université PARIS-SACLAY

Comparer une liste de publications Scopus ou OpenAlex avec une collection HAL

Journées CasuHAL 2024

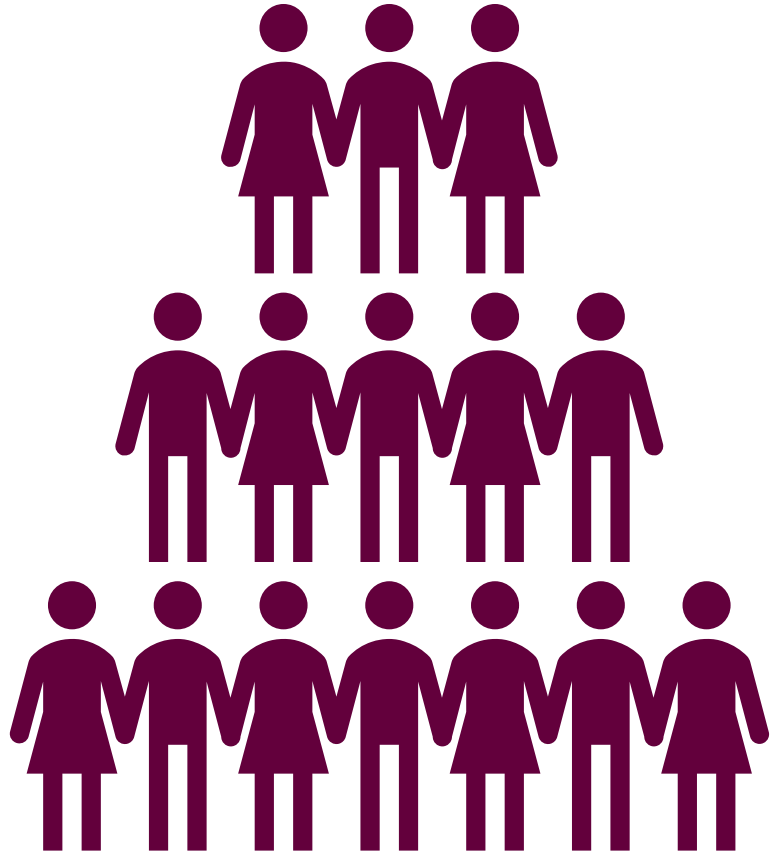
université PARIS-SACLAY

Contexte : année d'évaluation HCERES

Université Paris-Saclay = beaucoup d'unités très diverses, pas une base où on retrouve toutes les publications, et aucun moyen d'accorder suffisamment de temps à chaque unité

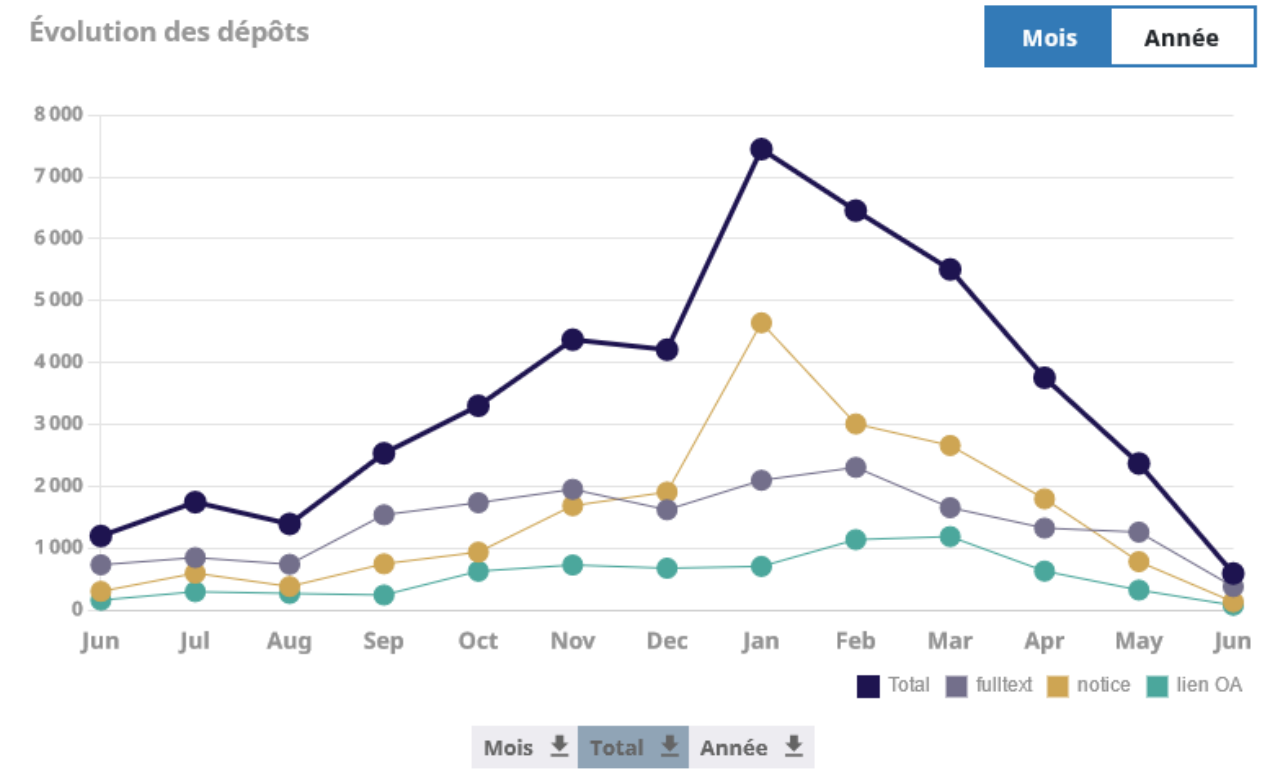
L'équipe

- 17 personnes + collègues en réseau
- Temps par personne : entre 0,1 et 0,5 ETP
- Chacun établit le lien avec plusieurs unités et apporte le soutien nécessaire



Notre mission

- On ne dépose pas à la place des auteurs
- On veut aider les auteurs au maximum sans faire à leur place
- On passe uniquement par les DU
- Il ne s'agit pas de surveiller et punir



La méthode choisie

- Exporter les publications de chaque unité depuis une BDD
- Comparer la liste exportée avec ce qui est présent dans HAL
- Tenir le DU informé de :
  - Ce qui est correctement dans la collection
  - Ce qui est dans HAL mais mal affilié/non tamponné
  - Ce qui n’est pas dans HAL et doit être déposé
- Le DU s’organise avec son équipe pour progresser vers 100% de « correctement dans la collection »

Comment ?

- Notebook utilisant les API HAL
- Fonctionne à partir d'exports Scopus ou OpenAlex
- Mis à disposition sur Colab dans un premier temps puis une instance Jupyterhub de l'université

Vérifier si des DOIs sont dans la collection HAL d'un laboratoire

Charger les "librairies" (boîtes à outils de code) qui seront nécessaires pour l'opération

```
[ ]: ! pip install Unidecode
     ! pip install langdetect
     ! pip install regex
```

```
[ ]: ! pip install openpyxl
```

```
[ ]: import requests
     import pandas as pd
     import numpy as np
     import regex as re
     from unidecode import unidecode
     from langdetect import detect
```

Indiquer dans la cellule suivante, à la place de MA_COLLECTION, le code collection du laboratoire. Les guillemets autour du code collection doivent rester.

```
[ ]: collection_a_chercher="CESP"
```

Dans la barre de gauche, cliquer sur l'onglet "Fichiers" et y importer (par cliquer-glisser ou en utilisant le bouton d'import) le fichier Excel qui contient les publications du laboratoire. Ensuite, en cliquant droit sur le fichier importé, copier le chemin d'accès à ce fichier à la place de mon/fichier dans la cellule suivante (garder les guillemets autour)

```
[ ]: fichier="cesp_scopus_20240130.xlsx"
```

```
[ ]: endpoint="https://api.archives-ouvertes.fr/search/"
```

Démonstration ?

- Notebook en local
- Notebook sur Colab
- Notebook sur JupyterHub

Fonctionnement du notebook pas-à-pas (1)

Etapes préparatoires : chargement des librairies et des données

1. Si nécessaire, installe les librairies à utiliser
2. Importe les librairies à utiliser
3. Choisit la collection HAL qui sera utilisée comme référence
4. Choisit le chemin du fichier (Scopus ou OpenAlex) à comparer
5. Charge le fichier à comparer

Fonctionnement du notebook pas-à-pas (2)

Fonction principale : comparaison des publications par DOI

- Prend en compte chaque ligne du fichier externe (idéalement le DOI et le titre)
- Si un DOI existe dans le fichier externe, il est recherché dans la collection (rapide car collection chargée en mémoire)
- S'il n'est pas trouvé dans la collection, il est recherché sur l'API HAL générale (parfois lent, dépend de l'API)
- S'il n'est pas trouvé dans HAL ou s'il n'existe pas dans le fichier, le titre est traité

Fonctionnement du notebook pas-à-pas (3)

Fonction principale : comparaison des publications par titres

1. Le titre normalisé est cherché dans la collection
2. S'il n'est pas trouvé, un titre à la graphie proche (95% similarité) est cherché dans la collection
3. Si rien n'est trouvé dans la collection, le titre est cherché dans tout HAL via l'API
4. S'il n'est pas trouvé, un titre à graphie proche (opérateur ~) est cherché via l'API
5. Selon le résultat, une mention est inscrite dans le tableau

Résultats du processus

- Le tableau complété est exporté au format .xlsx
- En filtrant par statut, la personne référente peut vérifier que les résultats sont cohérents par rapport aux attentes
- En fonction du nombre de statuts « à vérifier » et de sa charge de travail, elle peut choisir de vérifier manuellement ce que le système a indiqué comme incertain ou de laisser ces publications de côté
- Une fois les vérifications manuelles faites, elle peut produire des statistiques et envoyer le fichier ainsi que les statistiques au DU concerné.

Retours des personnes référentes

Initialement : grosses craintes sur la charge de travail et la prise en main induites

- Après formation et accompagnement : la plupart on pu s'approprier le notebook même si un accompagnement ponctuel reste nécessaire
- Sur la charge de travail, la vérification automatique y compris des titres est un allègement significatif et bienvenu

Un bilan positif malgré quelques limites difficiles à dépasser

- Les unités mal représentées dans Scopus (spécialement SHS) restent difficiles à traiter malgré l'apparition d'OpenAlex car OpenAlex a encore des progrès à faire en matière de lien publications-institutions et nos unités n'ont pas toutes un ROR
- Les titres très brefs et non significatifs (par ex. « Introduction ») ne peuvent pas être traités automatiquement s'ils ne disposent pas d'un DOI

Améliorations à apporter au code dans les années à venir

- Pour éviter les erreurs dans les noms et les chemins de fichiers, passer par des « input » et boîtes de dialogue
- Pour éviter les modifications accidentelles de code, limiter le notebook à l'appel de fonctions extérieures
- Pour améliorer la réutilisabilité, passer les fonctions en programmation orientée objet
- Dans l'idéal, déployer une version en ligne avec serveur

Améliorations à apporter au process à l'avenir

- Elargir le modèle de fichiers lisibles par le processus voire permettre l'interrogation automatique des BDD sans export manuel
- Proposer des visuels de statistiques dans le temps pour encourager les unités
- Combiner le processus existant avec d'autres procédés automatisés (comme celui présenté par Maximilien) pour déposer à la place de ceux qui n'ont pas le temps
- Encourager le dépôt dans HAL de métadonnées complètes en intégrant autant que possible les métadonnées du fichier externe dans les dépôts automatiques