Supplemental file to QSS-2021-0048 paper (Chaignon & Egret, 2022): Identify scientific publications country-wide and measure their open access: the case of the French Open Science Barometer (BSO)

The goal of this notebook is helping potential readers replicate the study with their own data.

Collect data for Publications France 2015-2020

Objective: to collect data from French publications per year (from 2015 to 2020) for 6 databases + the BSO, analyze the overlaps (cross-matching by DOI), then search the Open Access (OA) data in the Unpaywall snapshot.

Collect and process Scopus data

Scopus export is limited to batches of 20,000 records. This is why we have favored SciVal, which provides a very similar corpus.

- **Scopus Query :** AFFILCOUNTRY(FRANCE) AND (PUBYEAR > 2014, PUBYEAR < 2021)

	All	2015	2016	2017	2018	2019	2020
Scopus	*749 134*	123 608	125 486	126 543	126 202	123 181	124 114
SciVal	*738 054*	121 068	122 587	124 248	124 210	122 321	123 620

- Data extracted from SciVal by years (in batches of 100,000) in csv format.
- Import the .csv files and keep the DOI + years in a spreadsheet

After selection of DOIs, processing of duplicates, and some residual cleaning:

Table of DOIs

Scopus	2015	2016	2017	2018	2019	2020	Total
	110 681	113 288	115 284	116 567	115 259	118 312	**689 391**

Collect and process Web of Science data

WoS export is very limited. As a shortcut we use InCites which provides a very similar corpus.

- **WOS query:** CU=FRANCE AND PY =(2015-2020) Timespan: 2015-2020. Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI, CCR-EXPANDED, IC

WoS	*All*	2015	2016	2017	2018	2019	2020
PY	*723 873*	117 047	119 891	120 062	120 207	124 790	121 876
FPY	*715 538*	117 047	119 891	120 062	120 207	124 284	114 047
InCites	*698 532*	113 647	117 001	117 330	116 573	119 638	114 343

- Data extracted from InCites by year (in batches of 50,000) in csv format.
- Import the .csv files and keep the DOI+year in a spreadsheet

After selection of DOIs, processing of duplicates, and some residual cleaning:

Table of DOIs

WoS	2015	2016	2017	2018	2019	2020	Total
	91 453	96 475	96 621	97 928	101 375	102 664	**586 516**

Collect HAL data

- Use of the API, with filter on year and on France (fr) : fq = producedDateY_i: 2015 and fq = structCountry_s: fr (queries per year, per batch of 10,000)

http://api.archives-ouvertes.fr/search/?q=*:*&rows=170000&wt=csv&fq=producedDateY_i:2015&fq=structCountry_s:fr&fl=halId_s,doiId_s,docType_s,submitType_s,title_s,producedDa

- Import the .csv files and keep the DOI+years in a spreadsheet

HAL API	*All*	2015	2016	2017	2018	2019	2020
all	*978 906*	163 578	171 637	171 425	172 298	165 962	134 006
fr	*932 150*	154 998	163 327	162 279	164 437	158 937	128 814
avec DOI	*58 493*	64 193	64 193	65 262	68 471	71 657	69 543

After selection of DOIs, processing of duplicates, and some residual cleaning (wrong DOI format):

HAL	2015	2016	2017	2018	2019	2020	Total
	54 332	60 114	60 905	64 082	66 673	64 804	**370 910**

Collect ADS data

- https://ui.adsabs.harvard.edu/
- **ADS query:** ((aff:"France") AND year:2015-2020)
- Export in Custom Format : % R |%> 25.5N | "% \ T" |% \ J |% Y |% W |% d in batches of 500
- Import the result .csv file and keep the DOI + year in a spreadsheet

The custom format suggested here exports the BibCode number, the list of authors (up to 25 authors), Title, Journal, Year, Document type, and DOI number.

ADS	*All*	2015	2016	2017	2018	2019	2020
	120 053	19 065	19 968	20 846	20 048	19 997	20 129

After selection of DOIs and processing of duplicates:

ADS	2015	2016	2017	2018	2019	2020	Total
	15 343	16 216	16 982	16 304	15 731	16 625	**97 201**

Collect PubMed data

- https://pubmed.ncbi.nlm.nih.gov/advanced/
- **Pubmed query (per year) :** (France[Affiliation]) AND (("2015"[Date - Publication] : "2020"[Date - Publication]))
- Import the result .csv file and keep the DOI + years in a spreadsheet. We may have to split by month for export, example for January 2015 : (France[Affiliation]) AND ("2015/1"[Date - Publication]) in batches of 5000.

PubMed	*All*	2015	2016	2017	2018	2019	2020
	329 693	48 365	52 743	54 222	54 412	56 038	63 913

After selection of DOIs and processing of duplicates: (The query 2015-2020 provides some data for the year 2021.)

PUBMED	2015	2016	2017	2018	2019	2020	2021	Total
	47 923	45 125	46 291	46 506	48 388	47 357	8 505	**290 095**

Collect MAG (Microsoft Academic Graph) data

Data provided by the Curtin Open Knowledge Initiative (COKI) team

- **MAG query:** SELECT mag.Year, doi FROM `academic-observatory.observatory.doi20211002 WHERE mag.Year > 2014 and mag.Year < 2021 AND((SELECT COUNT(1) FROM UNNEST(mag.authors) as auth WHERE REGEXP_EXTRACT(auth.OriginalAffiliation, r'Fran(ce|kreich|cia)(?:\W|\s+|$)') is not null) > 0)
- Import the result .csv file and keep the DOI + years in a spreadsheet

MAG	2015	2016	2017	2018	2019	2020	Total
	93 107	96 896	99 749	95 116	101 885	109 762	**596 515**

France 2015-20 corpus : FR15-20

The corpus is formed by cross-matching and removing duplicates on the basis of DOIs. It may be necessary to **systematically lowercase all DOIs** before comparing them.

Constitution of the general corpus (based on DOIs) : spreadsheet FR15-20

Year	FR-15-20	HAL	PUBMED	ADS	WoS	Scopus	MAG
	829 678	370 933	290 097	97 210	586 519	689 392	596 515
2015	*130 243*	54 337	47 925	15 345	91 455	110 711	93 107
2016	*133 505*	60 122	45 125	16 216	96 475	113 239	96 896
2017	*134 150*	60 909	46 291	16 982	96 622	115 306	99 749
2018	*134 916*	64 085	46 506	16 310	97 928	116 567	95 116
2019	*137 480*	66 677	48 388	15 731	101 375	115 259	101 885
2020	*151 646*	64 803	47 357	16 626	102 664	118 310	109 762
2021	*7 738*		8 505				
	829 678	370 933	290 097	97 210	586 519	689 392	596 515

Collect BSO data

- Download on the MESRI Open Data site : https://data.enseignementsup-recherche.gouv.fr/explore/dataset/open-access-monitor-france/information/?disjunctive.oa_host_type&disjunctive.year
- Download by year in csv format
- Some data cleaning : non ASCII characters in DOIs... Unix command to find non-ASCII characters : iconv -f UTF-8 file.txt -o /dev/null

BSO	2015	2016	2017	2018	2019	2020	Total
	141 672	149 858	147 707	160 880	155 580	0	**755 697**

Collate FR15-20 and BSO (file All-FR)

- Unix command: cat DOI-BSO.txt DOI-FR1520.txt | awk '{ print tolower ($0) }' | sort -u > DOI-All.txt

Cross with the Unpaywall snapshot

- Extract data from the snapshot into a csv file data.txt with separators | through a Python program: upw.py

In [ ]:
```
# Reading json results from Unpaywall
# Writing a .txt file with DOI, Year, Genre, Status

import jsonlines

i, j, n = 0, 0, 0
ligne = " "
nomF = input('Name of file to read : ')
annee = input('Year : ')
an = int(annee)
nomS = input('Name of file to write : ')
of = open(nomS, 'a')

with jsonlines.open('.../unpaywall_snapshot_2021-02-18T160139.jsonl') as f:
    for line in f.iter():
        n = n + 1
        j = line['year']
        if j  == an :
            i = i + 1
                ligne = str(line['doi']) + "|" + str(line['year']) + "|" + str(line['genre'])
                + "|" + str(line['is_oa']) + "|" + str(line['published_date']) + "|" + str(line['journal_name']
                + "|" + str(line['publisher']) + "\n"
                ligne = str(line['doi']) + "|" + str(line['year']) + "|" + str(line['genre'])
                + "|" + str(line['is_oa']) + "|" + str(line['oa_status']) + "\n"
                of.write(ligne)
                if i % 100000 == 0 :
                    print(n,i)
                        print(line['doi'], "|", line['year'], "|", line['genre'], "|", line['is_oa'], "|",
                              line['published_date'], "|", line['journal_name'], "|", line['publisher'])

print (n, "lines read", i, "lines found")
f.close()
of.close()
```

- Make sure the DOI file is lower case. Unix command: awk -F '|' 'BEGIN{OFS="|"} FNR==NR{a[1] =2 FS 3FS4;next}{ print 0, a[1]}' data.txt file-DOI.txt > file-result.txt

Statistical assessment of the general corpus (France 2015-20 + BSO), with Unpaywall as the baseline for the year

985,474 records

958,028 found in Unpaywall

General table of crossings by DOI

Year	All	FR-15-20	HAL	PUBMED	ADS	WoS	Scopus	MAG	BSO
	979 474	848 113	359 231	288 685	96 826	580 112	674 604	594 046	748 588
%	103%	87%	37%	29%	10%	59%	69%	61%	76%
2014	11 973	11 972	4 096	3 621	488	8 954	10 516	6 244	15
2015	157 053	133 817	51 734	41 287	15 387	91 028	108 195	92 722	140 493
2016	164 772	138 885	57 851	44 785	16 396	96 186	112 486	96 850	148 476
2017	162 179	138 845	59 451	46 057	16 806	95 731	113 077	95 808	146 179
2018	171 987	141 059	61 997	46 490	16 254	97 012	114 069	99 356	159 380
2019	167 412	139 514	63 413	48 047	15 410	96 712	111 422	102 338	153 705
2020	139 851	139 843	59 796	55 293	16 077	94 237	104 533	100 608	234
2021	4 247	4 178	893	3 105	8	252	306	120	106

Open Access									
Year	**is_OA**	**FR-15-20**	**HAL**	**PUBMED**	**ADS**	**WoS**	**Scopus**	**MAG**	**BSO**
	979 474	780 798	359 231	288 685	96 826	580 112	674 604	594 046	748 588
%	49%	52%	62%	58%	73%	55%	52%	49%	50%
2014	*41%*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*
2015	44%	47%	57%	52%	67%	49%	47%	45%	46%
2016	47%	50%	60%	57%	72%	53%	50%	47%	48%
2017	49%	52%	62%	59%	74%	55%	52%	49%	50%
2018	50%	53%	63%	61%	75%	56%	53%	51%	51%
2019	53%	55%	67%	62%	78%	59%	56%	53%	54%
2020	53%	55%	64%	63%	75%	56%	54%	51%	0
2021	*45%*	*35%*	*0*	*35%*	*0*	*0*	*0*	*0*	*0*

journal-article									
Year	**is_OA**	**FR-15-20**	**HAL**	**PUBMED**	**ADS**	**WoS**	**Scopus**	**MAG**	**BSO**
	797 449	717 896	318 286	286 201	89 421	541 874	579 106	522 035	606 930
%	53%	53%	64%	59%	76%	56%	55%	52%	53%
2014	*46%*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*
2015	48%	48%	58%	54%	72%	51%	50%	47%	49%
2016	51%	51%	61%	57%	75%	55%	53%	49%	52%
2017	53%	53%	64%	60%	76%	57%	56%	52%	54%
2018	55%	55%	66%	61%	78%	58%	57%	53%	55%
2019	56%	57%	69%	63%	80%	60%	59%	55%	57%
2020	55%	55%	66%	59%	78%	57%	58%	53%	*0*

DOIs not found in Unpaywall database :

Non-Upw	Year	All	FR-15-20	HAL	PUBMED	ADS	WoS	Scopus	MAG	BSO
		27 750	27 266	11 676	1 411	375	6 402	14 768	2 471	6 908
	%		98%	42%	5%	1%	23%	53%	9%	25%
	2014	0	0	0	0	0	0	0	0	0
	2015	0	4 017	1 503	262	79	1 111	2 538	458	1 070
	2016	0	3 933	1 676	160	18	1 028	2 070	102	1 331
	2017	0	3 862	1 776	180	48	719	2 141	178	1 486
	2018	0	3 859	1 778	163	55	831	2 143	189	1 436
	2019	0	4 372	2 159	286	109	801	2 204	102	1 585
	2020	0	6 473	2 784	286	66	1 912	3 672	1 442	0
	2021	0	72	0	74	0	0	0	0	0
	Non	27 750								0
		27 750	26 588	11 676	1 411	375	6 402	14 768	2 471	6 908

Search article language

The parameter is included in some of the sources (for example HAL : language_s).

For a general query, starting from the titles of the articles (available via Unpaywall) : use of the Python program langdetect.py .

In [ ]:
```
import pandas as pd
import csv
```

In [ ]:
```
pip install langdetect
from langdetect import detect_langs
```

In [ ]:
```
#open the file with the titles
with open("C:/.../titre.csv") as f:
    titlang = pd.read_csv("titre.csv", sep=',', encoding='utf-8')
```

In [ ]:
```
#select the title's column
titlang1=titlang[['title']]
titlang1.head()
```

In [ ]:
```
#lowercase the titles
titlang2 = titlang1['title'].str.lower()
titlang2.head()
```

In [ ]:
```
languages = []

# Loop over the rows of the dataset and append language code

def det(x):
    try:
        lang = detect_langs(x)
    except:
        lang = 'Other'
    return lang

for row in range(len(titlang2)):
    languages.append(det(titlang2.iloc[row]))

print(titlang2.head())
```

In [ ]:
```
#Convert "languages" in a dataframe
titrelangue=pd.DataFrame(languages)
titrelangue.head()
```

In [ ]:
```
#Add the first column to your title file
titlang['language'] = titrelangue[0]
titlang.head()
```

In [ ]:
```
#Write the result in a new csv file
titlang.to_csv("titre_langue.csv",index=False,sep = ';', encoding='utf-8')
```

The program provides a language and a probability of detection. We only keep detections assigned with a probability greater than 0.99.

Table for all years 2015-2020

Years 2015-2020	Total	en	fr	%en	%fr
All	979 474	724 436	147 114	74,0%	15,0%
Proba > 0,99	894 788	724 436	147 114	81,0%	16,4%
journal-article	737 165	590 713	132 290	80,1%	17,9%
book-chapter	65 912	48 482	12 208	73,6%	18,5%
proceedings-article	65 715	63 830	855	97,1%	1,3%
other	25 996	21 411	1 761	82,4%	6,8%

Table for year 2019

2019		en	fr	%en	%fr
Proba > 0,99	153 272	126 812	22 285	83%	15%
journal-article	123 874	101 138	20 335	82%	16%
book-chapter	11 280	8 699	1 593	77%	14%
proceedings-article	11 632	11 334	100	97%	1%
other	6 486	5 641	257	87%	4%

All	Total	en	%en	fr	%fr	All
OA et % OA	480 337	392 366	54,2%	37 231	25,3%	49,1%

Analysis by Research Area

The analysis is based on the determination of the disciplines by the BSO and is therefore only valid for the part in common with the BSO.

Below is the result of the crossing of languages and disciplines :

	Nombre	en	fr	%fr	Non OA	%OA	Non OA	%OA
Chemistry	38 181	34 202	308	0,9%	21 074	38,4%	196	36,4%
Computer and informa	56 286	48 170	702	1,4%	22 967	52,3%	427	39,2%
Mathematics	21 079	17 347	604	3,3%	4 798	72,3%	296	51,0%
Medical research	239 906	165 383	52 428	23,6%	74 837	54,7%	48 428	7,6%
Biology (fond.)	110 614	94 962	4 125	4,1%	34 454	63,7%	2 522	38,9%
Social sciences	45 125	16 106	24 521	57,8%	9 483	41,1%	15 073	38,5%
Physical sciences, Astro	87 362	77 293	738	0,9%	31 333	59,5%	244	66,9%
Earth, Ecology, Energy	65 242	55 176	2 690	4,5%	23 719	57,0%	1 469	45,4%
Engineering	26 579	22 981	924	3,8%	14 671	36,2%	564	39,0%
Humanities	57 766	14 964	31 612	61,0%	8 845	40,9%	18 131	42,6%