

Une introduction au WebScraping

*Naviguer,
Aspirer,
Extraire,
Nettoyer*

 @Ben_Guinaudeau
 @benjaminguinaudeau
 benguinaudeau.com/

27 août 2024



Objectif: Un tremplin non technique vers le webscraping

1. Apprendre la structure d'un projet de webscraping
2. Introduire les requêtes HTTP et les langages HTML/CSS
3. Présenter les packages `{xml2}` et `{rvest}`

Programme de l'atelier

1. *Webscraping* comme dans "grattage du web"?
2. *Naviguer* Localiser les données brutes
3. *Aspirer* Lire une page HTML
4. *Extraire* Raffiner les données brutes
5. *Nettoyer*
6. 9 conseils pour devenir une "gratteuse" professionnelle
7. Exercices pratiques

Webscraping

Environ 733 000 résultats (0,44 secondes)

Langue détectée : Anglais

Français

webscraping

grattage web



Ouvrir dans Google Traduction • Commentaires

Webscraping

Webscraping

Aspiration/extraction automatique de données web

Données non-structurées,
lisibles par l'humain et
publiquement
accessibles

[actualites/20211209/314963.html#_Toc90474448](https://www.sénat.fr/actualites/actualite/2021/12/09/lesage-accuse-le-gouvernement-de-n'avoir-pas-prime-la-sante-des-personnes-vulnerables-dans-les-chsld.html)

Accès à l'information sur l'évolution de la pandémie de COVID-19

Mme Dominique Anglade

Mme Anglade : Merci, M. le Président. Dans le scandale des CHSLD, M. le Président, on sait que depuis le mois de janvier des gens savaient que les personnes âgées allaient être particulièrement touchées, notamment dans les CHSLD. On le sait puisque l'ancienne ministre de la Santé nous l'a confirmé. On le sait parce que l'ancien sous-ministre l'a confirmé. On le sait parce que l'INSPQ avait envoyé des signaux. On le sait parce que la Protectrice du citoyen l'a mentionné dans son rapport. Pourtant, le premier ministre lui-même nous dit que, dans le fond, il ne l'avait pas vu, on n'avait pas présenté ça, autour de la table, l'enjeu des CHSLD, pour aller s'assurer que les personnes vulnérables dans ces lieux puissent être protégées.

M. le Président, le 9 mars, quatre jours avant qu'il y ait l'adoption du décret d'urgence, il y a des scénarios qui ont été présentés au gouvernement par rapport à ce qui allait se passer dans la crise, différents scénarios. Ces scénarios-là n'ont pas été rendus publics.

Aujourd'hui, je demande au premier ministre de nous confirmer qu'il a bel et bien ces scénarios en sa possession et je lui demande de les rendre publics.

Le Président : M. le premier ministre.



Jeux de données
systématiques, structurés,
rectangulaires et lisibles
par la machine

leg	date	name	speech_id	text
27	1965-02-10	LESAGE	33550	M. le Président, j'invoque une question de privilège basée sur l'...
27	1965-02-10	JOHNSON	33551	M. le Président, est-ce que vous avez admis que le premier mi...
27	1965-02-10	LESAGE	33552	Le député de Brome, à la suite de ces choses, a subi une atta...
27	1965-02-10	JOHNSON	33553	Nous sommes peinés de ça, M. le Président.
27	1965-02-10	LESAGE	33554	... il est actuellement à l'hôpital et il est de mon devoir comme ...
27	1965-02-10	JOHNSON	33555	Le problème, M. le Président, que je soulevais, c'était...
27	1965-02-10	LESAGE	33556	J'étais pour dire à l'instant.
27	1965-02-10	JOHNSON	33557	Bien voici, il me semble que cela aurait dû être dit au début, on...
27	1965-02-10	LESAGE	33558	Je n'ai pas de raison à donner, M. le Président.
27	1965-02-10	JOHNSON	33559	... soulevait. Je n'ai pas compris les raisons qu'avait données le...
27	1965-02-10	LESAGE	33560	J'ai cité le règlement.
27	1965-02-10	JOHNSON	33561	S'il avait dit: « Le député de Brome ne peut pas être ici on aurai...
27	1965-02-10	LE PRESIDENT	33562	A l'ordre, messieurs!
27	1965-02-10	JOHNSON	33563	Et, deuxièmement, je pense que le premier ministre fait une er...
27	1965-02-10	LESAGE	33564	Je m'en doutais. En tout cas, ça n'a pas d'importance. Le journa...
27	1965-02-10	PINARD	33565	M. le Président, je ne voudrais faire aucune déclaration qui pou...
27	1965-02-10	LESAGE	33566	M. le Président, c'est une déclaration ministérielle.
27	1965-02-10	JOHNSON	33567	Je m'excuse, M. le Président. Nous allons laisser l'Orateur juger.
27	1965-02-10	LESAGE	33568	Il ne veut pas la vérité. M. JOHNSON: Au contraire. Bon, si vous ...
27	1965-02-10	JOHNSON	33569	M. le Président, je voudrais savoir en vertu de quel règlement; ...
27	1965-02-10	PINARD	33570	M. le Président, si on me permet de dire un mot sur la questi...
27	1965-02-10	LE PRESIDENT	33571	Je crois que c'est la pratique régulière de permettre à un minist...
27	1965-02-10	JOHNSON	33572	M. le Président, qu'on comprenne bien; il n'y a aucun député d...
27	1965-02-10	LESAGE	33573	C'est une question de privilège, voyons!
27	1965-02-10	JOHNSON	33574	... sub judice.
27	1965-02-10	LESAGE	33575	Il n'y a rien de ça sub judice.
27	1965-02-10	JOHNSON	33576	Non, mais ce sont tout de même des propos qui dérivent de ca...

Webscraping A quoi bon?



La révolution digitale est une opportunité unique pour les chercheurs.euses

1 Accès exhaustif à d'immense corpus de documents

- médias, accords internationaux, discours et procédures parlementaires, rapports annuels d'entreprises, jurisprudences, etc.

2 La trace digitale des réseaux sociaux permet l'étude de nombreux phénomènes sociaux

- polarisation affective/idéologique (Facebook), transmission de l'information parmi les acteurs d'un marché (Twitter), parcours professionnels (Linkedin), ...

Les sites web sont optimisés pour une navigation humaine

A quel âge Justin Trudeau est-il devenu premier ministre?

- A. naviguer vers un moteur de recherche
- B. chercher "Justin Trudeau dates importantes"
- C. sélectionner parmi les résultats un site contenant l'information
- D. extraire rapidement l'information cherchée en exploitant le formatage

Mais l'extraction systématique de données peut rapidement devenir laborieux

A quel âge les premiers ministres du Canada sont-ils devenus premier ministre?

Répéter A-D pour chacun des 23 premiers ministres

- 1 Justin Trudeau: A, B, C,D
- 2 Stephen Harper: A, B, C, D
- 3 Paul Martin: A, B, C, D
- ...
- 23 John Macdonald: A, B, C, D

Aspiration automatique de données web

Avantages

- efficace (même pour des ensembles de mégadonnées)
- systématique (données disponibles pour la population ; pas besoin d'échantillonage)
- précis
- (reproductible)

Inconvénients

- détournement de l'objectif initial des sites internet (processus technique)
- potentielle violation des CGU

Comment approcher un projet d'aspiration automatique?

There is no one solution to all problems. It's the problem itself that can lead to the solution.

~Jay Maisel

Chaque projet est unique, mais tous reposent sur les mêmes compétences:

- ingénierie URL
- connaissances minimales des requêtes HTTP et du langage HTML
- manipulation des sélecteurs CSS

Une approche unifiée pour les projets d'aspiration automatique

1 *Naviguer* Localiser les données brutes

2 *Aspirer* Accéder à l'information brute

3 *Extraire* Raffiner les données brutes

4 *Nettoyer*

1 *Naviguer* Localiser les données brutes



Naviguer Localiser les données brutes

- Quelles données doivent être prélevées?
- Quelles pages devront être visitées afin de collecter les données?
- Est-ce que la collecte devra être répétée chaque jour, chaque semaine, chaque année?

Exemple 1 Dates des élections fédérales canadiennes et allemandes

Site officiel d'Elections Canada

The screenshot shows the homepage of the Elections Canada website. The header includes the logo, a search bar, and navigation links for Accueil, À propos de nous, and Contactez-nous. The main menu features categories like Électeurs, Élections, Centre de ressources, Médias, Emplois, Financement politique, and Participez. A sidebar on the left provides information about voter turnout and election campaigns. The central content area displays the 'Élections passées' section, which lists previous general elections:

Choisissez une élection générale :

- [43^e élection fédérale, 21 octobre 2019](#)
- [42^e élection fédérale, 19 octobre 2015](#)
- [41^e élection fédérale, 2 mai 2011](#)
- [40^e élection fédérale, 14 octobre 2008](#)
- [39^e élection fédérale, 23 janvier 2006](#)
- [38^e élection fédérale, 28 juin 2004](#)
- [37^e élection fédérale, 27 novembre 2000](#)
- [36^e élection fédérale, 2 juin 1997](#)

Exemple 1 Dates des élections fédérales canadiennes et allemandes

Site officiel des élections allemandes



Deutscher Bundestag

Abgeordnete	Parlament	Ausschüsse	Internationales	Dokumente	Mediat
-------------	-----------	------------	-----------------	-----------	--------

Startseite ▶ Parlament ▶ Wahlen ▶

Bundestagswahlergebnisse seit 1949 – Zweitstimmen

Jahr	CDU/CSU	SPD	FDP	Die Grünen	Bündnis 90/Die Grünen	Die Linke. PDS	AfD	Sonstige
2021	24,1	25,7	11,5		14,8	4,9	10,3	8,7
2017	32,9	20,5	10,7		8,9	9,2	12,6	5,0
2013	41,5	25,7	4,8		8,4	8,6		11
2009	33,8	23,0	14,6		10,7	11,9		6,0
2005	35,2	34,2	9,8		8,1	8,7		4,0
2002	38,5	38,5	7,4		8,6	4,0		3,0
1998	35,2	40,9	6,2		6,7	5,1		5,9
1994	41,5	36,4	6,9		7,3	4,4		3,5
1990	43,8	33,5	11,0	3,8	1,2	2,4		4,3
1987	44,3	37,0	9,1	8,3				1,3
1983	48,8	38,2	7,0	5,6				0,4
1980	44,5	42,9	10,6	1,5				0,5
1976	48,6	42,6	7,9					0,9
1972	44,9	45,8	8,4					0,9

Exemple 1 Dates des élections fédérales canadiennes et allemandes

Liste des élections canadiennes sur Wikipédia

en.wikipedia.org/wiki/List_of_Canadian_federal_general_elections	
20th	1945 <i>populaire Canadien</i> wins two seats in Quebec on an anti-conscription and Quebec nationalism platform; future Prime Minister Pierre Trudeau and future mayor of Montreal Jean Drapeau are young party members.
21st	1949 Liberals, led by Liberal Prime Minister Louis St-Laurent, are re-elected with a majority, defeating Progressive Conservatives led by George Drew.
22nd	1953 Prime Minister St-Laurent's Liberals are re-elected with a majority, defeating Drew's Progressive Conservatives.
23rd	1957 Progressive Conservatives, led by John Diefenbaker, defeat Liberals led by Prime Minister St-Laurent with an upset minority victory.
24th	1958 Progressive Conservatives, led by Prime Minister Diefenbaker, are re-elected with the largest majority to date in Canadian history, defeating Liberals and their new leader Lester Pearson.
25th	1962 Progressive Conservatives, led by Prime Minister Diefenbaker, are re-elected, but with a minority. Under "father of Canadian medicare" Tommy Douglas, the New Democratic Party, evolved from the CCF, wins 19 seats but fails to achieve a hoped-for breakthrough. Social Credit makes unprecedented gains in Quebec, but only a modest recovery in the West.
26th	1963 Liberals, led by Lester Pearson, defeat Prime Minister Diefenbaker's Progressive Conservatives, winning a minority.
27th	1965 Liberals, led by Prime Minister Pearson, are re-elected with a second minority, defeating former Prime Minister Diefenbaker's Progressive Conservatives.
28th	1968 Liberals, led by new Prime Minister Pierre Trudeau, are re-elected with a majority, defeating Progressive Conservatives led by Robert Stanfield.
29th	1972 Liberals, led by Prime Minister P. Trudeau, are re-elected, but with a minority, defeating Stanfield's Progressive Conservatives by only two seats. The NDP pick up several seats under new leader David Lewis.
30th	1974 Liberals, led by Prime Minister P. Trudeau, defeat Stanfield's Progressive Conservatives with a majority. Progressive Conservatives, led by Joe Clark, defeat Liberals, led by Prime Minister P. Trudeau, and win a minority, despite winning a significantly smaller share of the vote than the Liberals. The PCs
31st	1979 win the popular vote in seven provinces, but the Liberals capture an enormous lead in Quebec. Ed Broadbent makes his debut as leader of the NDP, which wins 10 more seats than in 1974 in a Parliament enlarged by 18 seats.
32nd	1980 Liberals, led by former Prime Minister P. Trudeau, defeat Progressive Conservatives, led by Prime Minister Clark. Social Credit fades into history after an almost unbroken 45-year run, leaving Canada with a three-party system.
33rd	1984 Progressive Conservatives, led by Brian Mulroney, defeat Liberals, led by new Prime Minister John Turner and win the most seats in Canadian history. The election is both the best showing ever for the Progressive Conservatives and the second-worst showing ever for the Liberals (by total seats).
34th	1988 Progressive Conservative Prime Minister Mulroney is re-elected with a second majority, contending with a much stronger performance from former Liberal Prime Minister Turner and a strong third-party showing from Broadbent's New Democrats, who score that party's third best result ever. Liberals, led by Jean Chrétien, win a majority and soundly defeat Progressive Conservatives, led by new Prime Minister Kim Campbell, who are left in fifth place with just two seats, their worst ever showing.
35th	1993 The separatist Bloc Québécois under ex-Mulroney cabinet minister Lucien Bouchard becomes the official opposition, and the right-wing Reform Party, led by Preston Manning, becomes the third party. Audrey McLaughlin's New Democrats also post their worst ever results with just nine seats. The election marks the end of the predominantly three-party system of the Liberals, Progressive Conservatives, and NDP.
36th	1997 Liberals, led by Prime Minister Chrétien, are re-elected with a second majority. Manning's Reform Party becomes the official opposition. Bloc Québécois falls to third place under new leader Gilles Duceppe. NDP under Alexa McDonough win 21 seats, 12 more than in 1993. Progressive Conservatives under Jean Charest win nearly as many votes as Reform, but only one-third the seats.
37th	2000 Liberals, led by Prime Minister Chrétien, are re-elected with a third majority, defeating Stockwell Day's Canadian Alliance, the unsuccessful attempt to unite the Reform Party and the Progressive

Exemple 1 Dates des élections fédérales canadiennes et allemandes

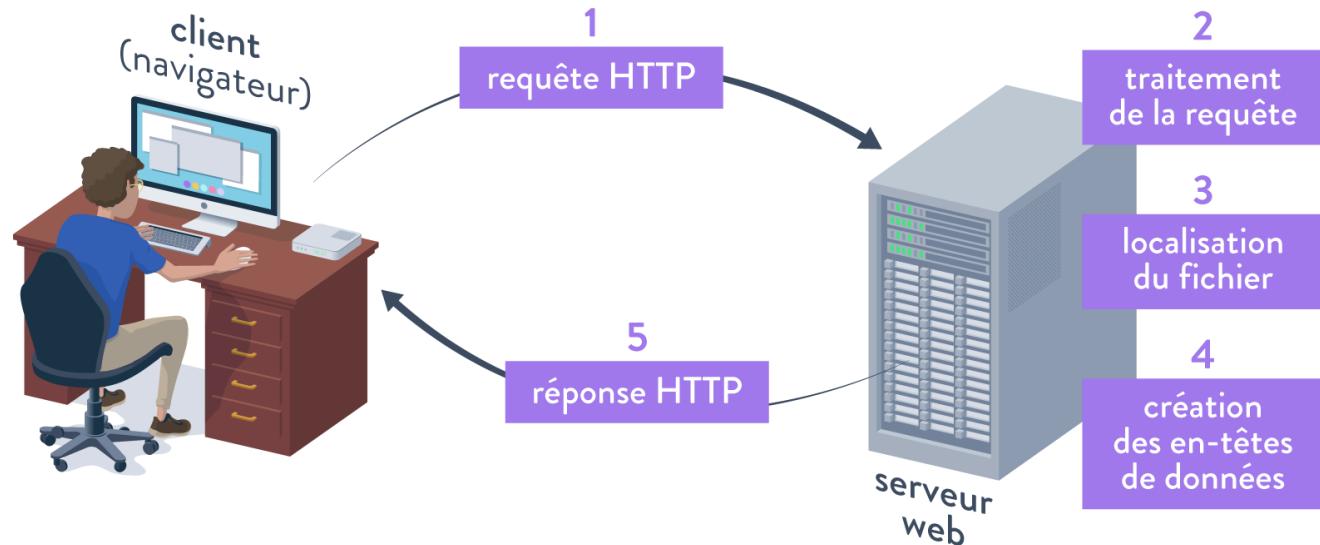
Liste des élections allemandes sur Wikipédia

The screenshot shows a table from the German Wikipedia page on federal elections (Bundestagswahl). The table lists the date of each election, the number of mandates won by various parties, and other relevant information. The parties listed include CDU/CSU, SPD, FDP, Grüne, PDS/Linke, AfD, DP, Z, and Sonstige.

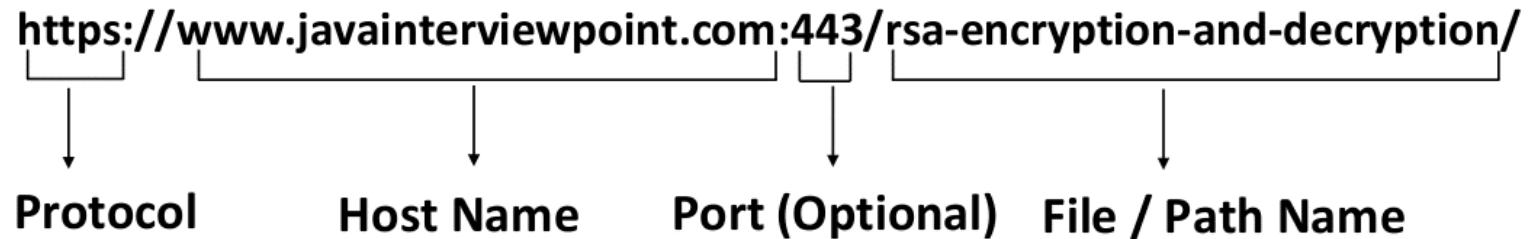
Wahltag	Mandate	CDU/CSU	SPD	FDP	Grüne	PDS/Linke	AfD	DP	Z	Sonstige
14. August 1949	402	139	131	52	—	—	—	17	10	KPD 15; BP 17; WAV 12; DKP-DRP 5; SSW 1; Unabhängige 3
6. September 1953	487	243	151	48	—	—	—	15	3	GB/BHE 27
15. September 1957	497	270	169	41	—	—	—	17	—	
17. September 1961	499	242	190	67	—	—	—	—	—	
19. September 1965	496	245	202	49	—	—	—	—	—	
28. September 1969	496	242	224	30	—	—	—	—	—	
19. November 1972	496	225	230	41	—	—	—	—	—	
3. Oktober 1976	496	243	214	39	—	—	—	—	—	
5. Oktober 1980	497	226	218	53	—	—	—	—	—	
6. März 1983	498	244	193	34	27	—	—	—	—	
25. Januar 1987	497	223	186	46	42	—	—	—	—	
2. Dezember 1990	662	319	239	79	8	17	—	—	—	
16. Oktober 1994	672	294	252	47	49	30	—	—	—	
27. September 1998	669	245	298	43	47	36	—	—	—	
22. September 2002	603	248	251	47	55	2	—	—	—	
18. September 2005	614	226	222	61	51	54	—	—	—	
27. September 2009	622	239	146	93	68	76	—	—	—	
22. September 2013	631	311	193	—	63	64	—	—	—	
24. September 2017	709	246	153	80	67	69	94	—	—	
26. September 2021	736	197	206	92	118	39	83	—	—	SSW 1

2 Aspirer Charger la donnée brute

Le protocole HTTP



Aspirer Comprendre la structure d'une URL



Aspirer Comprendre la structure d'une URL

https://en.wikipedia.org:80/wiki/List_of_Canadian_federal_general_elections

Au moins quatre éléments composent une URL

1. un protocole `https://`
2. un nom de domaine `en.wikipedia.org`
3. un port `:80`
4. un chemin vers le contenu spécifique
`/wiki>List_of_Canadian_federal_general_elections`

Aspirer avec R et le package {xml2}

Input

```
url <-  
"https://en.wikipedia.org/wiki/List_of_Canadian_federal_general_elections";  
page <- xml2::read_html(url)  
page
```

► Run

Output

```
xml2::write_html(page, file = "test.html")  
# file.show("test.html")  
# browseURL("test.html")
```

3 Extraire Raffiner les données avec `{rvest}`



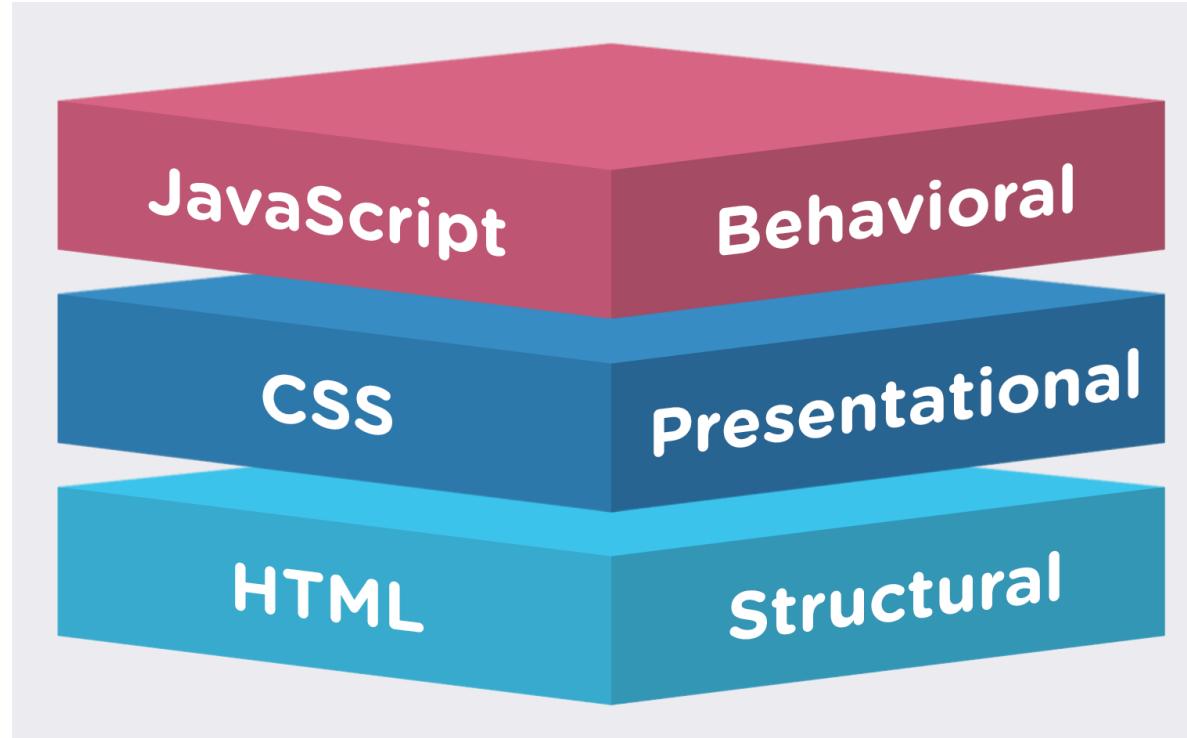
Exemple 2 Extraire les noms, partis et liens des députés canadiens

Exemple 2

noscommunes.ca/members/fr/constituencies

Aurora—Oak Ridges—Richmond Hill Ontario  Leah Taylor Roy Libéral	Avalon Terre-Neuve-et-Labrador  Ken McDonald Libéral	Avignon—La Mitis—Matane—Matapedia Québec  Kristina Michaud Bloc Québécois
Baie de Quinte Ontario  Ryan Williams Conservateur	Banff—Airdrie Alberta  Blake Richards Conservateur	Barrie—Innisfil Ontario  John Brassard Conservateur
Barrie—Springwater—Oro-Medonte Ontario  Doug Shipley Conservateur	Battle River—Crowfoot Alberta  Damien C. Kurek Conservateur	Battlefords—Lloydminster Saskatchewan  Rosemarie Falk Conservateur
Beaches—East York Ontario  Nathaniel Erskine-Smith Libéral	Beauce Québec  Richard Lehoux Conservateur	Beauport—Côte-de-Beaupré—Île d'Orléans—Charlevoix Québec  Caroline Desbiens Bloc Québécois

Extraire Qu'est-ce qu'une page web?



Extraire Qu'est-ce qu'une page web?

Une page web a trois composantes principales:

1. **HTML (Hyper Text Markup Language)**

prend en charge le contenu (textes, images, liens internes/externes)

2. **CSS (Cascading Style Sheets)**

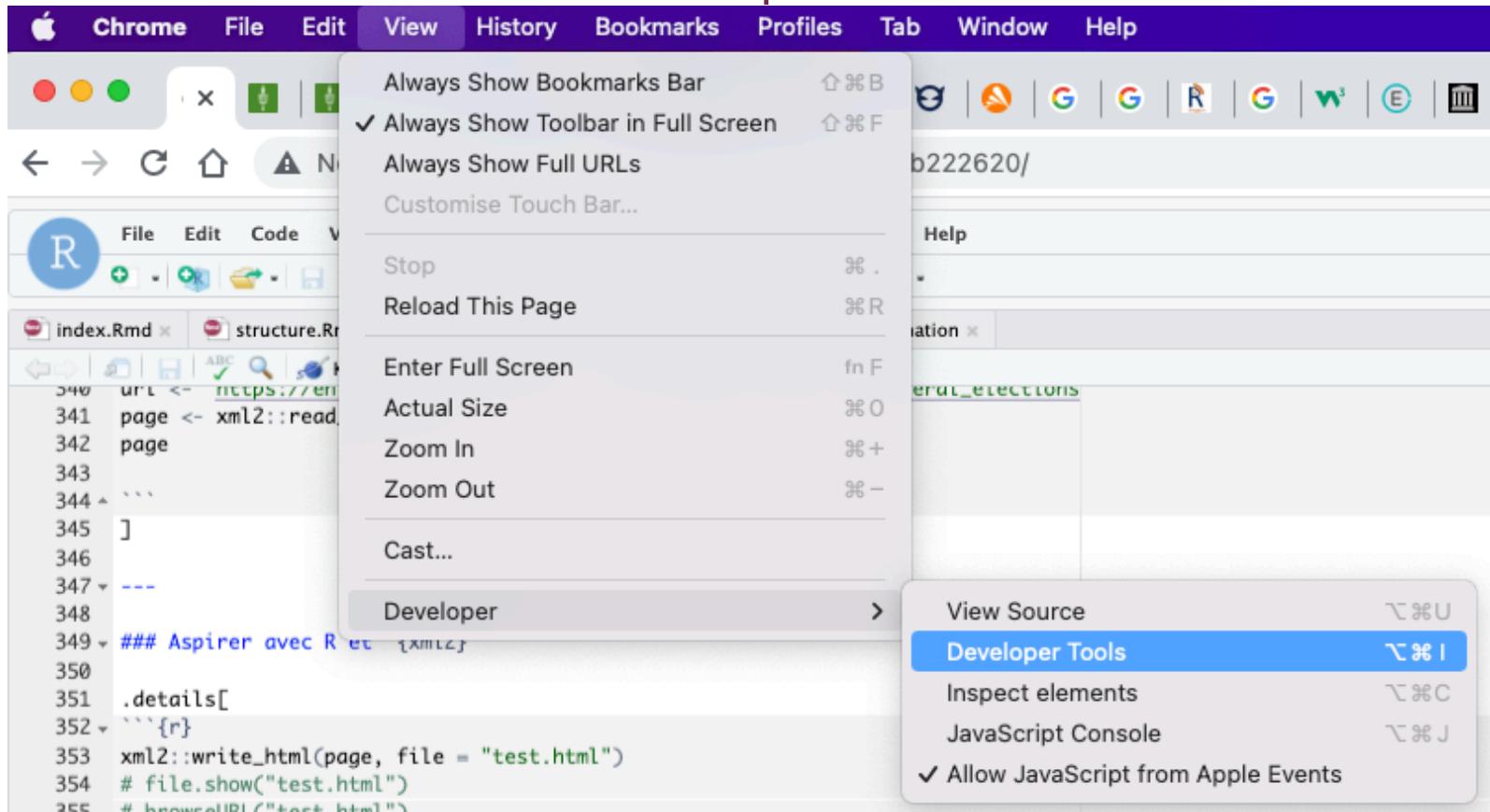
définit l'apparence des éléments HTML

3. **JS (Javascript)**

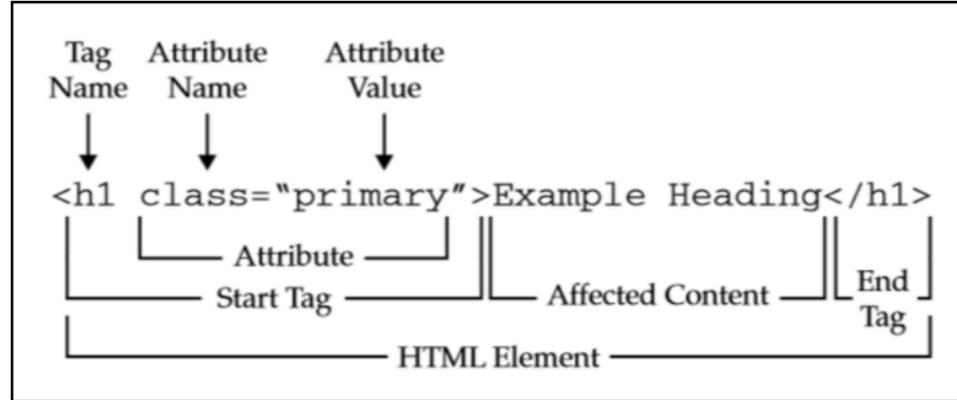
modifie l'apparence/le contenu en fonction des interactions avec l'utilisateur

Extraire Disséquer une page web avec les outils developer

Example 2



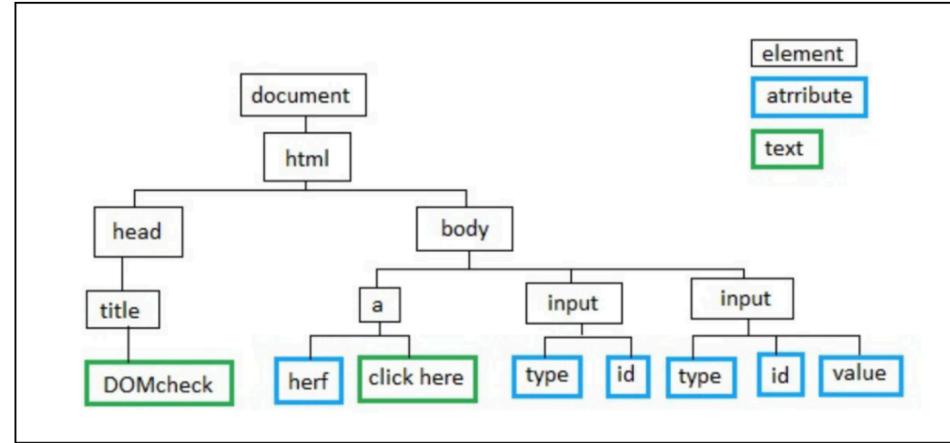
Extraire Une mosaïque d'éléments HTML



Chaque élément HTML

- est délimité par des balises de début `<tag>` et de fin `</tag>`
- est caractérisé par des attributs insérés dans la balise de début (id, class, type, style, etc.)
- peut contenir un ou plusieurs éléments enfants (dont il est le parent)

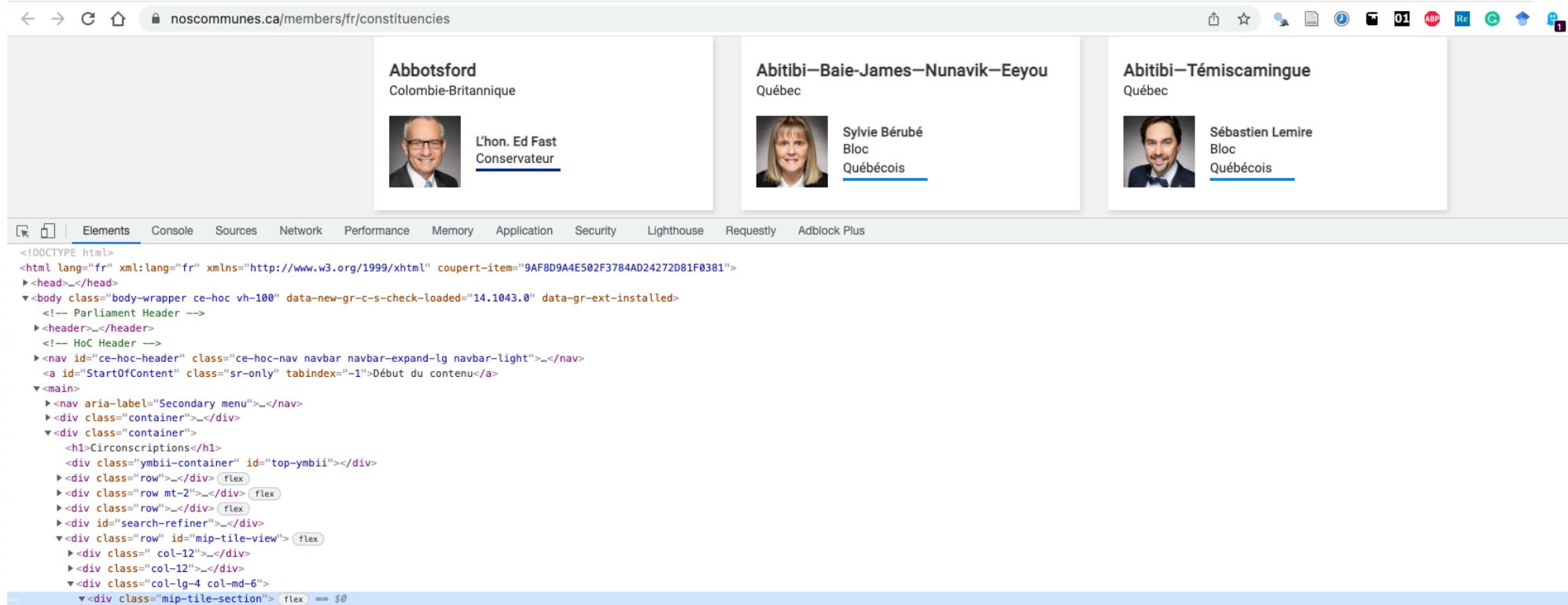
Extraire Une mosaïque d'éléments HTML



Chaque page web peut être représentée comme une arborescence:

- avec un sommet initial unique (**document**),
- qui contient au moins deux enfants **head** et **body**

Extraire Une mosaïque d'éléments HTML



The screenshot shows a web browser displaying a grid of political constituency profiles. The profiles are arranged in three columns. The first column contains the profile for Abbotsford, Colombie-Britannique, featuring L'hon. Ed Fast, Conservateur. The second column contains the profile for Abitibi-Baie-James-Nunavik-Eeyou, Québec, featuring Sylvie Bérubé, Bloc Québécois. The third column contains the profile for Abitibi-Témiscamingue, Québec, featuring Sébastien Lemire, Bloc Québécois. Below the browser window, the browser's developer tools are visible, with the 'Elements' tab selected. The HTML code is displayed in a tree structure, showing the hierarchical structure of the web page, including elements like the header, main content area, and individual constituency tiles.

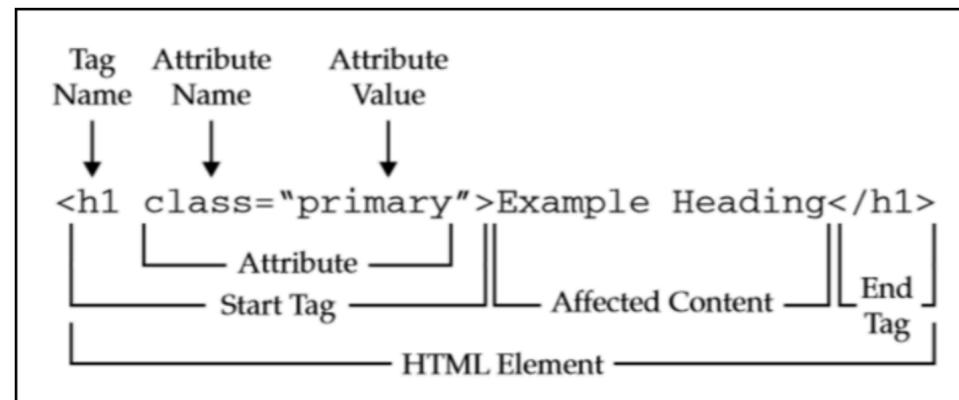
Extraire Parcourir l'arborescence HTML avec `{rvest}`

1 Isoler l'élément contenant la donnée

```
webpage <- xml2::read_html(url)
elem <- isolate_element(webpage, "id_of_the_element_ofContaining_the_data")
```

2 Extraire l'information

```
value <- extract_data(elem)
```



Extraire Parcourir l'arborescence HTML avec `{rvest}`

2 fonctions d'isolation

- `html_element(parent, sélecteur)` isole le premier élément enfant de l'élément `parent` correspondant à `sélecteur`
- `html_elements(parent, sélecteur)` isole l'ensemble des éléments enfant de l'élément `parent` correspondant à `sélecteur`

2 fonctions d'extraction

- `html_text(element)` extrait le texte de `element`
- `html_attr(element, attr)` extrait l'attribut `attr` de `element`

Extraire isoler avec `rvest::html_element`

`html_element(parent, sélecteur)` isole le premier élément enfant de l'élément `parent` correspondant à `sélecteur`

```
page <- xml2::read_html("https://www.noscommunes.ca/members/fr/constituencies")
rvest::html_element(page, "body")
```

► Run

Extraire isoler avec `rvest::html_element`

`html_element(parent, sélecteur)` isole le premier élément enfant de l'élément `parent` correspondant à `sélecteur`

```
rvest::html_element(page, "header")
```

► Run

Extraire isoler avec `rvest::html_element`

`html_element(parent, sélecteur)` isole le premier élément enfant de l'élément `parent` correspondant à `sélecteur`

```
rvest::html_element(page, "[class='mip-mp-name'])")
```

► Run

```
# Shortcut for class  
rvest::html_element(page, ".mip-mp-name")
```

► Run

Extraire isoler avec `rvest::html_elements`

`html_elements(parent, sélecteur)` isole l'ensemble des éléments enfant de l'élément `parent` correspondant à `sélecteur`

```
rvest::html_elements(page, "[class='mip-mp-name'])")
```

► Run

Extraire extraire le texte `rvest::html_text`

`html_text(element)` extrait le texte de `element`

```
# Isole l'ensemble des éléments dont la classe contient `mip-constituency-tile`  
mp_tuiles <- rvest::html_elements(page, "[class='mip-mp-name'])")  
# Pour chaque élément de l'ensemble, extrait l'attribut href  
rvest::html_text(mp_tuiles)
```

► Run

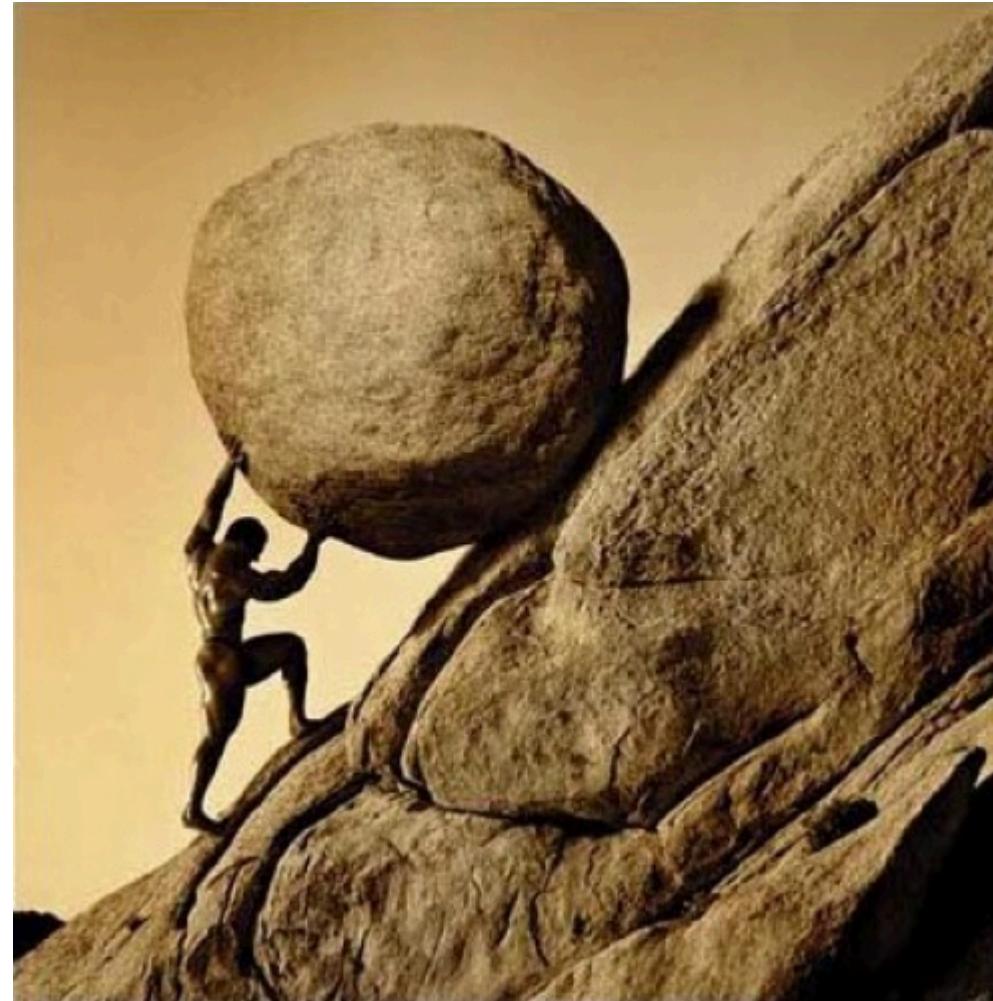
Extraire extraire des attributs avec `rvest::html_attr`

`html_attr(element, attr)` extrait l'attribut `attr` de `element`

```
# Isole l'ensemble des éléments dont la classe contient `mip-constituency-tile`  
mp_tuiles <- rvest::html_elements(page, "[class='mip-constituency-tile']]")  
# Pour chaque élément de l'ensemble, extrait l'attribut href  
rvest::html_attr(mp_tuiles, "href")
```

► Run

4 Nettoyer L'éternel recommencement



Nettoyer L'éternel recommencement

Une fois extraite, il faut encore:

- transformer les valeurs numériques en `numeric`
- transformer les dates en `Date`
- ajouter de nouvelles variables (notamment `lgl`)
- enlever les espaces superflus des vecteurs `chr`

Tout cela, à l'aide des suspects habituels:

- `{dplyr}` pour manipuler les jeux de données
- `{stringr}` pour manipuler les vecteurs `chr`
- `{lubridate}` pour manipuler les dates
- `{purrr}` pour manipuler/itérer sur les listes

Pour résumer

Un projet d'aspiration automatique s'articule autour de quatre axes:

1. *Naviguer* collecter l'ensemble des adresses à visiter
2. *Aspirer* charger le code HTML formant la page brute
3. *Extraire* raffiner les données brutes
4. *Nettoyer* finaliser les données selon le format souhaité

9 conseils pour devenir une "gratteuse" professionnelle



1 Utilisez SelectorGadget pour se familiariser avec les sélecteurs CSS

Pour en apprendre plus [1](#) ; [2](#)

Affiliation politique Province/Territoire Nom de circonscription

.mip-constituency-tile Clear (338) Toggle Position XPath ? X

Toutes Toutes Toutes Réinitialiser

Résultats 338 de 338

Abbotsford Colombie-Britannique  L'hon. Ed Fast Conservateur	Abitibi—Baie-James—Nunavik—Eeyou Québec  Sylvie Bérubé Bloc Québécois	Abitibi—Témiscamingue Québec  Sébastien Lemire Bloc Québécois
Acadie—Bathurst Nouveau-Brunswick  Serge Cormier Libéral	Ahuntsic-Cartierville Québec  L'hon. Mélanie Joly Libéral	Ajax Ontario  L'hon. Mark Holland Libéral
Alfred-Pellan Québec  Angelo Iacono	Algoma—Manitoulin—Kapuskasing Ontario  Carol Hughes	Argenteuil—La Petite-Nation Québec  Stéphane Lauzon

2 Considérez les implications techniques et légales de l'aspiration automatique

Les noms de domaines incluent un fichier `robots.txt` précisant les sous-domaines pouvant être aspirés automatiquement.

- {robotstxt}

2 Considérez les implications techniques et légales de l'aspiration automatique

Chaque requête HTTP exerce un poid technique sur le serveur, évitez de surcharger avec des requêtes rapides et réduisez le rythme de requête.

```
for(link in links){  
  aspire(link)  
  # Ajouter une second de pause entre chaque aspiration  
  Sys.sleep(1)  
}
```

2 Considérez les implications techniques et légales de l'aspiration automatique

Qu'en est-il des termes légaux?

- Jurisprudence en mouvement et qui dépend de la jurisdicition
- Principes à garder en tête:
 - Utilisation académique vs. utilisation commerciale
 - Données publiques vs. données privées (authentification requise?)
 - Informations personnelles identifiables (PII)
- Nécessité d'évaluer le risque d'être attaqué en justice

3 Aspirez directement les tableaux avec `rvest::html_table`

Les tableaux HTML (y compris ceux de Wikipédia) peuvent être aspirés directement.

Germany Canada

```
"https://de.wikipedia.org/wiki/Bundestagswahl" %>%  
xml2::read_html() %>%  
rvest::html_table() %>%  
purrr::chuck(2)
```

► Run

4 Apprenez à maîtriser les expressions régulières

Les expressions régulières (regex) sont un langage pour décrire des schémas de caractère.

Par exemple, `\d{2}\s\d{2}` correspond à deux chiffres, suivis d'un espace, suivi de deux chiffres. Si elles semblent parfois obscures et aléatoires, elles permettent un nettoyage efficace et rapide de données textuelles.

Pour en apprendre plus sur les regex [1](#) ; [2](#)

5 Peaufinez vos compétences en sélecteurs CSS

Les sélecteurs CSS sont souvent (comme ici) évoqués de façon superficielle. Ce sont, en réalité, des outils à la syntaxe très puissante.

Pour en apprendre plus [1](#) ; [2](#) ; [3](#) ; [4](#) ; [5](#)

6 Si elles existent, utilisez les API officielles

Une API (Application Programming Interface) permet la dissémination de données formatées pour les machines.

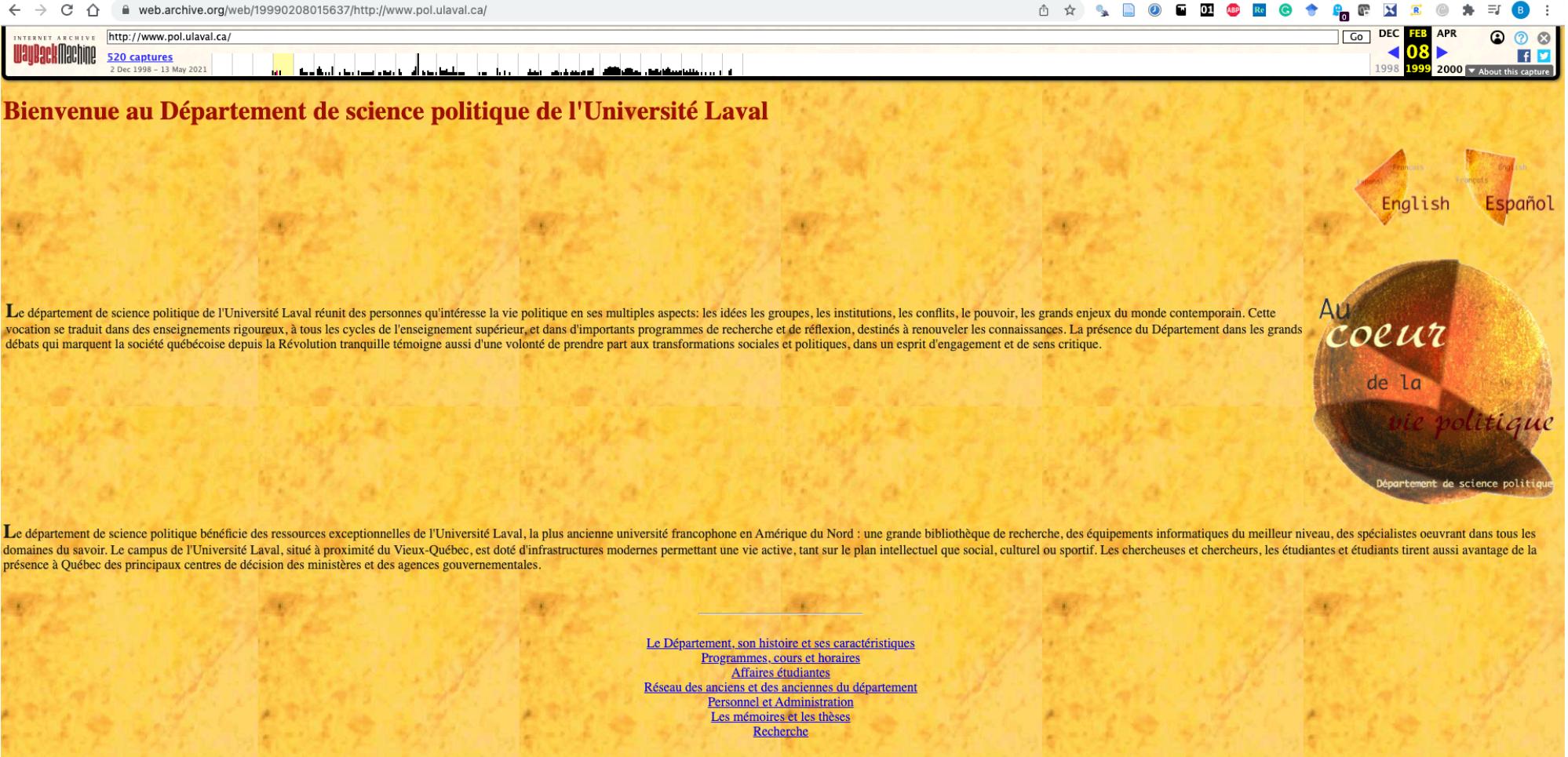
- [academictwitteR](#): Twitter
- [riingo](#): Données boursières (Tiingo)
- [tuber](#): Youtube
- [WikipediR](#): Wikipédia
- [RedditExtractor](#): Reddit

7 Remontez le temps avec la Wayback Machine

Le département de science politique de l'Université Laval réunit des personnes qu'intéresse la vie politique en ses multiples aspects: les idées les groupes, les institutions, les conflits, le pouvoir, les grands enjeux du monde contemporain. Cette vocation se traduit dans des enseignements rigoureux, à tous les cycles de l'enseignement supérieur, et dans d'importants programmes de recherche et de réflexion, destinés à renouveler les connaissances. La présence du Département dans les grands débats qui marquent la société québécoise depuis la Révolution tranquille témoigne aussi d'une volonté de prendre part aux transformations sociales et politiques, dans un esprit d'engagement et de sens critique.

Le département de science politique bénéficie des ressources exceptionnelles de l'Université Laval, la plus ancienne université francophone en Amérique du Nord : une grande bibliothèque de recherche, des équipements informatiques du meilleur niveau, des spécialistes oeuvrant dans tous les domaines du savoir. Le campus de l'Université Laval, situé à proximité du Vieux-Québec, est doté d'infrastructures modernes permettant une vie active, tant sur le plan intellectuel que social, culturel ou sportif. Les chercheuses et chercheurs, les étudiantes et étudiants tirent aussi avantage de la présence à Québec des principaux centres de décision des ministères et des agences gouvernementales.

[Le Département, son histoire et ses caractéristiques](#)
[Programmes, cours et horaires](#)
[Affaires étudiantes](#)
[Réseau des anciens et des anciennes du département](#)
[Personnel et Administration](#)
[Les mémoires et les thèses](#)
[Recherche](#)



The screenshot shows a Wayback Machine interface with a timeline from December 1998 to April 2000. The specific capture is from February 8, 1999. The URL in the address bar is <http://www.pol.ulaval.ca/>. The page title is "Bienvenue au Département de science politique de l'Université Laval". The page content includes a yellow background with text about the department's mission and resources, and a circular logo on the right side with text in English and Spanish, and "AU coeur de la vie politique" in the center.

8 Révélez les API cachées dans les site web grâce à la rétro-ingénierie des sites web

Beaucoup de sites web conservent leurs données dans des bases de données accessibles via une API personnalisée. Ces API "officieuses" peuvent être exploitées afin d'extraire un grand nombre de données.

Pour en apprendre plus [1](#) ; [2](#)

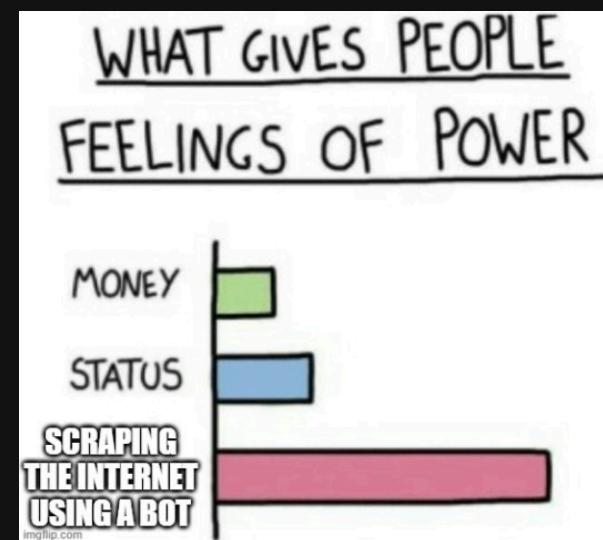
9 Simulez de réels navigateurs web afin de dépasser les problèmes de JS

Pour les plus acharnées souhaitant aspirer des pages impliquant des interactions complexes:

- [Selenium](#)
- [Puppeteer](#)
- [Playwright](#)/[Playwrightr](#)

Merci pour votre attention!

et n'oubliez pas



Questions



Exercice pratiques

1. Tableau des médailles:

fr.wikipedia.org/wiki/Tableau_des_m%C3%A9dailles_des_Jeux_olympiques_d%27%27

2. Vélo en libre service: aveloquebec.ca/

3. Résultats détaillés des JO: olympics.com/fr/paris-2024/calendrier/5-aout

4. Twitter: x.com/pol_ulaval

