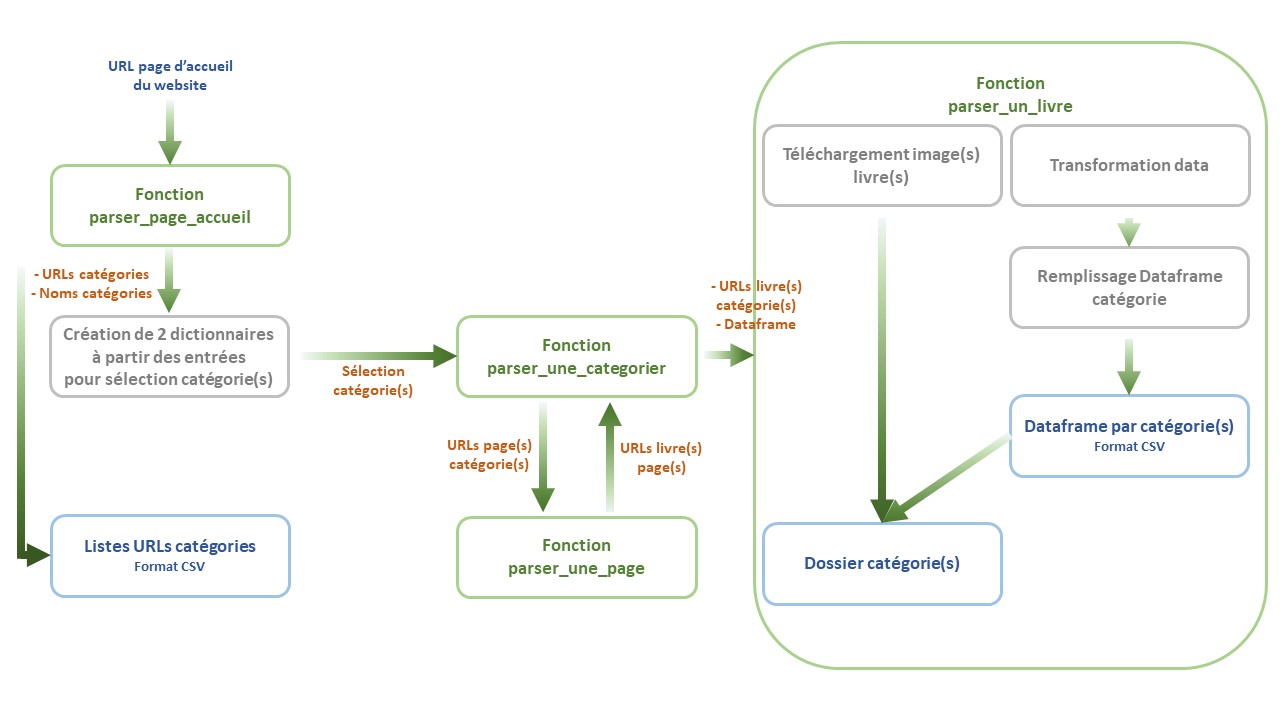
**Objectif de développement :**

Ce programme a été développé dans le but de concevoir un pipeline ETL à partir du site web <http://books.toscrape.com/index.html>. L’objectif étant de pouvoir extraire et analyser les données pour chacun des livres commercialisés par le revendeur.

**Schéma de principe :**



**État du développement :**

Le développement est actuellement terminé. Cependant des évolutions suivantes seraient les bienvenues :

* La mise en place d’une sélection multiple de catégories. Pour le moment le programme permet de "scraper" soit l’ensemble du site, soit une seule catégorie.
* Une simplification de cette même partie avec l’usage d’un unique Dataframe apporterait de meilleures performances. En effet, aujourd’hui la sélection se fait à partir de 2 dictionnaires dont les 2 possèdent un « index/clé » en commun.

**Environnement de développement :**

Le programme a été développé à partir de Python 3.10, et utilise les paquets suivants :

* beautifulsoup4==4.11.2
* bytesbufio==1.0.3
* lxml==4.9.2
* pandas==1.5.3
* Pillow==9.5.0
* requests==2.28.2

**Exécution du programme :**

Le programme peut être lancé via la ligne de commande *python .\web\_scraping\_BooksToScrape.py*,

Puis il faudra renseigner l’adresse URL de la page d’accueil du site, et suivre les indications fournies dans le déroulement du programme.

**Description des fonctions :**

Le programme se compose de 4 fonctions principales :

* ***parser\_page\_accueil*** : cette fonction permet de répertorier toutes les catégories du site. Afin de pouvoir les traiter dans les fonctions suivantes. Cette fonction permet également de préparer le fichier URL\_categories.csv listant l’ensemble des liens URL de chaque catégorie du site.
* ***parser\_une categorie*** : cette fonction appelle la fonction ***parser\_une\_page*** autant de fois qu’il y a de pages dans la catégorie. Une fois la liste de livre à analyser établi, la fonction appelle la fonction ***parser\_un\_livre***
* ***parser\_une\_page*** : cette fonction récupère les liens URL de tous les livres situés sur une page HTML. Puis elle retourne ce résultat à la fonction ***parser\_une\_categorie***.
* ***parser\_un\_livre*** : Cette fonction permet d’extraire les informations relatives à un livre, puis après transformation elle les met en forme de tableau pour les exports au format csv. Cette fonction permet également de télécharger les images de chaque livre.

**Fichiers de sortie :**

L’ensemble des éléments générés sont stockés dans un dossier « output », puis classés en catégorie, où chaque catégorie dispose des images extraites et du fichier csv relatif à la catégorie.