

Data-journalisme

MMI 2 – TP#2 S4

Danielo JEAN-LOUIS

Data-journaliste

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

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-scraping**

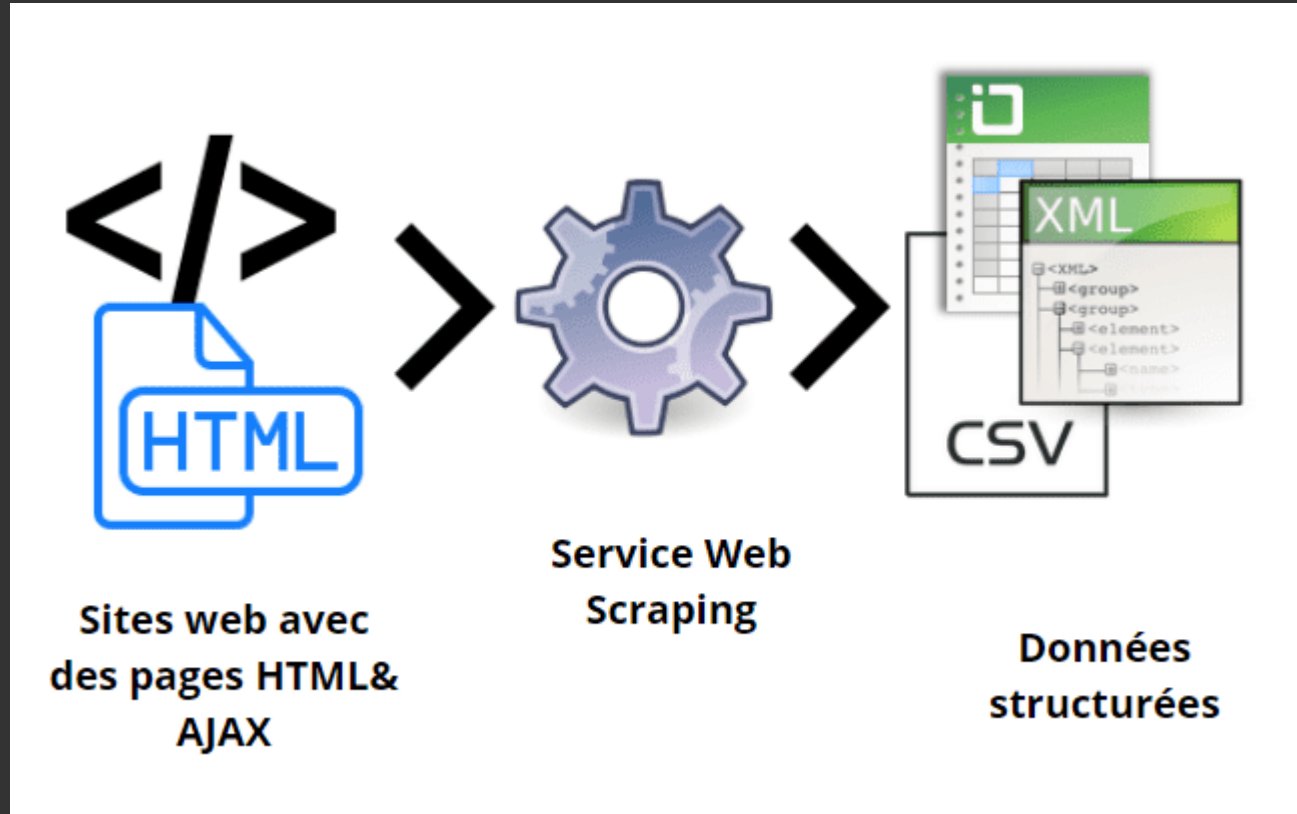
Web-scrapping

- Technique d'extraction **automatisée** du contenu de sites web (n'importe quel site)
- Permet 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 - Schéma



Source(s) :

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

Web-scraping – 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-scraping – Outils

- Javascript – Nécessite Nodejs
 - Selenium
 - puppeteer

Web-scraping – Précautions

- Scrapez 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-scraping – Précautions

- En absence de délai entre les crawls mettez +~30 secondes
- Scrapez 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 :

<https://download-directory.github.io/?url=https%3A%2F%2Fgithub.com%2FDanYellow%2Fcours%2Ftree%2Fmain%2Fdata-journalisme-s4%2Ftravaux-pratiques%2Fnumero-2%2Fressources%2Fweb-scraping>

Web-scraping

- Google Sheets permet de faire du scraping
 - Assez limité en terme de scrap
- Autres outils (freemium) – liste non exhaustive :
 - <https://phantombuster.com/>
 - <https://www.octoparse.com/>

Source(s) :

- <https://www.octoparse.com/blog/simple-web-scraping-using-google-sheets> – anglais
- <https://www.youtube.com/watch?v=WxKuWOPcC70>
- <https://support.google.com/docs/answer/3093339?hl=en> - anglais
- <https://www.youtube.com/watch?v=-nSEDTklvkQ> - Didacticiel sur le scraping sans code

Data-journalisme

MMI 2 – TP#2 S4





Danielo **JEAN-LOUIS**

Data-journaliste

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

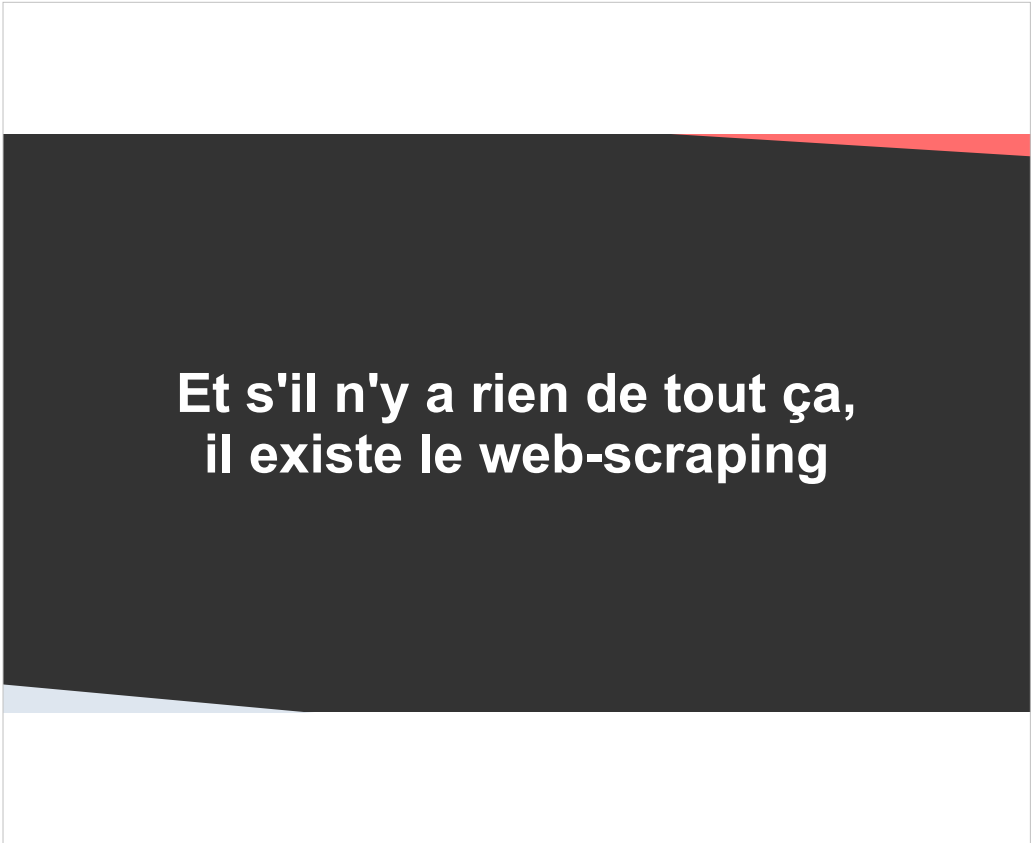
- On retrouve Apple, Netflix ou encore Microsoft au W3C. Netflix va notamment contribuer à une technologie (<https://www.w3.org/TR/media-source/>) utilisée dans l'épisode Bandersnatch de Black Mirror
- Le W3C définit également les standards du javascript ou du CSS
- Face à la lenteur du W3C, Apple (Safari), Mozilla (Firefox) et Opera (Opera) forment le WHATWG (Web Hypertext Application Technology Working Group). L'idée étant de définir ensemble les standards du web.

-

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

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

- On retrouve Apple, Netflix ou encore Microsoft au W3C. Netflix va notamment contribuer à une technologie (<https://www.w3.org/TR/media-source/>) utilisée dans l'épisode Bandersnatch de Black Mirror
- Le W3C définit également les standards du javascript ou du CSS
- Face à la lenteur du W3C, Apple (Safari), Mozilla (Firefox) et Opera (Opera) forment le WHATWG (Web Hypertext Application Technology Working Group). L'idée étant de définir ensemble les standards du web.
-



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

- Tim Berners-Lee a également inventé :
 - le www (world wide web) ou web, système incluant :
 - adresses web
 - protocole HTTP (communication client-serveur)
 - Le HTML
 - Il invente également le W3C
- Le HTML n'est pas un langage de programmation. Il ne permet pas de mettre en place des conditions (if, else, boucle while...)

Web-scraping

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

- On retrouve Apple, Netflix ou encore Microsoft au W3C. Netflix va notamment contribuer à une technologie (<https://www.w3.org/TR/media-source/>) utilisée dans l'épisode Bandersnatch de Black Mirror
- Le W3C définit également les standards du javascript ou du CSS
- Face à la lenteur du W3C, Apple (Safari), Mozilla (Firefox) et Opera (Opera) forment le WHATWG (Web Hypertext Application Technology Working Group). L'idée étant de définir ensemble les standards du web.

-

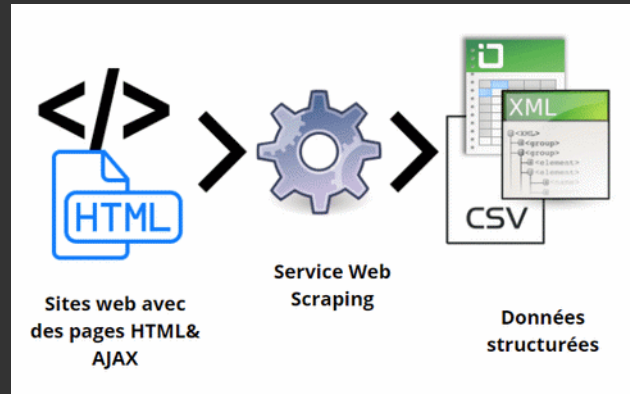
Web-scraping

- 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

- On retrouve Apple, Netflix ou encore Microsoft au W3C. Netflix va notamment contribuer à une technologie (<https://www.w3.org/TR/media-source/>) utilisée dans l'épisode Bandersnatch de Black Mirror
- Le W3C définit également les standards du javascript ou du CSS
- Face à la lenteur du W3C, Apple (Safari), Mozilla (Firefox) et Opera (Opera) forment le WHATWG (Web Hypertext Application Technology Working Group). L'idée étant de définir ensemble les standards du web.

-

Web-scraping - Schéma



Source(s) :

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

- On retrouve Apple, Netflix ou encore Microsoft au W3C. Netflix va notamment contribuer à une technologie (<https://www.w3.org/TR/media-source/>) utilisée dans l'épisode Bandersnatch de Black Mirror
- Le W3C définit également les standards du javascript ou du CSS
- Face à la lenteur du W3C, Apple (Safari), Mozilla (Firefox) et Opera (Opera) forment le WHATWG (Web Hypertext Application Technology Working Group). L'idée étant de définir ensemble les standards du web.

-

Web-scraping – Outils

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

- On retrouve Apple, Netflix ou encore Microsoft au W3C. Netflix va notamment contribuer à une technologie (<https://www.w3.org/TR/media-source/>) utilisée dans l'épisode Bandersnatch de Black Mirror
- Le W3C définit également les standards du javascript ou du CSS
- Face à la lenteur du W3C, Apple (Safari), Mozilla (Firefox) et Opera (Opera) forment le WHATWG (Web Hypertext Application Technology Working Group). L'idée étant de définir ensemble les standards du web.
-

Web-scraping – Outils

- Javascript – Nécessite Nodejs
 - Selenium
 - puppeteer

- On retrouve Apple, Netflix ou encore Microsoft au W3C. Netflix va notamment contribuer à une technologie (<https://www.w3.org/TR/media-source/>) utilisée dans l'épisode Bandersnatch de Black Mirror
- Le W3C définit également les standards du javascript ou du CSS
- Face à la lenteur du W3C, Apple (Safari), Mozilla (Firefox) et Opera (Opera) forment le WHATWG (Web Hypertext Application Technology Working Group). L'idée étant de définir ensemble les standards du web.
-

Web-scraping – Précautions

- Scraperez 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

- On retrouve Apple, Netflix ou encore Microsoft au W3C. Netflix va notamment contribuer à une technologie (<https://www.w3.org/TR/media-source/>) utilisée dans l'épisode Bandersnatch de Black Mirror
- Le W3C définit également les standards du javascript ou du CSS
- Face à la lenteur du W3C, Apple (Safari), Mozilla (Firefox) et Opera (Opera) forment le WHATWG (Web Hypertext Application Technology Working Group). L'idée étant de définir ensemble les standards du web.

-

Web-scraping – Précautions

- En absence de délai entre les crawls mettez +~30 secondes
- Scrapez de façon maligne
 - Changez de DNS entre chaque "scrap"
 - Limite les risques de se faire bannir son adresse IP

- On retrouve Apple, Netflix ou encore Microsoft au W3C. Netflix va notamment contribuer à une technologie (<https://www.w3.org/TR/media-source/>) utilisée dans l'épisode Bandersnatch de Black Mirror
- Le W3C définit également les standards du javascript ou du CSS
- Face à la lenteur du W3C, Apple (Safari), Mozilla (Firefox) et Opera (Opera) forment le WHATWG (Web Hypertext Application Technology Working Group). L'idée étant de définir ensemble les standards du web.

-

Pratiquons ! - Web-scraping

Pré-requis :

- Avoir la ressource [ressources/web-scraping](#)
- A télécharger ici :
<https://download-directory.github.io/?url=https%3A%2F%2Fgithub.com%2FDanYellow%2Fcours%2Ftree%2Fmain%2Fdata-journalisme-s4%2Ftravaux-pratiques%2Fnumero-2%2Fressources%2Fweb-scraping>

Web-scraping

- Google Sheets permet de faire du scraping
 - Assez limité en terme de scrap
- Autres outils (freemium) – liste non exhaustive :
 - <https://phantombuster.com/>
 - <https://www.octoparse.com/>

Source(s) :

- <https://www.octoparse.com/blog/simple-web-scraping-using-google-sheets> – anglais
- <https://www.youtube.com/watch?v=WxKuWOPcC70>
- <https://support.google.com/docs/answer/3093339?hl=en> - anglais
- <https://www.youtube.com/watch?v=nSEDTklvKQ> - Didacticiel sur le scraping sans code

- On retrouve Apple, Netflix ou encore Microsoft au W3C. Netflix va notamment contribuer à une technologie (<https://www.w3.org/TR/media-source/>) utilisée dans l'épisode Bandersnatch de Black Mirror
- Le W3C définit également les standards du javascript ou du CSS
- Face à la lenteur du W3C, Apple (Safari), Mozilla (Firefox) et Opera (Opera) forment le WHATWG (Web Hypertext Application Technology Working Group). L'idée étant de définir ensemble les standards du web.

-

