# Exploration documentaire : à la recherche de thèmes émergents dans un domaine scientifique

L'objectif est de dresser une cartographie de thèmes présents dans un ensemble de publications d'un domaine scientifique donné en exploitant un ensemble de données et de méta-données : titres des publications, résumés textuels, ensemble de mots clés, dates de publication, noms des laboratoires des auteurs etc.

1) La première étape consiste à extraire un ensemble de publications depuis la plateforme HAL à partir des APIs d'interrogation.
2) La seconde étape consiste à enregistrer les données et méta-données recueillies dans un Dataframe Pandas et à y effectuer des analyses globales : nombre de publications par année, nombre de publications par laboratoires, fréquences d'usage des mots-clés etc.
3) La troisième étape consiste à exploiter des approches neuronales à base de modèles de langue pour représenter graphiquement les thèmes des publications sous forme de graphes.

## Etape 1.

La documentation de l'API HAL est disponible ici : https://api.archives-ouvertes.fr/docs/search

L'obtention des publications se fera via des appels de requête contenant les mots-clés et les propriétés des articles recherchés.

Il faudra utiliser la bibliothèque Python « requests » (client HTTP) : https://pypi.org/project/requests/

Les données seront obtenues par l'appel à la méthode get de requests en passant la requête en argument. Les réponses sont au format JSON, données que l'on pourra ensuite transférer dans un data frame pour les analyser, filtrer etc.

Exemple proche :

```
keywords = "(quantum OR qubit OR qbit)"
query =
"q=title_t:"+keywords+"&q=abstract_t:"+keywords+&fq=producedDateY_i:
["2015 TO 2020 »]& fl=authFullName_s,title_s & facet=true &
facet.pivot=producedDateY_i,docType_s & wt=json"
try:
    requete = "https://api.archives-ouvertes.fr/search/?"+query
    reponses = requests.get(requete, timeout=(300,300))
except requests.ReadTimeout:
    print ("temps limite atteint")
j = json.loads(reponses.text)
liste = j['facet_counts']['facet_pivot']['producedDateY_i,docType_s']
```

## Etape 2.

L’extraction des informations utiles de la réponse en JSON doit donner lieu à un data frame de la forme :

11 rows | 1453 rows × 7 columns | Static output

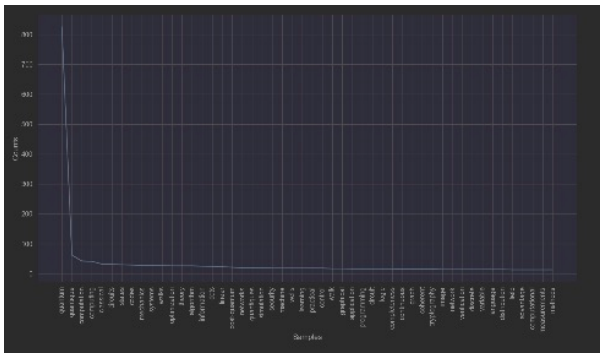
| | title_s | journalTitle_s | publicationDateY_i | publicationDateM_i | publicationDateD_i | abstract_s | authFullName_s |
|---|---|---|---|---|---|---|---|
| 0 | Quantum computing : a short introduction.,Info… | NaN | 2019 | 12.0 | 30.0 | In this short introd… | Thomas Cluzel,Claude Mazel,David R.C. Hill |
| 1 | Feedback exponential stabilization of open qua… | NaN | 2019 | 10.0 | 30.0 | In this thesis\, we … | Weichao Liang |
| 2 | Measurement-based quantum computation beyond q… | NaN | 2022 | 2.0 | 22.0 | Measurement-based qu… | Robert Ivan Booth |
| 3 | Photonic Resources for the Implementation of Q… | NaN | 2022 | 12.0 | 2.0 | The security of mode… | Simon Neves |
| 4 | Industrial approach to quantum dots in fully-d… | NaN | 2022 | 12.0 | 14.0 | Electron spin qubits… | Ioanna Kriekouki |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 1448 | Control of light emission of quantum emitters … | Light: Science and Applicati… | 2023 | 7.0 | 13.0 | Light emission of eu… | Martin Montagnac,Yoann Brûlé,Aurélien Cuche,Je… |
| 1449 | On the Bi diffusion from (001) GaAsBi-GaAs qua… | 5th international workshop o… | 2014 | 7.0 | NaN | NaN | Alexandre Arnoult,Aurélien Kuck,Hajer Makhlouf… |
| 1450 | Nanotechnology practical teaching at school an… | 2016 IEEE Nanotechnology Mat… | 2016 | 10.0 | NaN | In the last two deca… | Marc Respaud,Hélène Tap,Jérémie Grisolia,Gérar… |
| 1451 | Integration of the Rhombohedral BiSb(0001) Top… | ACS Applied Materials & Inte… | 2021 | 8.0 | 4.0 | Bismuth-Antimony all… | Dima Sadek,Daya S Dhungana,Roland Coratger,Cor… |
| 1452 | All-sky search for long-duration gravitational… | Classical and Quantum Gravity | 2018 | 3.0 | 22.0 | NaN | B Abbott,R Abbott,T Abbott,M Abernathy,F Acern… |

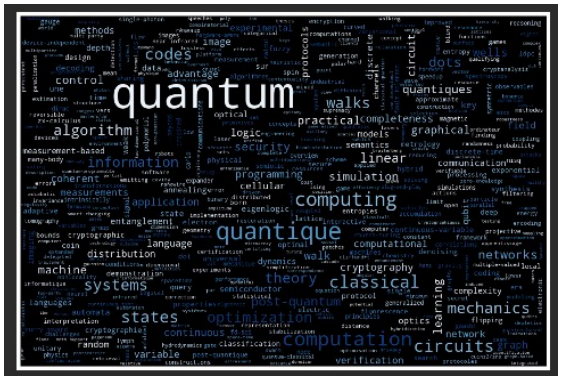
Il devra avoir au minimum pour colonnes :
['Title', 'Source', 'Publication Year', 'Publication Month', 'Publication Day', 'Abstract', 'Authors']

De manière à rendre plus pertinente l’analyse des résumés et des titres, il sera utile de supprimer les mots outils (stopwords).
On emploiera pour cela la bibliothèque Python NLTK : https://www.nltk.org/
Et l’on pourra s’aider de : https://medium.com/analytics-vidhya/removing-stop-words-with-nltk-library-in-python-f33f53556cc1

Des statistiques globales sur la distribution des mots pourront ensuite être réalisées en utilisant la méthode FreqDist de NLTK, puis un nuage de mots avec la bibliothèque WordCloud : https://pypi.org/project/wordcloud/

Ces deux étapes permettront la création de graphiques comme :






Cette analyse pourra être conduite sur les mots des titres et/ou des résumés mais aussi sur celles des noms des auteurs et de leurs laboratoires.

On se propose ensuite de représenter graphiquement l’évolution de l’usage des mots les plus fréquents par périodes temporelles. Cela permettra de savoir quels sont les mots clés et les thèmes qui apparaissent et disparaissent au fil des ans.

Enfin, on s’intéressera à mesurer les collaborations entre laboratoires et leur évolution. De quels pays ou villes sont-ils issus ? Quelles sont les collaborations les plus fréquentes ?

# Partie 3.