Pratiques de plus en plus utilisées par les entreprises et les indépendants, le **web scraping**et le **data scraping** permettent de **collecter des données** pour une infinité de finalités possibles telles que la constitution ou l’enrichissement de bases de données de prospection ou de produits par exemple.

Il existe pourtant une **confusion commune**entre les termes **web scraping**et **data scraping**. Voyons ensemble la **distinction** avant d’en apprendre plus sur le **web scraping**.

Avant tout, clarifions immédiatement une **erreur très commune** chez les francophones (et même à travers le monde) : on parle de **scraping** (avec un “p”) et non de **scrapping**(avec deux “p”). Le terme correct, “**Scraping**”, vient du verbe anglais to scrape qui signifie l’acte de gratter, racler voire érafler une surface. Le mauvais terme, à ne pas utiliser, “scrapping” vient du verbe to scrap qui signifie plutôt “se débarrasser de quelque chose”. **Ne faites plus l’erreur !**

**Définition du Data Scraping**

Le **data scraping** (ou capture, collecte ou extraction de données en français) est une technique par laquelle un **programme informatique extrait des informations depuis une source lisible par un humain et produite par un autre programme informatique**.

Le **data scraping** permet d’extraire des données et de les structurer depuis une source prévue à l’origine pour être**lue par un humain** et où l’information n’est donc pas structurée, pas documentée, et pas optimisée pour être extraite facilement. En général, le **data scraping**est utilisé pour interfacer un programme avec un programme plus ancien qui ne propose pas d’API ou pour **extraire des données depuis une source tierce**, parfois sans son accord.

## Web scraping : définition

Il existe **plusieurs types de data scraping** parmi lesquels le **screen scraping**(extraction des données depuis un écran), le **report mining** (extraction des données depuis un rapport en fichier texte) et le plus répandu : le **web scraping**.

On l’aura deviné, le **web scraping est la technique permettant d’extraire des informations depuis des pages web,**via le**protocole HTTP**ou depuis un **navigateur internet**. Bien qu’il puisse être fait à la main, le web scraping tire évidemment tout son intérêt dans le fait qu’il est **automatisable**.

Tout programme de web scraping suit toujours **trois étapes importantes**:

* Le **Fetching** (ou l’étape de **requête**) : il s’agit tout simplement du téléchargement de la page en vue de son analyse. La façon dont notre programme va **parcourir**le site visé (**crawling**), stocker les URL à traiter et les appeler pour le fetching est donc primordiale.
* Le **Parsing**(ou l’étape de **traitement**) : une fois la page téléchargée, notre programme a accès à son contenu (code HTML) et peut le **parser**, c’est-à-dire le traiter pour n’extraire que les données voulues (par exemple : un nom, un prix, une URL…). On utilisera pour cela des **sélecteurs CSS ou XPath** permettant de sélectionner un élément bien précis du code.
* Le **stockage** : une fois les données récupérées, notre programme doit les structurer pour les exporter et les stocker au format de notre choix comme un **tableau clé-valeur**(par exemple un fichier JSON) ou une **base de données**.

Le monde du développement web évolue très vite et les sites web font aujourd’hui de plus en plus appel à des APIs pour dynamiser l’expérience utilisateur et ajouter des informations sur les pages web de **façon asynchrone** via des appels **AJAX**. Les techniques modernes de web scraping tendent à extraire des données depuis ces appels et leurs réponses : le but est de récupérer les appels API faites par une page et de les reproduire en dehors de cette page. C’est pourquoi, en pratique, les termes **data scraping**et **web scraping** peuvent aujourd’hui facilement s’interchanger.

La pratique s’étant fortement démocratisée, de nombreuses solutions “clé en main” existent aujourd’hui pour faire du **web scraping**. Souvent très coûteuses et peu flexibles, ces solutions sont loin des possibilités offertes par les technologies de web scraping telles que les **frameworks**et les **librairies** disponibles dans de nombreux langages. Découvrez la suite de notre série sur le web scraping avec notre article dédié : [**Quels langages et technologies choisir pour le web scraping ?**](https://practicalprogramming.fr/web-scraping-outils/)