pyrefsearch

Fonctions:

1. Recherche de publications (base de données Scopus) et de brevets US et canadiens (bases de données INPADOC et/ou USPTO, au choix) pour une liste d'auteur.e.s sur une gamme d'années donnée.
2. Recherche de profils Scopus pour une liste d'auteur.e.s

NB: La recherche de brevets dans INPADOC (*International Patent Documentation*) est préférable à l'USPTO parce que les brevets sont groupés par familles, ce qui permet notamment de retenir uniquement les premiers dépots et octrois de brevets pour une invention et éviter les redondances (brevets dans plusieurs pays, multiples numéros de brevets pour une même invention reliés à des variations mineures comme des changements aux revendications, des modifications cléricales, etc.).

Pré-requis:

- Une clé API est requise pour effectuer des recherches sur Scopus, laquelle peut être obtenue à partir de cette page web, click sur *I want an API Key* puis *Create API Key*. La première exécution du script demande de saisir la clé, celle-ci est ensuite sauvegardée sur le poste de travail. Le script doit être exécuté à partir d'un poste sur le réseau universitaire pour que l'accès à Scopus soit autorisé.
- Une clé API est requise pour faire des recherches de brevets dans la base de données INPADOC (*International Patent Documentation*) gérée par l'EPO (*European patent office*) et accessible via *espacenet*. Il faut d'abord ouvrir un compte en ligne pour accéder aux *Open Patent Services (OPS)* de l'EPO. Une fois le compte activé, il faut définir les variables environnementales *PATENT_CLIENT_EPO_API_KEY* et *PATENT_CLIENT_EPO_SECRET* sur le poste de travail tel qu'expliqué dans *How to get an EPO OPS API key*.

Utilisation:

- Appel du script: *pyrefsearch\pyrefsearch.py data\pyrefsearch.toml*
- Les paramètres d'exécution sont lus dans le fichier *pyrefsearch.toml*
- Les données d'auteur.e.s sont lues dans un fichier Excel d'entrée (voir *pyrefsearch.toml*)
- Les résultats de la recherche sont écrits dans un fichier Excel de sortie (voir *pyrefsearch.toml*)
- Tous les fichiers spécifiés sont lus/écrits dans le même répertoire que le fichier .toml

Fichier des paramètres d'exécution *pyrefsearch.toml* (voir le fichier donné en exemple pour plus d'infos sur les paramètres) :

- *search_type* ("Publications" ou "Profils"): Type de recherche (publications & brevets ou profils d'auteur.e.s)
- *in_excel_file* : Fichier Excel d'entrée
- *in_excel_file_author_sheet* : Feuillet dans le fichier Excel d'entrée contenant les informations des auteur.e.s
- *first_year*, *last_year* : Première/dernière année de la gamme d'années (recherche de publications & brevets)
- *publication_types* : Descriptions des types de publications et codes Scopus correspondants pour la recherche de publications dans Scopus
- *local_affiliations* : Liste des institutions considérées comme des affiliations "locales"
- *scopus_database_refresh*: Intervalle de mise à jour de la copie locale de la base de données Scopus

Fichier Excel d'entrée *in_excel_file* :

Noms et identifiants Scopus des auteur.e.s spécifiés dans les colonnes suivantes du feuillet *in_excel_file_sheet_name*:

- *Nom* : Nom de famille de l'auteur.e
- *Prénom* : Prénom de l'auteur.e
- *ID Scopus* : Identifiant Scopus de l'auteur.e pour une recherche de publications (laisser la cellule vide si aucun identifiant Scopus n'est disponible)

Il a 2 options pour le format du fichier:

1. Base de données du 3IT, où le statut des auteur.e.s est spécifié par année fiscale (*Régulier* ou *Collaborateur*): seuls les membres ayant eu un statut *Régulier* pendant au moins une année dans la gamme d'années spécifiée seront inclus dans les recherches
2. Une simple liste d'auteur.e.s

Fichier Excel de sortie :

- Recherche de publications & brevets:
  - Nom du fichier: *<in_excel_file mantissa>_publications_<first_year>-<last_year>.xlsx*
  - Feuillet *Résultats* : résumé des résultats de la recherche
  - Feuillets *Articles*, *Conférences*, *Lettres*, *Livres*, *Chap. de livres*, *Rapports*, *Brevets US (en instance)*, *Brevets US (délivrés)* : résultats de la recherche par type de document (omis si aucun résultat)
  - Feuillet *Profils par identifiants*, colonne *Erreurs*: erreurs le cas échéant dans les profils d'auteur.e.s et/ou disparités entre les informations dans le fichier Excel d'entrée et la base de données Scopus
- Recherche de profils :
  - Nom du fichier: *<in_excel_file mantissa>_profils.xlsx*
- Recherche de publications & brevets ou de profils :
  - Feuillet *Homonymes* ou *Profils* selon le type de recherche, colonne *Affl/ID*: repérage des noms d'auteur.e.s correspondant à plusieurs identifiants Scopus, afin de s'assurer que les bons identifiants sont fournis dans le fichier Excel d'entrée pour la recherche de publications & brevets:
    - *Affl.* : l'affiliation est comprise dans la liste *local_affiliations* spécifiée dans *pyrefsearch.toml*
    - *ID* : l'identifiant Scopus est identique à celui dans le fichier Excel d'entrée

Problèmes / points à considérer :

- Il y a des erreurs fréquentes dans la base de données Scopus (ex : publications attribuées à tort à des auteur.e.s ayant le même nom, affiliation incorrecte, etc.). Les informations dans le fichier Excel de sortie (feuillet *Homonymes* ou *Profils* selon le type de recherche, colonne *Affl/ID*) aident à repérer ces erreurs. Le cas échéant, il faut demander à la personne concernée de faire une mise à jour de son profil Scopus. La mise à jour de la base de données peut prendre jusqu'à une semaine, il faut ensuite mettre à jour la copie locale de la base de données via le paramètre *scopus_database_refresh* dans le fichier *pyScopus.toml* avant de lancer une nouvelle recherche avec le script.
- Google Scholar semble plus complet que Scopus, en particulier pour les conférences, mais il ne semble pas y avoir de moyen facile de faire des recherches scriptées dans Google Scholar.

Installation:

- Projet *pyrefsearch* disponible sur github
- *pip install -r requirements.txt*