

Data-journalisme

MMI 2 – TP#2 S4



Danielo **JEAN-LOUIS**

Data-journaliste

- Personne *pluridisciplinaire*
 - Développement
 - Graphisme
 - Rédaction
 - Experte dans un domaine
 - **Journaliste**

Le data-journaliste interroge la donnée, il peut la trouver

- Rapports
- Fuite de données
 - Pandora papers, LuxLeaks...
- Données ouvertes
- API / Bases de données

**Et s'il n'y a rien de tout ça,
il existe le web-scaping**

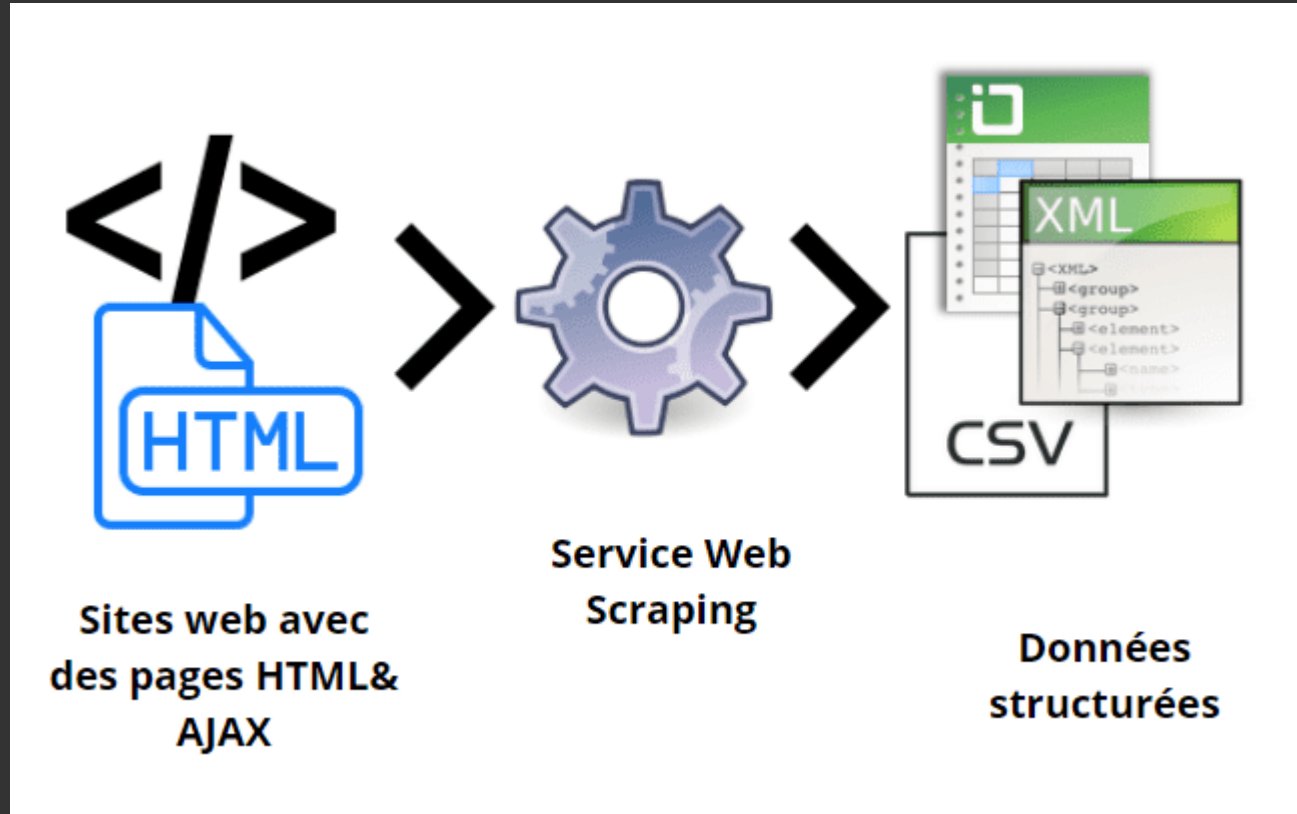
Web-scrapping

- Technique d'extraction **automatisée** du contenu de sites web (n'importe quel site)
- Permet de remplacer, de faire sa propre API / Base de données

Web-scraping

- Transforme des données de sites web en données structurées
 - Permet une exploitation plus simple
- Nécessite des connaissances en programmation
 - Connaissance du DOM / Sélecteurs HTML
 - Python / Javascript et d'outils dédiés

Web-scraping



Source(s) :

- <https://superdatacamp.com/big-data/web-scraping/>

Web-scrapping – Outils

- Python
 - Selenium
 - Puissant fait fonctionner un navigateur en mode "headless"
 - BeautifulSoup
 - Léger mais limité, ne permet pas de lire les éléments HTML asynchrones

Web-scrapping – Outils

- Javascript – Nécessite Nodejs
 - Selenium
 - puppeteer

Web-scrapping – Précautions

- Scrappez de façon éthique
 - Évitez de le faire de façon intensive
 - Instaurez un délai entre chaque "scrap"
 - Limite les risques de se faire IP ban
 - Vérifiez le robots.txt du site pour connaître le délai autorisé entre chaque crawl
 - {domaine}.{tld}/robots.txt pour accéder au fichier

Web-scrapping – Précautions

- Scrappez de façon maligne
 - Changez de DNS entre chaque "scrap"
 - Limite les risques de se faire bannir son adresse IP

Pratiquons ! - Web-scraping

Pré-requis :

- Avoir la ressource `ressources/web-scraping`
- A télécharger ici :

