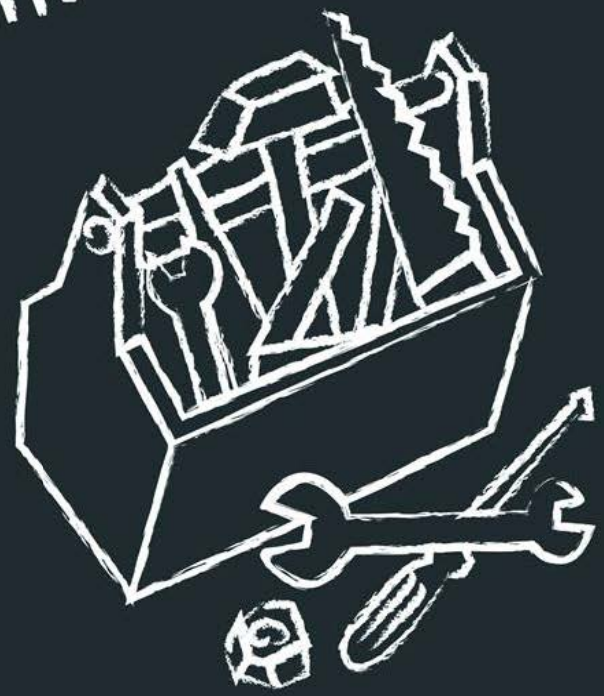
mshe
MAISON DES SCIENCES DE L'HOMME ET DE L'ENVIRONNEMENT
CLAUDE NICOLAS LEDOUX—USR 3124

LES TRUCS ET ASTUCES DE LA PLATE-FORME TECHNOLOGIQUE



Web scraping

Virginie Lethier
virginie.lethier@univ-fcomte.fr

PROGRAMME SUR
HTTPS://MSHE.UNIV-FCOMTE.FR

Pourquoi cette formation? Pour qui?

- Acquérir des données textuelles pour une recherche, un mémoire, pour organiser un projet (adresses mail, numéros de téléphone, liste d'articles) et/ou une veille informationnelle, pour archiver…
- Aucune compétence en programmation requise

Web scraper ? Harvesting ?

- Utiliser des « robots » pour « aspirer » des sites web

- Un **web scraper** est un programme informatique qui lit le code html des pages web pour extraire des données qui sont présentées sur des sites **web**.

Web scraper, c'est hacker ?

- Scraper n'est pas une activité illégale en soi

- Mais une utilisation des données webscrapées peut aller contre l'éthique ou être véritablement illégale...

  Rappel : toujours lire et respecter les conditions générales d'utilisation (CGU) et les conditions générales de vente (CGV) avant de web scrapper .

Image : UnderNews Actu

- Rappel des bonnes pratiques et des problématiques liées aux données de la recherche dans un contexte d'open data :
  http://www.oecd.org/fr/sti/inno/38500823.pdf
  http://www.donneesdelarecherche.fr/

De nombreux outils disponibles

- Parmi lesquels : Import.io, Scrapy, Outwit Hub, Gromoteur
- Webscraper est une extension disponible sous Google Chrome qui permet d'extraire les données d'un site internet très rapidement.

(1) Je télécharge sur mon ordinateur l'extension de Google Chrome : http://webscraper.io/

Web Scraper

Web Scraper is a company specializing in data extraction from web pages. We offer 2 great opti
Google Chrome Web Scraper Extension, and cloud based Web Scraper.

Web Scraper Extension (Free!)

Using our extension you can create a plan (sitemap) how a web site should be traversed and what should be extracted. Using these sitemaps the Web Scraper will navigate the site accordingly and extract all data. Scraped data later can be exported as CSV.

Download Free on Chrome Store »

Cloud Web Scraper

Cloud Web Scraper offers top quality resu
This option allows you to extract large amou
at once, and even run them on a set schedu

Read more about the Cloud Wel

(2) Je vais sur le site dont je souhaite aspirer le contenu

FDESOUCHE.COM EST UNE REVUE DE PRESSE

ACCUEIL SOCIÉTÉ SÉCURITÉ POLITIQUE ECONOMIE MONDE ECOLOGIE CULTURE SPORT INSOLITE FDÉSINTOX

Accueil / Mentions légales

Mentions légales

Ce blog est édité par Tilak RAJ.

1203/10 Govind Puri, *kalkaji*, New Delhi 110019, Inde

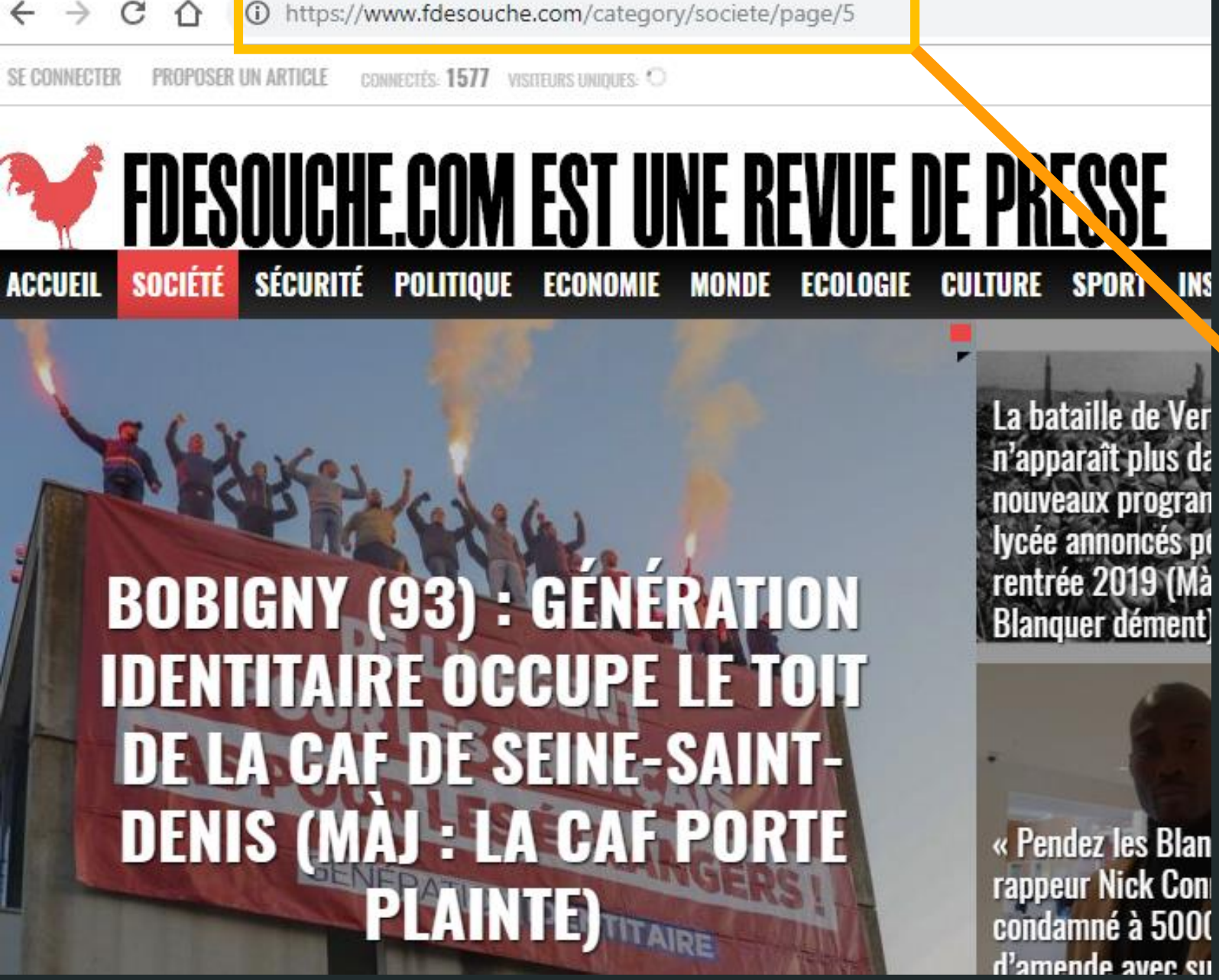
Pour toutes questions, requêtes ou réclamations vous pouvez contacter l'éditeur par courriel à l'adresse suivante : fdesouche@india.com

Ce site participe au Programme Partenaires d'Amazon EU, un programme d'affiliation conçu pour permettre à des sites de percevoir une rémunération grâce à la création de liens vers Amazon.fr.

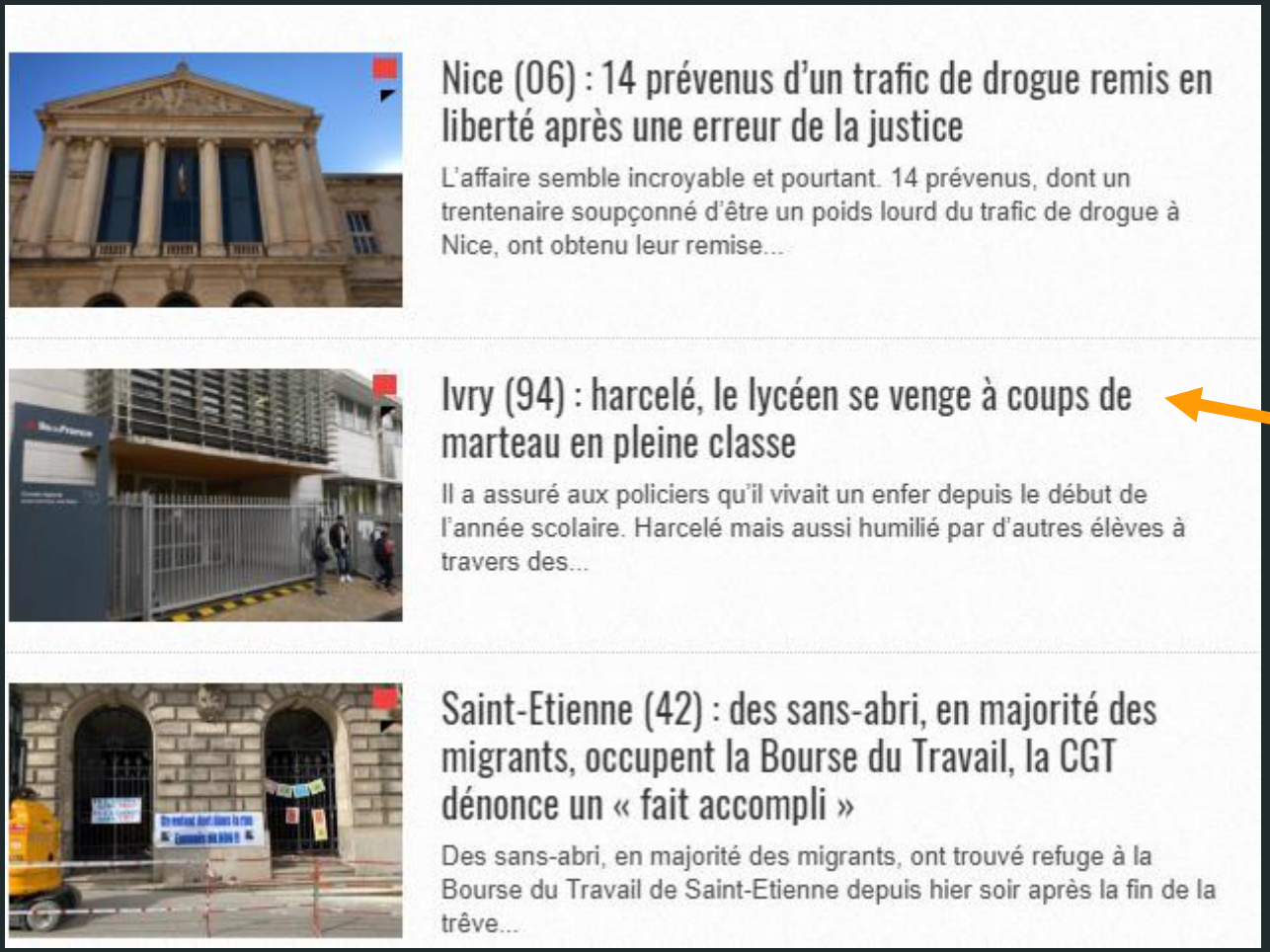
(3) J'analyse la présentation du site

J'observe la variation de l'adresse liée à la rubrique « Société »

J'observe le nombre de pages…



J'analyse l'adresse de la page 5 de la rubrique « Société »



Nice (06) : 14 prévenus d'un trafic de drogue remis en liberté après une erreur de la justice

L'affaire semble incroyable et pourtant. 14 prévenus, dont un trentenaire soupçonné d'être un poids lourd du trafic de drogue à Nice, ont obtenu leur remise...

Ivry (94) : harcelé, le lycéen se venge à coups de marteau en pleine classe

Il a assuré aux policiers qu'il vivait un enfer depuis le début de l'année scolaire. Harcelé mais aussi humilié par d'autres élèves à travers des...

Saint-Etienne (42) : des sans-abri, en majorité des migrants, occupent la Bourse du Travail, la CGT dénonce un « fait accompli »

Des sans-abri, en majorité des migrants, ont trouvé refuge à la Bourse du Travail de Saint-Etienne depuis hier soir après la fin de la trêve...

Chaque PAGE contient PLUSIEURS LIENS vers des articles. Le lien correspond au titre de l'article.

La mise en forme de chaque lien apparaît rigoureusement identique.

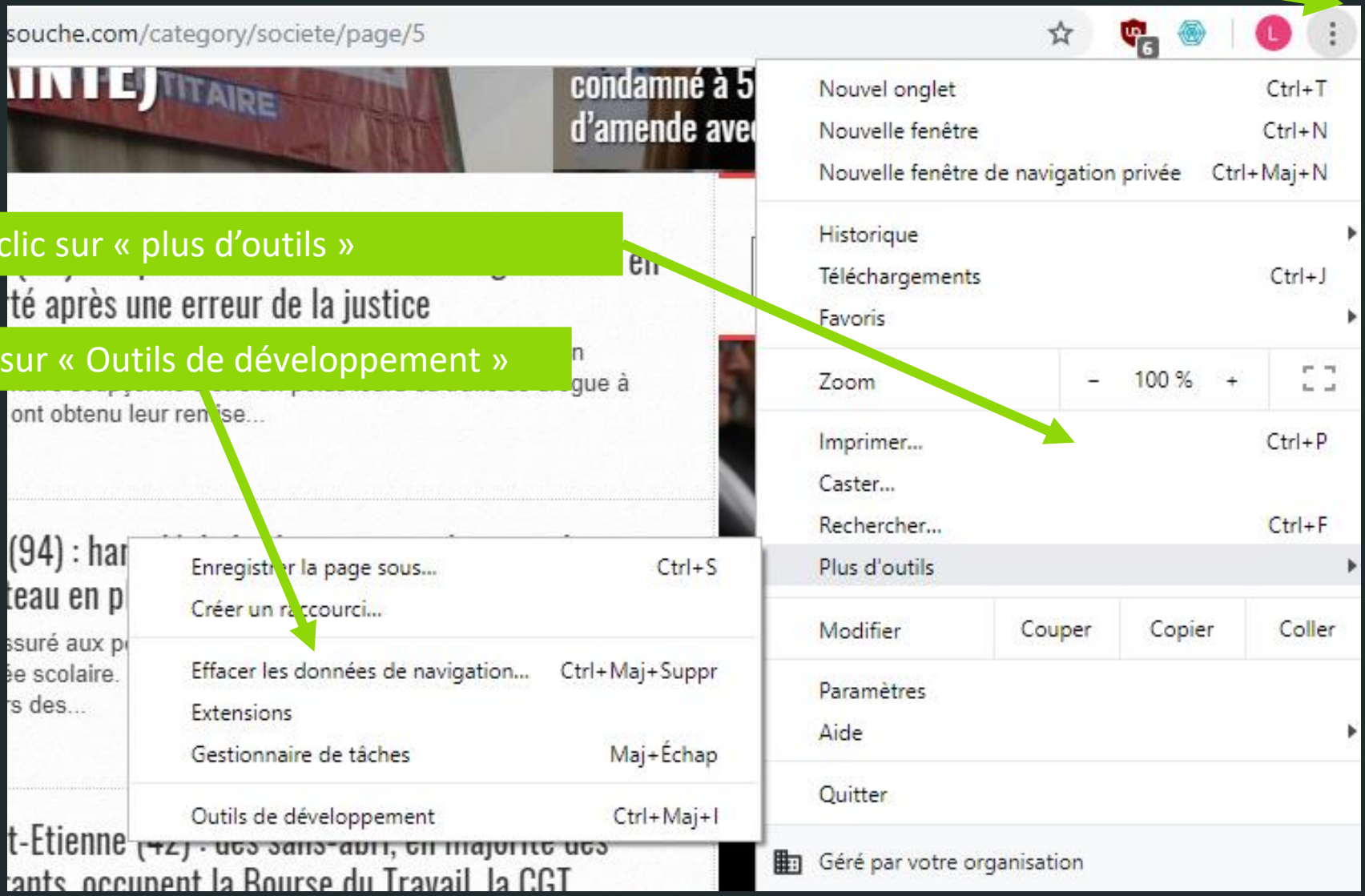
(4) J’ouvre Web Scraper

- Windows, Linux: Ctrl+Shift+I or F12
- Mac: Cmd+Opt+I
- Any OS: open Tools / Developer tools

(a) Je clic sur les points verticaux

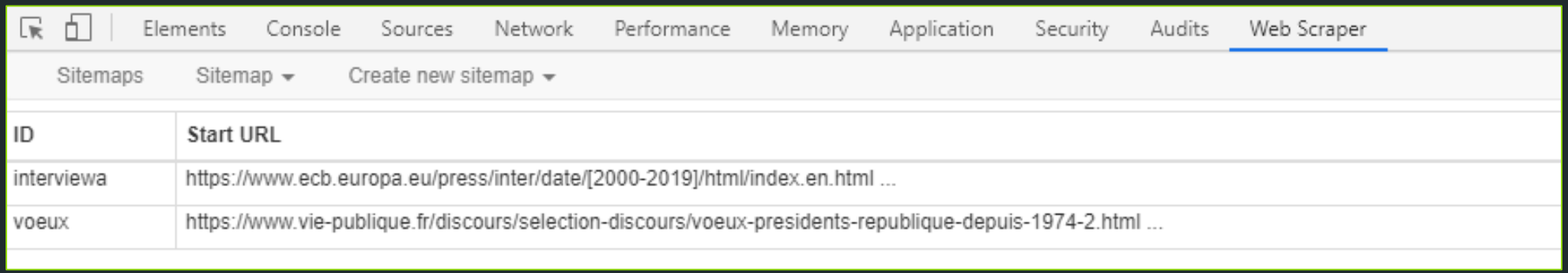
(b) Je clic sur « plus d'outils »

(c) Je clic sur « Outils de développement »

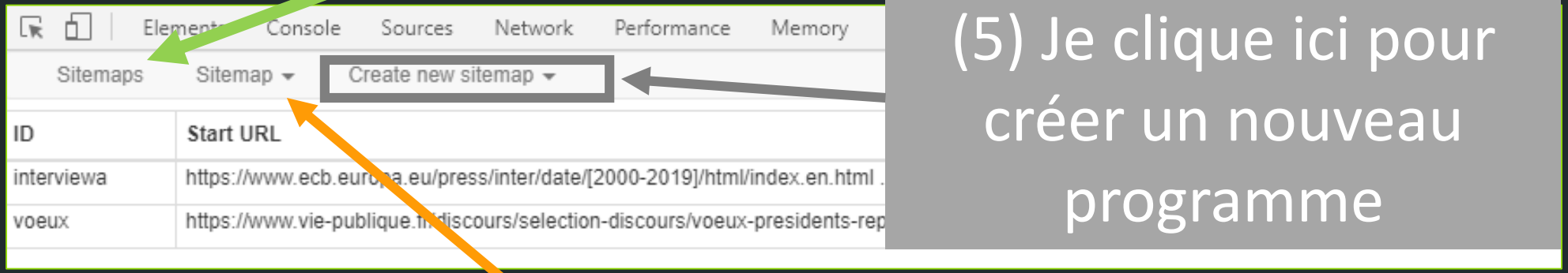


Une « fenêtre » s'ouvre en bas de votre écran

Je clic sur Web Scraper

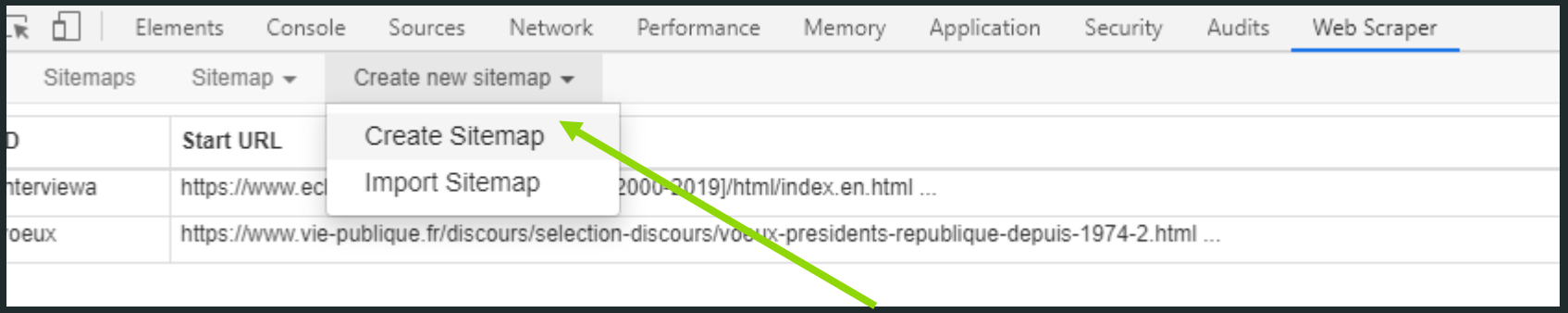


Sous l'onglet « SITEMAPS », seront rangés mes « programmes » spécifiques à un site

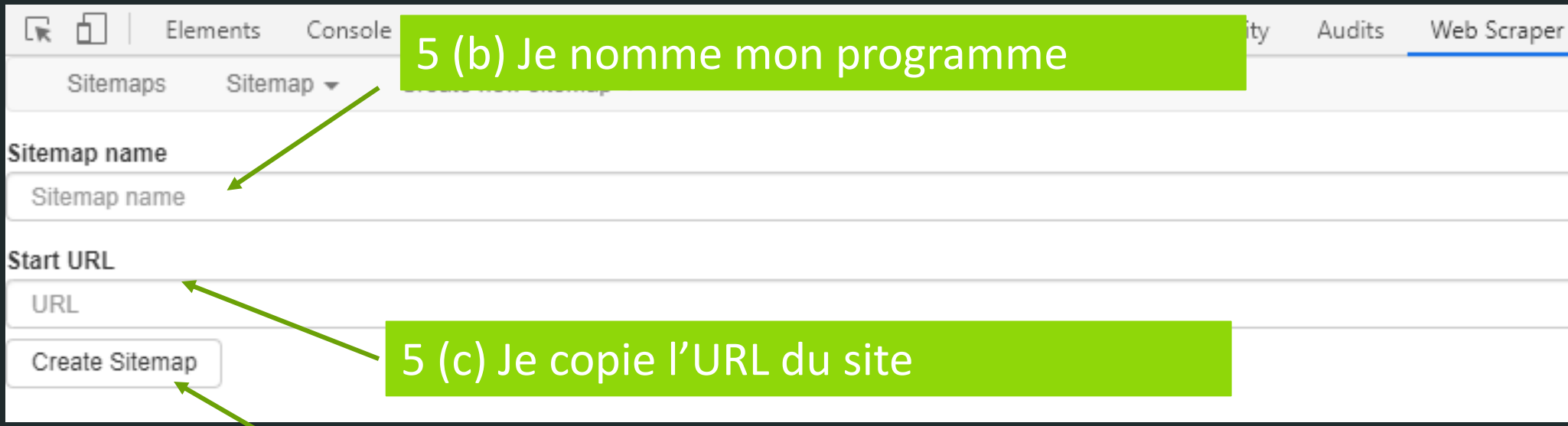


(5) Je clique ici pour créer un nouveau programme

Sous l'onglet « Sitemap » se trouvent mes commandes pour un programme ouvert



5 (a) Je clic sur « create Sitemap»

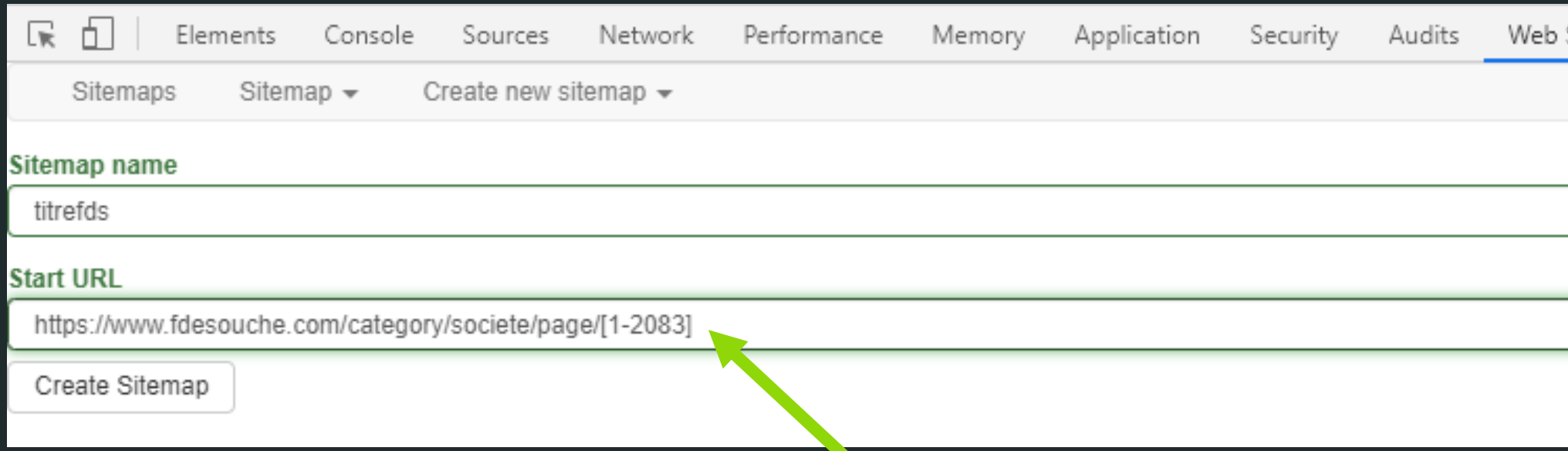


5 (b) Je nomme mon programme

5 (c) Je copie l'URL du site

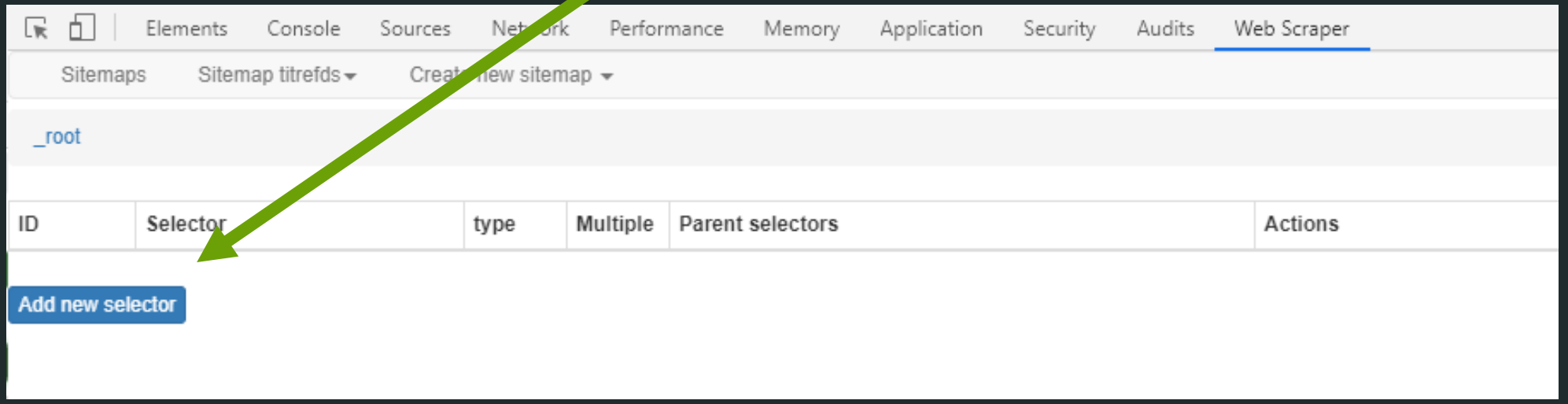
5 (d) Je valide pour créer le programme

Exemple : je veux aspirer les titres des articles de la rubrique « Société » du site. Cette rubrique s'étend sur 2083 pages.



Mon programme tournera les pages de 1 à 2083 et appliquera systématiquement mes commandes sur chaque page

(6) Je clic sur « add selectors »



(a) Je nomme l'objet à sélectionner

Sitemaps | Sitemap titrefds | Create new sitemap

Id

Selector Id

Type

Text

(b) Je choisis dans le menu déroulant le type de l'obje

Selector

Select | Element preview | Data preview

Multiple

(c) Je coche « multiple » si mon objet est présent plusieurs fo sur une page.

Exemple : sélectionner les liens des articles

Id

Titre

Type

Text

Text
Link
Popup Link
Image
Table
Element attribute
HTML
Element
Element scroll down
Element click
Grouped

0

Id

Titre

Type

Text

Selector

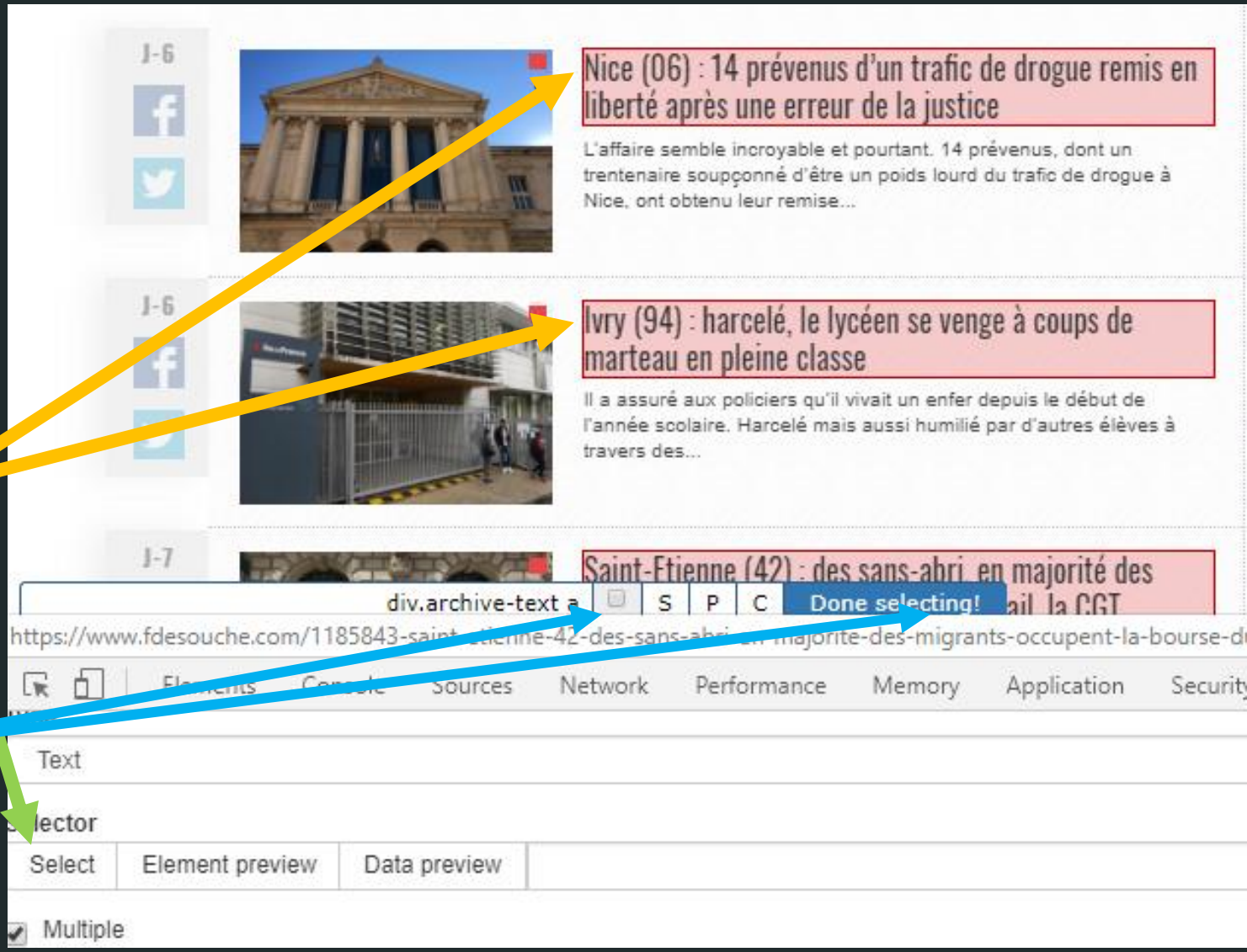
Select | Element preview | Data preview

Multiple

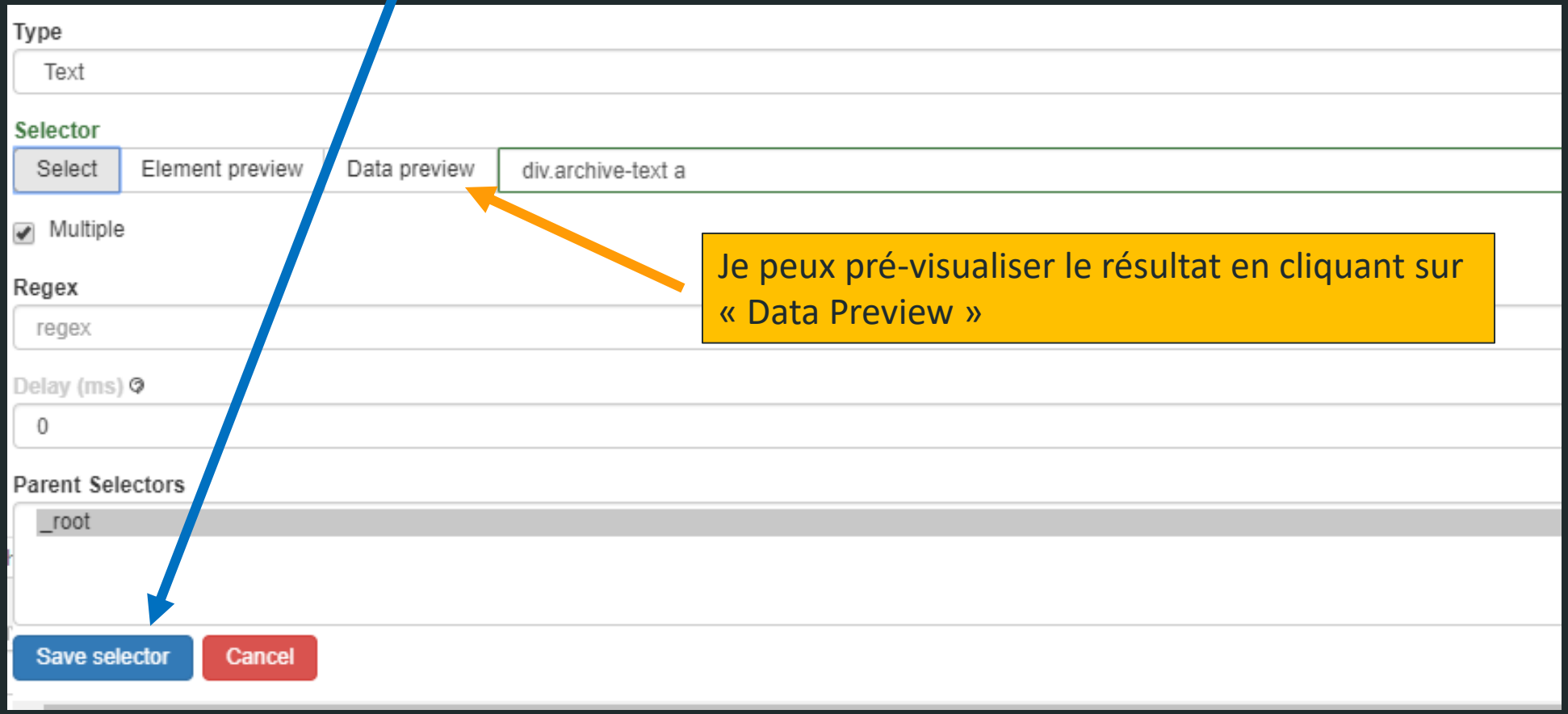
(d) Je clique sur SELECT **puis** je clique sur les TITRES du site.

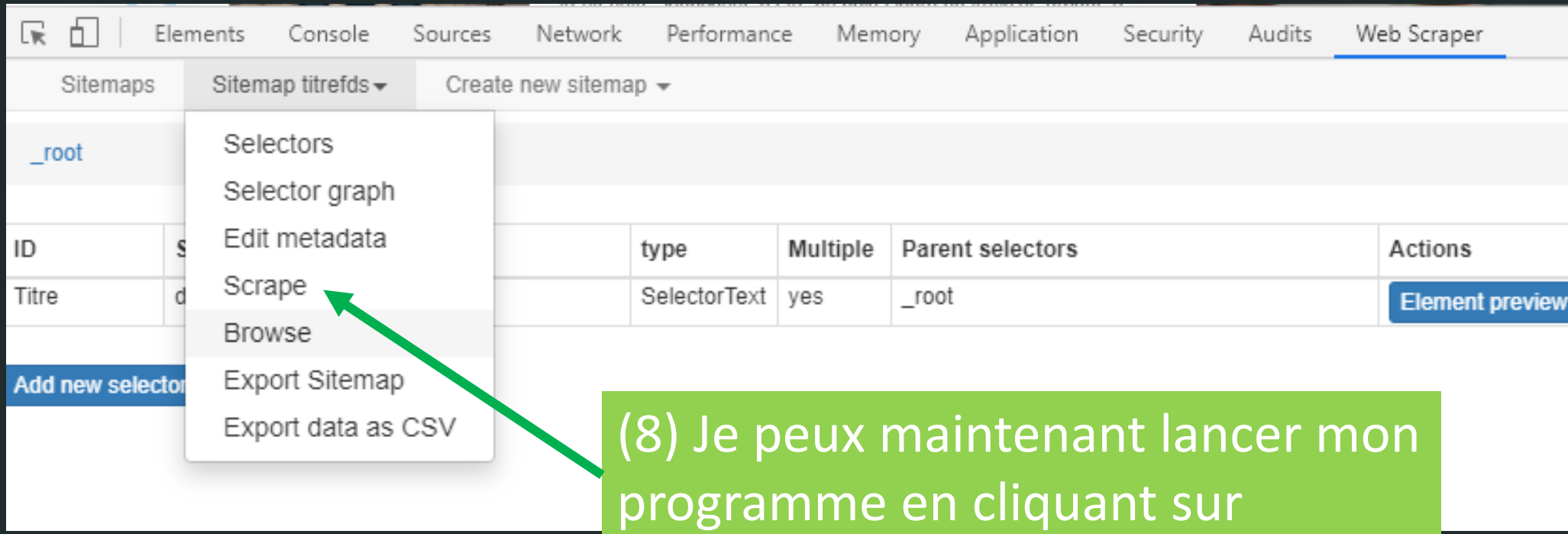
puis je clique sur les TITRES du site.

Enfin, je coche et je valide



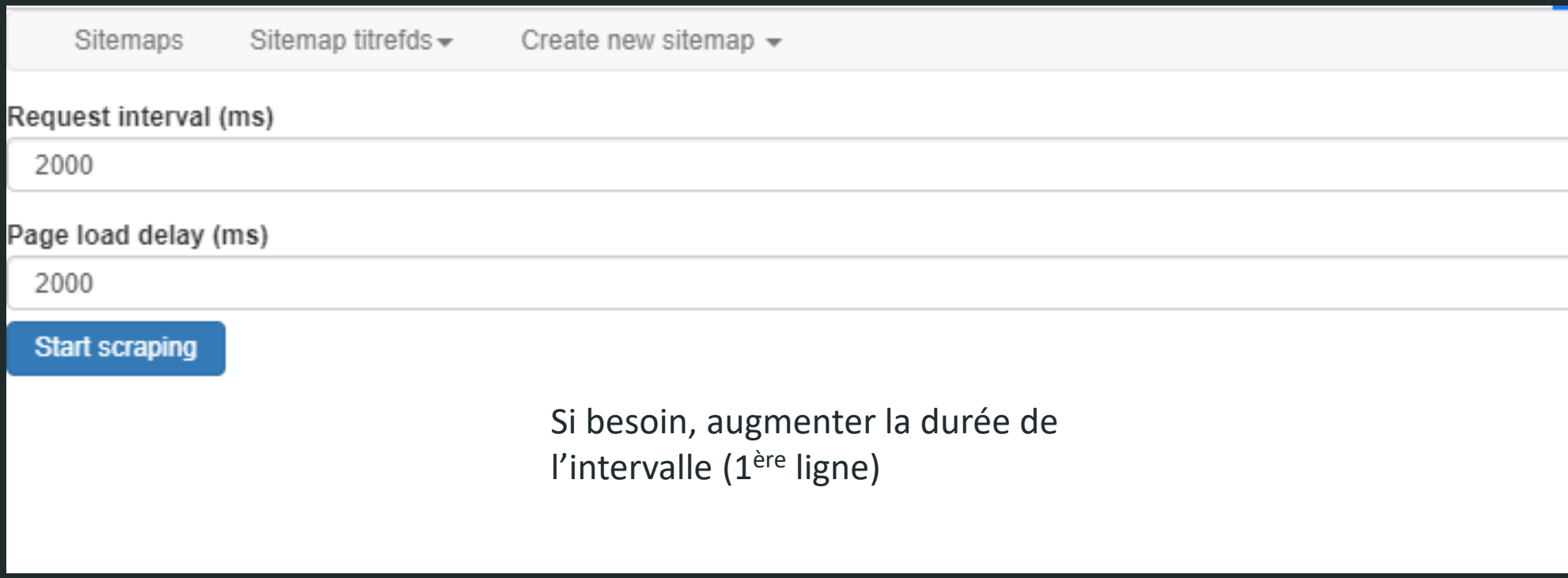
(7) Je valide en cliquant sur « Save selector »





(8) Je peux maintenant lancer mon programme en cliquant sur SCRAPE

(8) Je confirme en cliquant sur « start »



Si besoin, augmenter la durée de l'intervalle (1ère ligne)

Le fichier .csv est prêt !

Sitemaps | Sitemap titrefds | Create new sitemap

- Selectors
- Selector graph
- Edit metadata
- Scrape
- Browse
- Export Sitemap
- Export data as CSV

Refresh Data

b Scrapers User Interface (UI)? Sure!

Télécharger le fichier en cliquant sur « export data »

web-scraper-order	w	Titre
1554898321-68	ht…gory/societe/page/8	Migrants : le pape François exhorte les Romains à utiliser leurs énergies « pour transformer les tensions et les problèmes en opportunités »
1554898306-24	ht…gory/societe/page/10	Saïda se bat pour faire « rapatrier » sa petite-fille de 3 ans détenue en Irak
1554898340-128	https://www.fdesouche.com/category/societe/page/5	Ivry (94) : harcelé, le lycéen se venge à coups de marteau en pleine classe
1554898348-175	https://www.fdesouche.com/category/societe/page/4	400 nouveaux radars tourelles débarquent: pourquoi il sera bien plus difficile de leur échapper
1554898356-200	https://www.fdesouche.com/category/societe/page/3	Des « taux choquants de viols » dans les pays nordiques, dénonce Amnesty International
1554898376-232	https://www.fdesouche.com/category/societe/page/1	L'artiste toulousain Colomina expose son « migrant » à Paris
1554898321-74	https://www.fdesouche.com/category/societe/page/8	Abdelkader Merah explique que ses parents n'ont pas été très vigilants sur la scolarité de leurs enfants: « C'est dans la culture des Arabes »

Vue sur le fichier texte après sélection de la colonne titre

"Le prophète lui a dit que les Roms travaillent pour les envoyés du diable et eux ils mangent les enfants »"

"Prostitution à Rome : « Les prêtres ne veulent pas d'Italiens (…) ils veulent des migrants, c'est plus discret »
(Sodoma)"

"Saïda se bat pour faire « rapatrier » sa petite-fille de 3 ans détenue en Irak"

"Le Royaume-Uni refuse l'asile à un Iranien converti au christianisme sous prétexte que la Bible contient des passages violents"

"Gilets jaune : Maxime Nicolle convoqué par la police, Eric Drouet entendu en audition libre mercredi",

Nouvel exemple : je souhaite télécharger tous les articles de la rubrique « société » du site FDSOUCHE

(1) J’analyse le site…

POLITIQUE ECONOMIE MONDE ECOLOGIE CULTURE SPORT

« Comportements inappropriés » à francetvinfo.fr : l’école de journalisme de Sciences Po va ouvrir une enquête interne

L’école de journalisme de Sciences Po a annoncé samedi 6 avril, sur sa page Facebook, l’ouverture d’une enquête interne. Une décision qui fait suite aux...

Vannes (56) : elle devait s’occuper d’un cochon sauvé par une association, elle en fait du pâté

Babe avait évité l’abattoir de justesse, mais il n’a pas échappé à un destin tragique. Pour avoir réalisé des conserves avec la viande d’un...

Ennery (95) : un projet « secret » de centre d’accueil provoque la colère des habitants

L’association Espérer 95 projette à Ennery une résidence d’accueil pour 80 personnes en situation de précarité. Le village a contesté fortement ce projet lors...

https://www.fdesouche.com/1188177-vannes-elle-devait-soccuper-dun-cochon-sauve-par-une-ass...

SE CONNECTER PROPOSER UN ARTICLE CONNECTÉS: 1620 VISITEURS UNIQUES:

FDESOUCHE.COM EST UNE REVUE DE PRESSE

ACCUEIL SOCIÉTÉ SÉCURITÉ POLITIQUE ECONOMIE MONDE ECOLOGIE CULTURE SPORT INSOLITE FDÉSINTOX

Accueil / Insolite / Vannes (56) : elle devait s’occuper d’un cochon sauvé par une association, elle en fait du pâté

Vannes (56) : elle devait s’occuper d’un cochon sauvé par une association, elle en fait du pâté

Par Francois le 07/04/2019

(1) J’analyse l’article

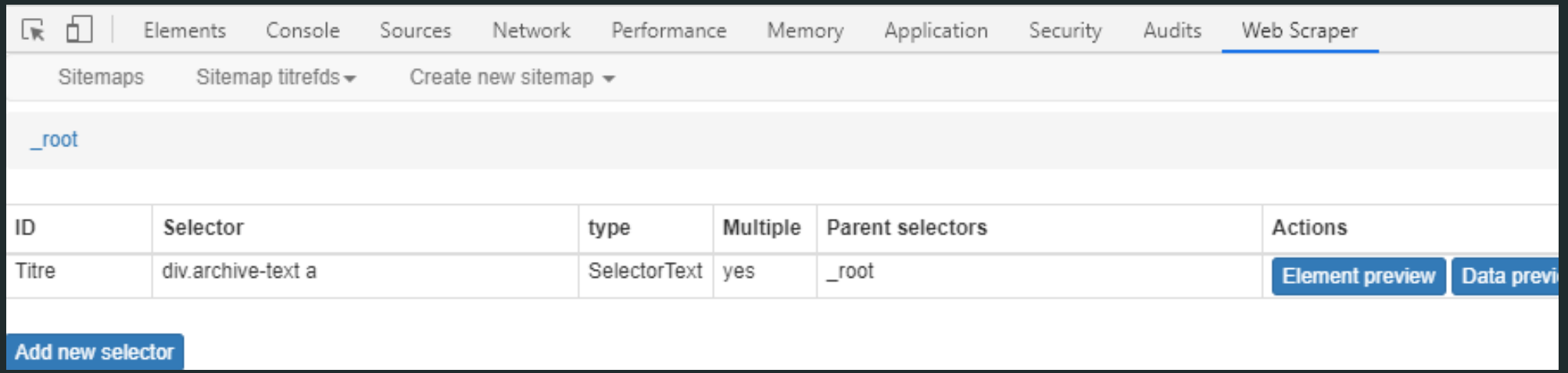
Babe avait évité l'abattoir de justesse, mais il n'a pas échappé à un destin tragique. Pour avoir réalisé des conserves avec la viande d'un cochon qu'elle était censée garder pour le compte d'une association de défense des animaux, une femme de 40 ans a été condamnée, à Vannes (Morbihan), à trois mois de prison avec sursis et 500 euros d'amende, a rapporté *Le Télégramme*, jeudi 4 avril.

Elle était poursuivie pour « abus de confiance et complicité d'abattage hors d'un abattoir dans des conditions illicites ». Le tribunal a par ailleurs fait état d'un risque sanitaire en cas de vente du pâté.

Il s’agit de TEXTE

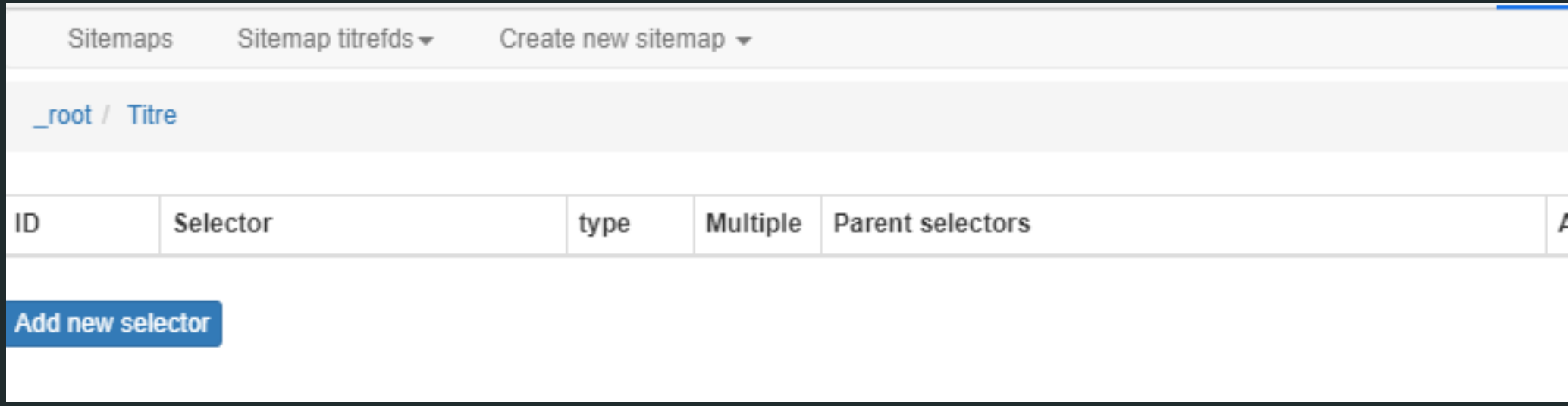
Plusieurs paragraphes par page
Plusieurs niveaux de mise en forme…

2) Je réitère la procédure de l'exemple précédent jusqu'à l'étape 7



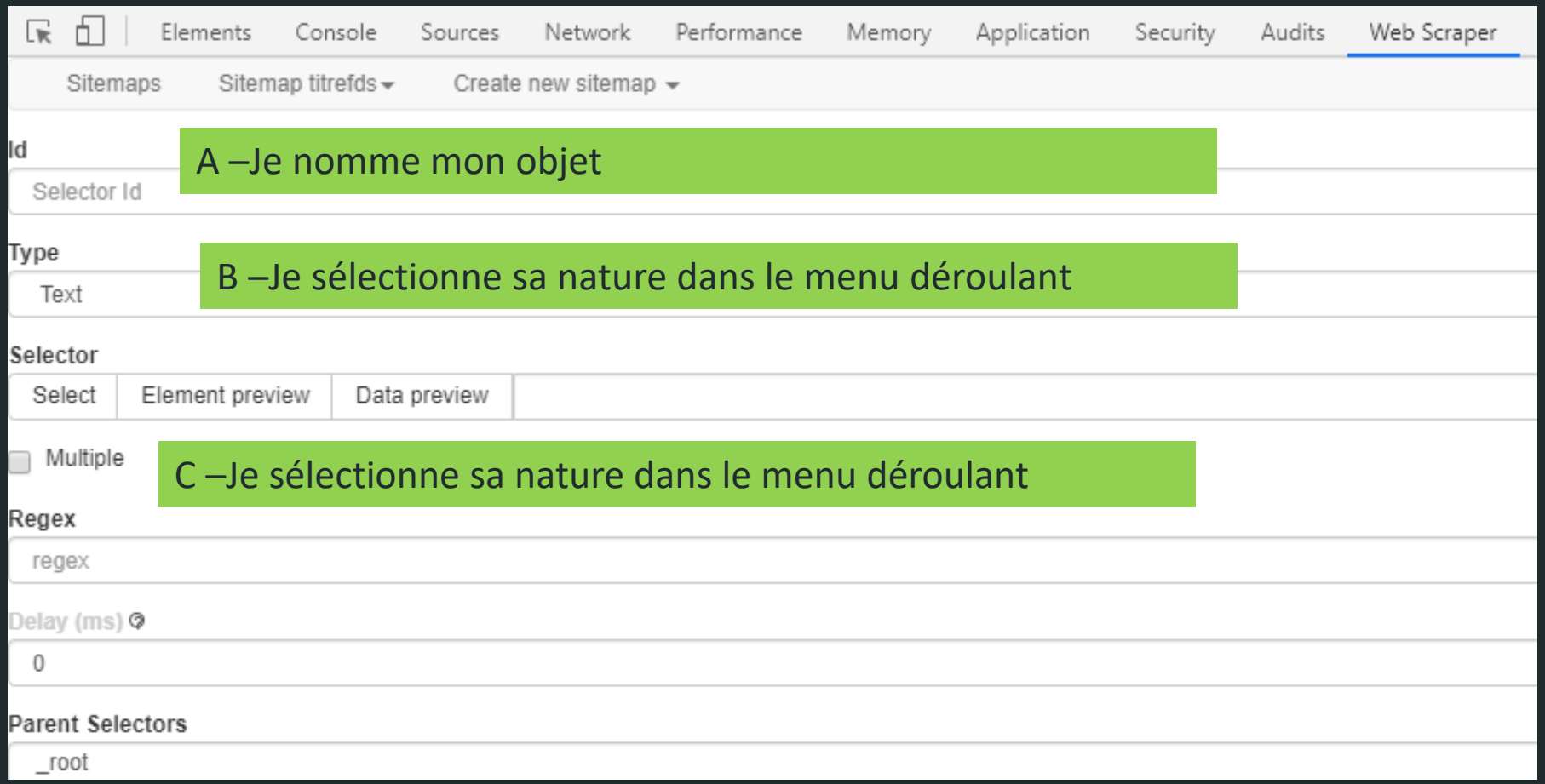
3) Je clique sur titre

Ma requête doit dépendre de celle formulée sur le niveau TITRE…

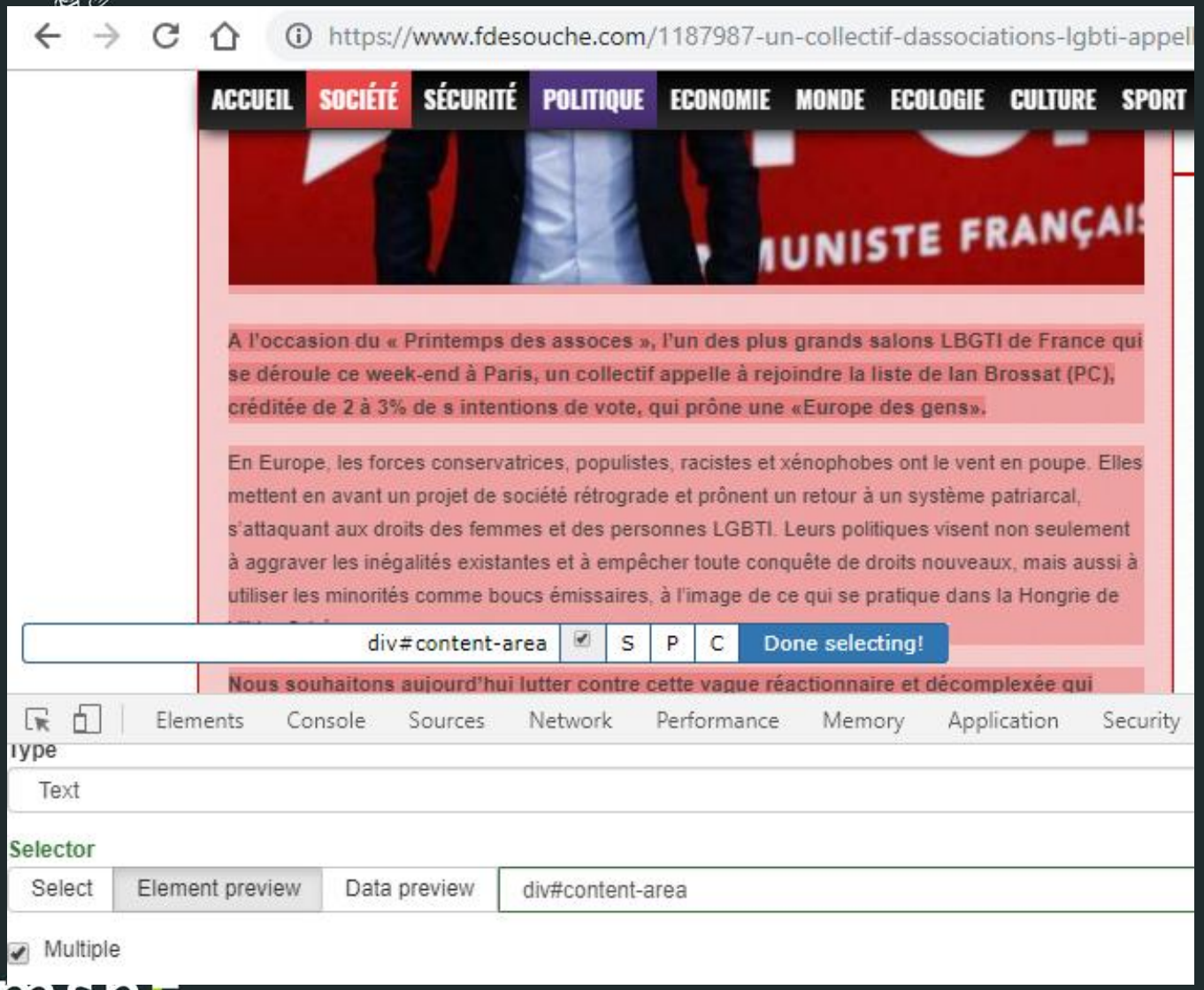


(4) Je clique sur « Add new selector »

(5) Je remplis le formulaire



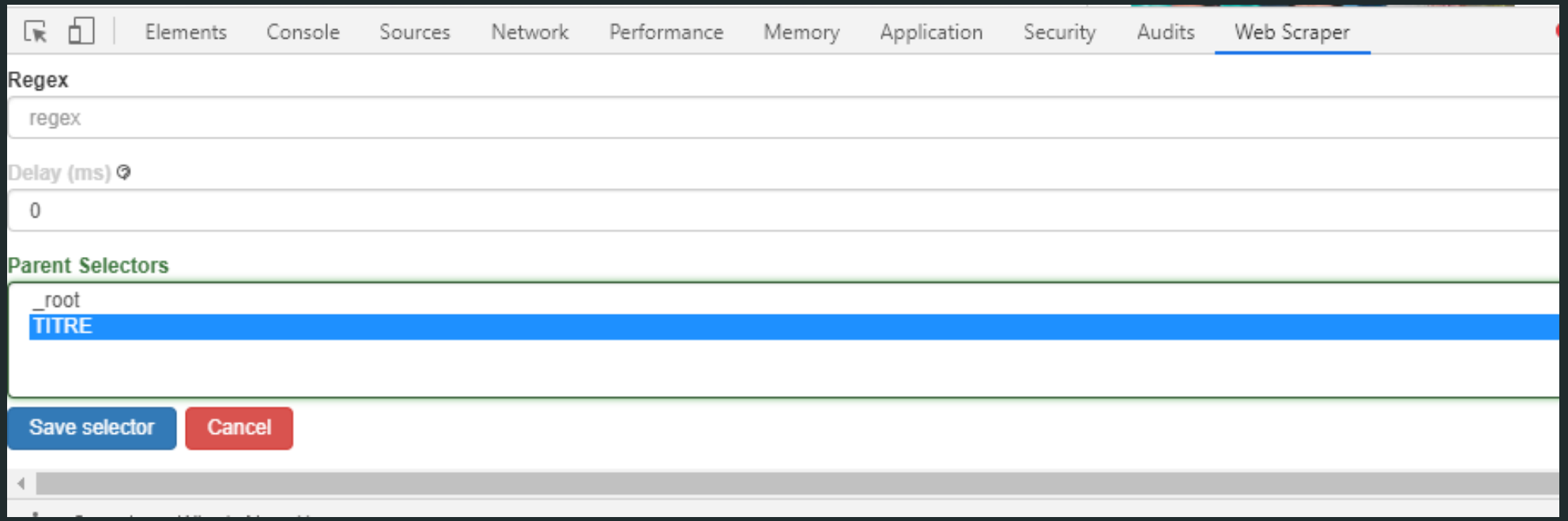
(6) Je sélectionne le texte



Je coche la validité de la structuration et je clique sur « Done selecting »

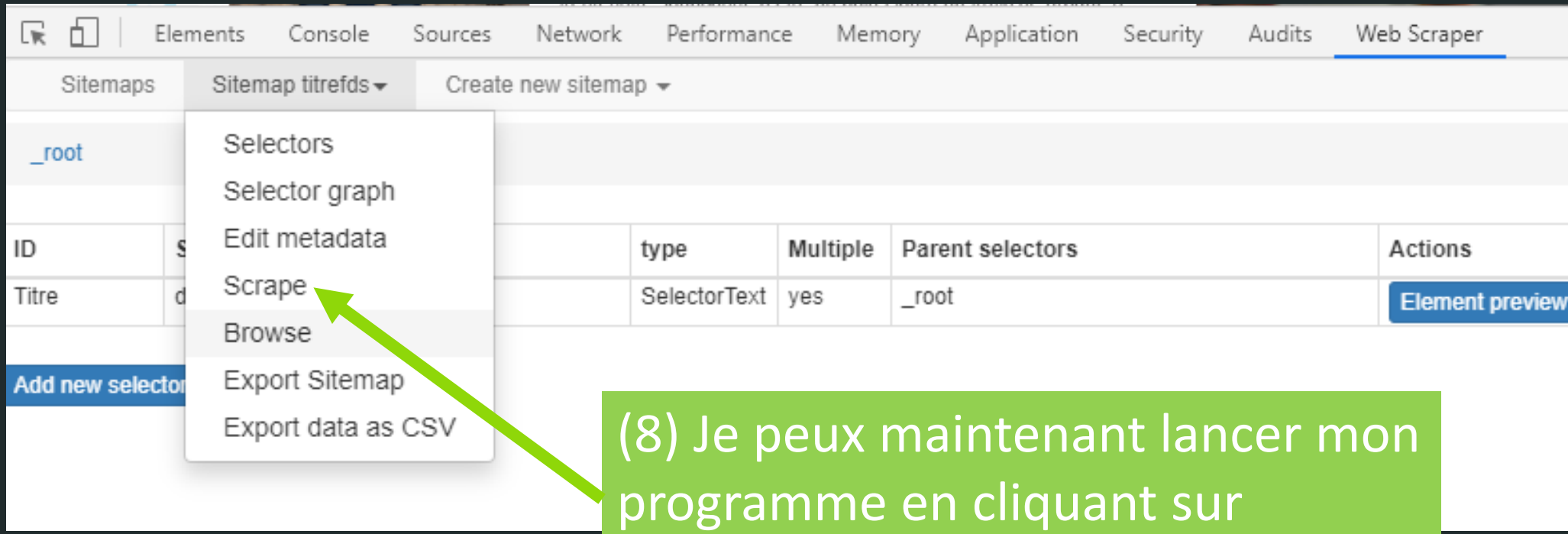
Je prévisualise le résultat

(7) Je choisis le sélecteur dont dépend ma requête



Ici, je choisis TITRE puis je confirme en cliquant sur « Save Selector »

(8) Je lance le scrap



(8) Je peux maintenant lancer mon programme en cliquant sur SCRAPE

C’est prêt !

Elements Console Sources Network Performance Memory Application Security Audits Web Scraper 16 1

				The Local Education : Depuis trente ans, le niveau de math des élèves de CM2 est en baisse, selon une étude L'association openDemocracy accuse les fondamentalistes chrétiens américains d'avoir financé l'extrême droite en Europe
900507-	https://www.fdesouche.com/category/societe/page/5	Beauvais (60) : un Afghan condamné pour avoir harcelé sexuellement des jeunes filles et agressé la gérante d’un bar	https://www.fdesouche.com/1185437-beauvais-un-afghan-condamne-pour-avoir-harcele-sexuellement-des-jeunes-filles-et-agresse-la-gerante-dun-bar-le-migrant-est-en-france-que-depuis-un-mois	Le tribunal de Beauvais a condamné ce lundi un homme pour violences et outrage sexiste. Cette dernière infraction, entrée en vigueur l’été dernier, se caractérise par le fait d’imposer à une personne tout propos ou comportement à connotation sexuelle ou sexiste qui porte atteinte à sa dignité. En clair, c’est ce qu’on appelle le harcèlement de rue, qui a été décrit par deux jeunes Beauvaisiennes devant le tribunal. Vendredi soir, elles sortaient d’un bar du centre-ville lorsqu’elles ont été interpellées par un homme de 28 ans, complètement

web-scraper-order,web-scraper-start-url,TITRE,TITRE-href,Article"1554900661
-361","https://www.fdesouche.com/category/societe/page/1","La Chine a rasé
plusieurs grandes mosquées en région ouïgoure, montrent des images
satellites","https://www.fdesouche.com/1188803-la-chine-a-detruit-
plusieurs-grandes-mosquees-en-region-ouigoure-montrent-des-images-
satellites",

"« Où est passée la mosquée ? » Shawn Zhang s'interroge. Cet activiste, étudiant à l'université de Vancouver au Canada, a pris l'habitude de dénoncer sur Twitter les exactions du régime chinois, notamment au Tibet e dans la région du Xinjiang, habitée par la minorité musulmane ouïgoure. En observant des images satellites sur Google Earth Pro, le jeune homme est arrivé à la conclusion que la grande mosquée Aitika de Keriya, une ville d 30.000 habitants dans cette province de l'ouest de la Chine, avait tout simplement disparu, au printemps 2018.Les spéculations autour du sort de l grande mosquée ont rapidement germé sur internet, d'autant que ce monument du XIIIe siècle avait été classé au patrimoine culturel national chinois : toute destruction ne pouvait être décidée que par le pouvoir central à Pékin.Comme le rapporte « Libération », l'étudiant d'origine chinoise a comparé la même vue satellite à différentes périodes, pour arriver à la conclusion que la mosquée avait été rasée entre mars et mai 2018. D'autres services de cartographie, tels que HERE, Planet Labs ou TerraServer, confirment la disparition du monument, selon « Libération » et l'associati

Pour aller plus loin et vous approprier l'outil…

- https://www.webscraper.io/documentation
- https://www.webscraper.io/tutorials

Pour un exemple sur une récupération d'adresses (tuto en langue anglaise) : https://www.youtube.com/watch?v=ZdOrH50WqEo