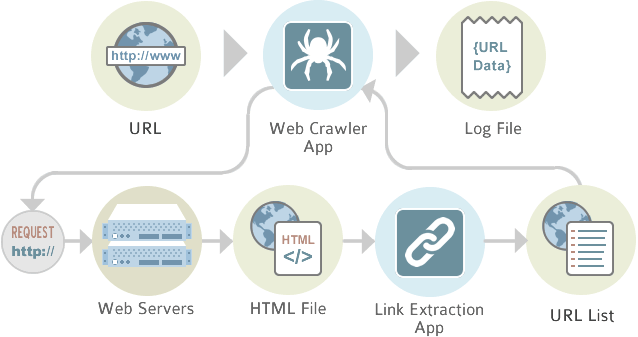
Innoscape Recruitment Challenge   
Data Collection

# Part 1: general questions

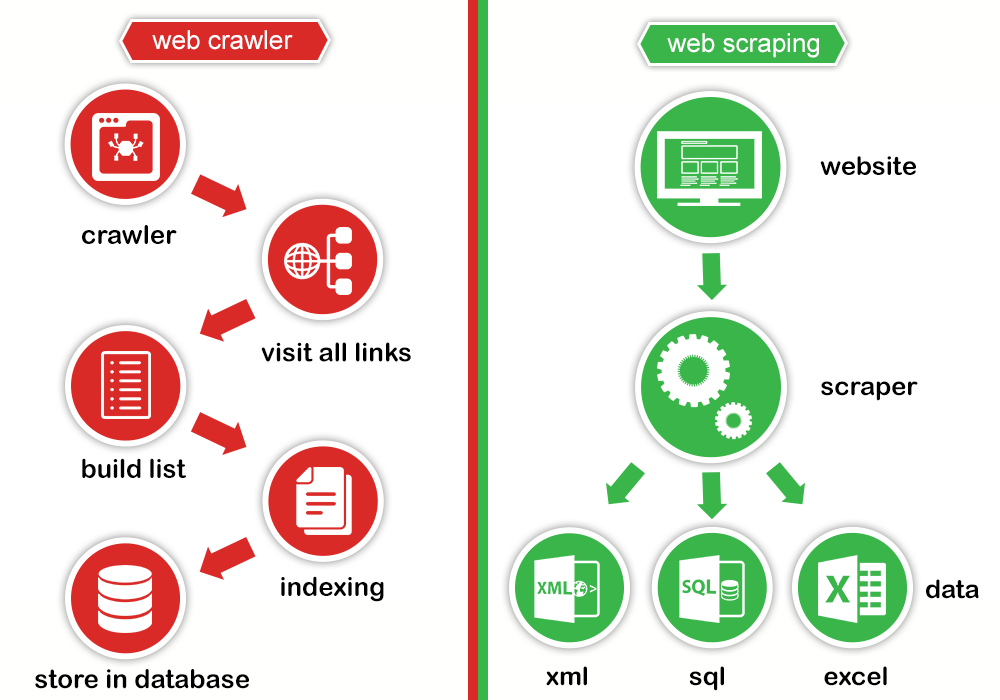
**a) Can you explain what is the general concept of web data crawling?**

Avant de s’intéresser au concept de *web data crawaling*, il faut tout d’abord comprendre l’idée général du *web crawling* représenté par l’illustration ci-dessous. Un web *craweler*, *spider*, ou robot d’indexation en français, est un programme parcourant les sites web à la recherche d’information général dans le but de les classer. Le processus est simple : parcourir un très grand nombre de sites web avec le rapidement possible afin de tirer les résultats les plus pertinents répertoriés dans un *Log File*.

[](https://seopressor.com/blog/how-to-control-web-crawlers/)

Lorsque que l’on traite la question de *web data crawling*, il s’agit de s’intéresser aux quatre principes du robot définissant son comportement en matière de recherche, collecte et traitement de la donnée. Ces principes sont résumés dans l’illustration ci-dessous.

Enfin, le *web crawling* doit être différencié du *web scraping*. L’objectif du *crawling* est de concevoir une liste de sites web, à la différence du *scraping* qui est utile pour récupérer des informations spécifiques sur une ou plusieurs pages web.

[](http://prowebscraping.com/web-scraping-vs-web-crawling/)

**b) What is the base concept of “xpath”?**

XPath signifie XML Path Language. Il peut être décris comme est un langage de requête permettant de sélectionner des nœuds et sous-nœuds d'un document XML. Par ailleurs, XPath peut être utilisé pour effectuer des calculs mathématiques avec des valeurs contenus dans un document XML.

La syntaxe utilisé par XPath est « path like », en d’autres termes, elle fonctionne comme une arborescence.

Enfin, XPath contient plus de 200 fonctions intégrées pour diverses thématiques.

**c) Can you present shortly the approach that you will be using in order to collect any data from a web site using its HTML / XML Dom tree?**

Une première approche serait de comprendre comment la page web fonctionne en lisant rapidement le code source. Puis, j’utiliserai un script python pour récupérer les balises qui m’intéresse en m’aidant de la libraire BeautifulSoup.

Une deuxième approche, plus rapide si la page est « bien » construite, serait d’utiliser des logiciels prêt à l’emploi. Je conseille donc [import.io](https://www.import.io/) qui prend la forme d’un navigateur internet et permet de sélectionner les champs que l’on souhaite étudier ; ou bien d’utiliser une extension de navigateur qui fonctionne de la même manière, mais est plus pratique pour extraire des informations sous forme de tableau. J’utiliserai ici [Instant Data Scraper](https://chrome.google.com/webstore/detail/instant-data-scraper/ofaokhiedipichpaobibbnahnkdoiiah), une extension disponible sous Google Chrome.

# Part 2 : Practical case

In this simple study case, we would like you to build a scraper using the free version of the scraping platform ‘Octoparse’ (https://www.octoparse.com/pricing). Your target will be to extract all user reviews from a sub-category of the Amazon.fr web site.

Voir pièce-jointe.