers": [ {"nom": "Étienne de Gerlache", "début": "26 février 1831", "fin": "23 mars 1831", "parti": "Catholique"}, {"nom": "Joseph Lebeau", "début": "23 mars 1831", "fin": "21 juillet 1831", "parti": "Libéral"}, {"nom": "Félix De Müelenaere", "début": "26 juillet 1831", "fin": "17 septembre 1832", "parti": "Catholique"}, {"nom": "Albert Goblet d'Alviella"}]}
```

# INFORMATION SEMI-STRUCTURÉE 1 (HTML)

WIKIPÉDIA  
L'encyclopédie libre

Article Discussion Lire Modifier Modifier le code Afficher l'historique Plus Rechercher dans Wikipédia

## Charles Michel (homme politique)

« Charles Michel » redirige ici. Pour les autres significations, voir Charles Michel (homonymie).

**Charles Michel**, né le 21 décembre 1975 à Namur, est un homme politique belge, membre du Mouvement réformateur. Premier ministre depuis le 11 octobre 2014, il est le chef de gouvernement le plus jeune de l'histoire de la Belgique<sup>2</sup>.

**Sommaire** [masquer]

- 1 Biographie
  - 1.1 Jeunesse et études
  - 1.2 Débuts en politique
  - 1.3 Ministre
  - 1.4 Premier ministre
- 2 Vie privée
- 3 Carrière politique
- 4 Mandats et fonctions<sup>[22]</sup>
- 5 Distinctions
- 6 Notes et références
- 7 Voir aussi
  - 7.1 Article connexe
  - 7.2 Liens externes

**Charles Michel**



Charles Michel en 2014.

**Fonctions**

Premier ministre belge

En fonction depuis le 11 octobre 2014 (3 ans et 15 jours)

| Monarque     | Philippe |
|--------------|----------|
| Gouvernement | Michel   |
| Législature  | 54e      |

Biographie [ modifier | modifier le code ]

La famille Michel est originaire d'Uccle (Bruxelles).<sup>3</sup> Fils de Martine et de Louis

# INFORMATION SEMI-STRUCTURÉE 2 (HTML)

```
<link rel="edit" title="Modifier" href="/w/index.php?title=Charles_Michel_(homme_politique)&action=edit" data-bbox="156 163 838 946"/>  
<link rel="apple-touch-icon" href="/static/apple-touch/wikipedia.png"/>  
<link rel="shortcut icon" href="/static/favicon/wikipedia.ico"/>  
<link rel="search" type="application/opensearchdescription+xml" href="/w/opensearch_desc.php" title="Wikipédia (fr)"/>  
<link rel="EditURI" type="application/rsd+xml" href="//fr.wikipedia.org/w/api.php?action=rsd"/>  
<link rel="license" href="//creativecommons.org/licenses/by-sa/3.0/"/>  
<link rel="canonical" href="https://fr.wikipedia.org/wiki/Charles_Michel_(homme_politique)"/>  
<link rel="dns-prefetch" href="//meta.wikimedia.org" />  
<!--[if lt IE 9]><script src="/w/load.php?debug=false&lang=fr&modules=html5shiv&only=scripts&skin=vector&sync=1"></script>  
<![endif]-->  
</head>  
<body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-subject page-Charles_Michel_homme_politique rootpage-Charles_Michel_homme_politique vector-experimental-print-styles vector-nav-directionality skin-vector action-view" data-bbox="156 368 838 418">  
    <div id="mw-page-base" data-bbox="156 418 838 468":class="noprint"></div>  
        <div id="mw-head-base" class="noprint"></div>  
        <div id="content" class="mw-body" role="main">  
            <a id="top"></a>  
  
            <div id="siteNotice" class="mw-body-content"><!-- CentralNotice --></div>  
            <div class="mw-indicators mw-body-content">  
                <div id="firstHeading" class="firstHeading" lang="fr">Charles Michel (homme politique)</h1>  
                    <div id="bodyContent" class="mw-body-content">  
                        <div id="siteSub" class="noprint">Un article de Wikipédia, l'encyclopédie libre.</div>  
                        <div id="contentSub"></div>  
                            <div id="jump-to-nav" class="mw-jump">  
                                Aller à :  
                                <a href="#mw-head">navigation</a>,  
                                <a href="#p-search">rechercher</a>  
                            </div>  
                            <div id="mw-content-text" lang="fr" dir="ltr" class="mw-content-ltr"><div class="mw-parser-output"><div class="homonymie"><a href="/wiki/Aide:Redirection" title="Aide:Redirection"></a> «Charles Michel» redirige ici. Pour les autres significations, voir <a href="/wiki/Charles_Michel_(homonymie)" class="mw-disambig" title="Charles Michel (homonymie)">Charles Michel (homonymie)</a>.</div>  
                            <table class="infobox_v2">  
                                <tr>  
                                    <td colspan="2" class="entete defaut" style="background-color:#6688FF; color:#FFFFFF">Charles Michel</td>  
                                </tr>  
                                <tr>  
                                    <td colspan="2" style="text-align:center; line-height: 1.5em;"><a href="/wiki/Fichier:Charles_Michel_(politician).jpg" class="image" title="Charles Michel en 2014."></a>  
                                </tr>  
                            </table>  
                        </div>  
                    </div>  
                </div>  
            </div>  
        </div>  
    </div>  
</body>
```

# INFORMATION SEMI-STRUCTURÉE 2 (HTML)

Article Discussion Lire Modifier Modifier le code Afficher l'historique Plus Rechercher dans Wikipédia

## Charles Michel (homme politique)

Charles Michel, né le 21 décembre 1975 à Namur, est un homme politique belge, membre du Mouvement réformateur. Premier ministre depuis le 11 octobre 2014, il est le chef de gouvernement le plus jeune de l'histoire de la Belgique<sup>2</sup>.

**Sommaire [masquer]**

- 1 Biographie
  - 1.1 Jeunesse et études
  - 1.2 Débuts en politique
  - 1.3 Ministre
  - 1.4 Premier ministre
- 2 Vie privée
- 3 Carrière politique
- 4 Mandats et fonctions<sup>[22]</sup>
- 5 Distinctions
- 6 Notes et références
- 7 Voir aussi
  - 7.1 Article connexe
  - 7.2 Liens externes



Charles Michel en 2014

**Fonctions**

Premier ministre belge

En fonction depuis le 11 octobre 2014  
(3 ans et 15 jours)

Biogra #Jeunesse\_et\_\.C3\.A9tudes , .toctext , .tocsection-1 , p , tc Clear (71) Toggle Position XPath ? X

# Expressions régulières (regexes)

À CHAQUE FOIS QUE  
J'APPRENDS QUELQUE CHOSE  
DE NOUVEAU, J'IMAGINE DES  
SCÉNARIOS ÉLABORÉS DÙ ÇA  
ME PERMET DE DEVENIR LE  
HÉROS DU JOUR.

OH NON ! LE TUEUR A  
DÙ LA SUIVRE SUR SON  
LIEU DE VACANCES !



MAIS POUR LES TROUVER IL FAUDRA FOUILLER  
PARMI 200 MB D'EMAILS EN CHERCHANT QUELQUE  
CHOSE QUI AURAIT LE FORMAT D'UNE ADRESSE !



QUE TOUT LE MONDE RECULE !



JE SAIS ME SERVIR DES  
EXPRESSIONS RÉGULIÈRES.



# LES EXPRESSIONS RÉGULIÈRES (REGEX)

- Mini-langage informatique
- Spécialisé dans la recherche/remplacement de texte
- Existent en différents « dialectes » (Open Refine = Regex Java ou Python)
- Indispensables pour travailler sérieusement sur du texte
- A première vue effrayantes :

\b((\+|00)32\s?|0)4(60|[789]\d)(-|\V|\s|\.|)(\d{2})\4(\d{2})\4(\d{2})\b

- Mais finalement pas plus que :

MMCCCLXXXVIII (2388)

Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz (« loi sur le transfert des obligations de surveillance de l'étiquetage de la viande bovine »)

# QUELQUES OUTILS

- RegExr ou Regex101: pour tester vos expressions
- Regexpert : pour visualiser des expressions complexes
- RegexOne : pour les apprendre
- Regex Cheat Sheet : en attendant de les retenir
- Regular-expressions.info : une référence
- Regular expressions library : des regex toutes faites

# EXERCICE REGEX

Suivez le lien ci-dessous et :

1. Essayez de « matcher » les **adresses email** (et seulement elles).
2. Essayez de matcher la date.

<https://regex101.com/r/LqTv7c/2>