The Hitchhiker's Guide to OpenAlex

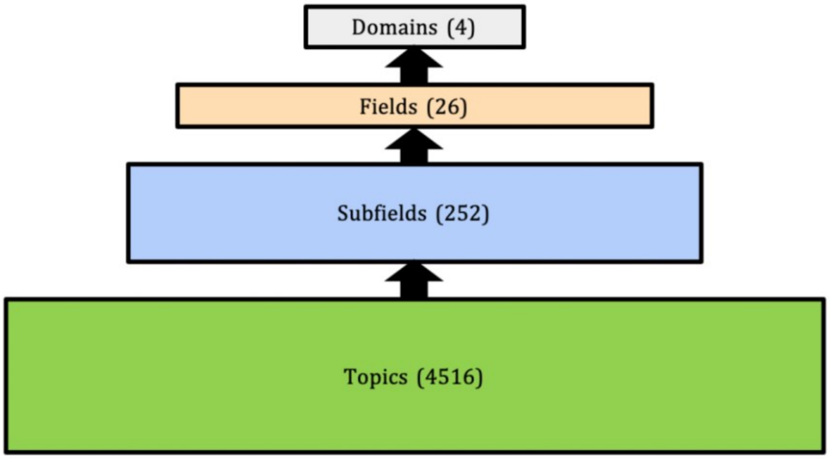
Introduction

This is a guide to using OpenAlex, an open-access bibliographic catalogue of scientific papers. OpenAlex was started in 2022 by the non-profit OurResearch. In this guide, one can find additional documentation and links to working with data about Works, Authors, Sources and Funding and Web of Science. This guide was credited and edited by Elisabetta Salvai (elsa@sodas.ku.dk),Jacob Dalsgaard, (jad@sodas.ku.dk), Marilena Hohmann (marilena.hohmann@sodas.ku.dk) and Sandro Sousa (ssou@itu.dk).

Works

Works classification

- **Classification logic: domain > field > subfield > topic.** For more information about the classification process, see the "OpenAlex Topic Classification"



  - **Domains**: If you query OpenAlex for the *domains,* you seemingly get 25 query results. However, there are duplicates, and only 4 unique domains are listed:
    - Life Sciences
    - Social Sciences
    - Physical Sciences
    - Health Sciences
  - **Fields:** There are the following 26 fields.
    - Agricultural and Biological Sciences, Arts and Humanities, Biochemistry, Genetics and Molecular Biology, Business, Management and Accounting,

    Chemical Engineering, Chemistry, Computer Science, Decision Sciences, Dentistry, Earth and Planetary Sciences, Economics, Econometrics and Finance, Energy, Engineering, Environmental Science, Health Professions, Immunology and Microbiology, Materials Science, Mathematics, Medicine, Neuroscience, Nursing, Pharmacology, Toxicology and Pharmaceutics, Physics and Astronomy, Psychology, Social Sciences, Veterinary

    - **Subfields:** Further classifications within the fields.
    - **Topics:** are assigned with an automated system. They report the three top-ranked topics for each work.
- **Concepts**: OpenAlex also lists *concepts* for each work. They are deprecated and shouldn't be used anymore (unless there is a very good reason to do so).

Coverage

- How far back in time do the records go? It is possible to find *works* until 1690.
- The OA documentation contains a comparison of the OpenAlex data set vs. similar databases: For an additional comparison of DOI coverage across different platforms see this paper

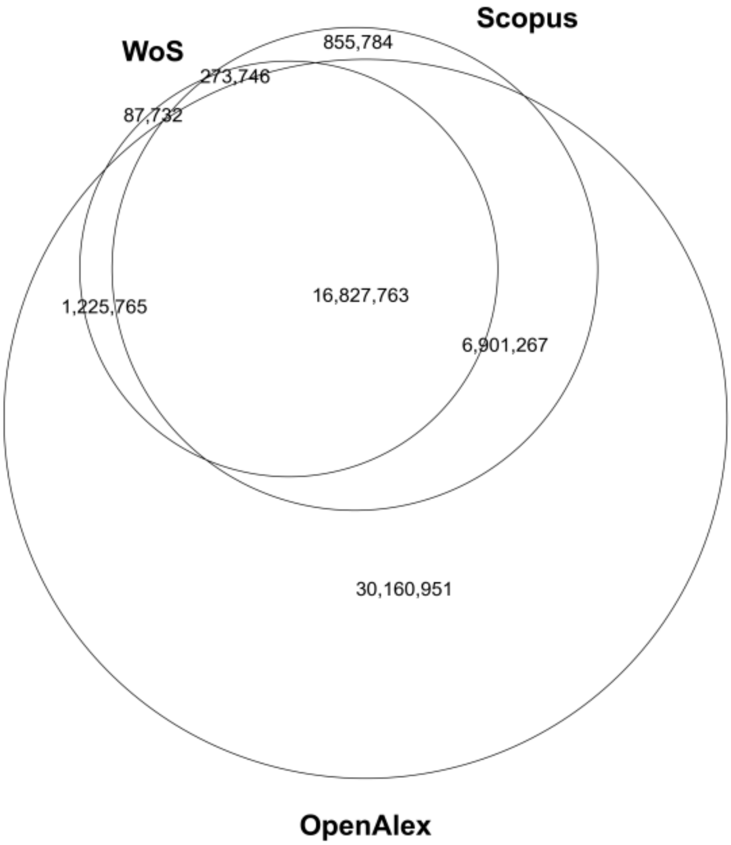
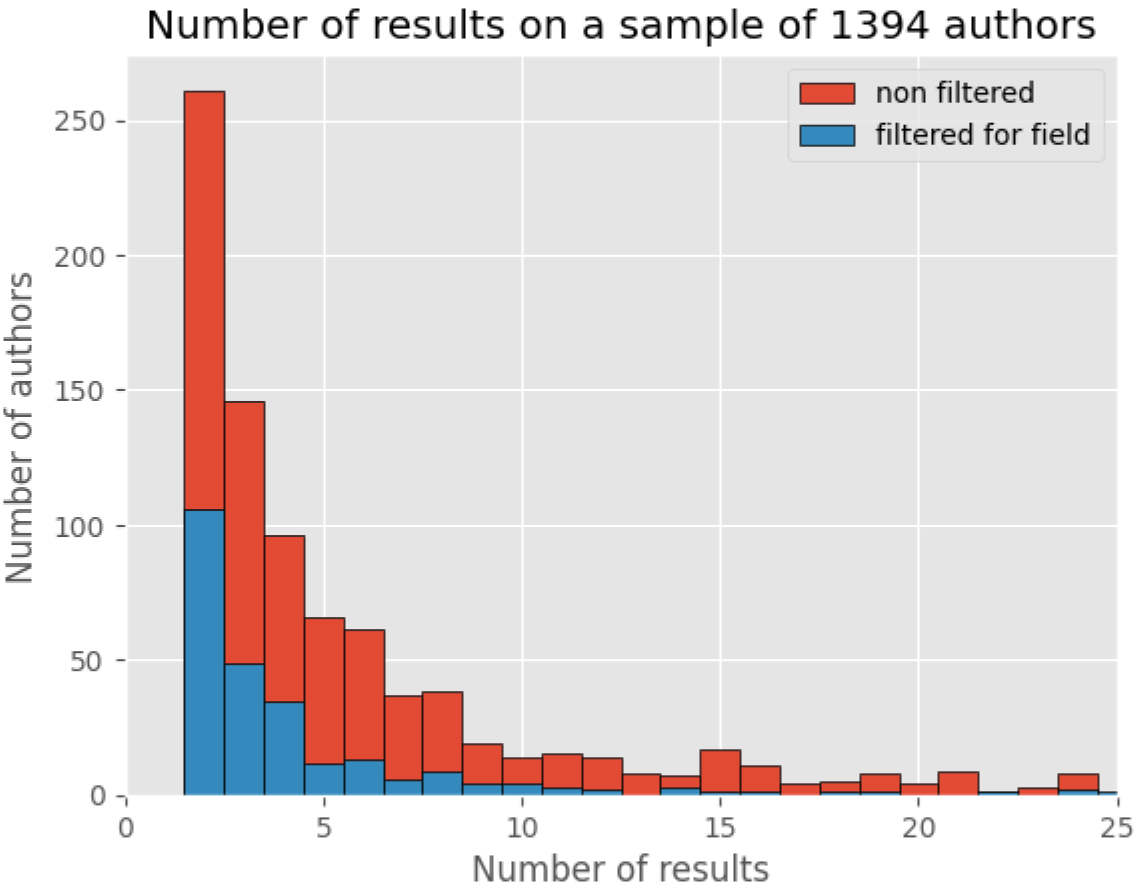


Fig. 1 Venn diagram of the intersection sizes of unique DOIs based in each database on exact DOI match, for records published between 2015 and 2022

Authors

When searching for an author using just their name, one can find multiple results. The plot below shows a comparison of the number of results obtained by searching the name of the author and adding a filter on the primary topic “computer science” (filter: at least one work published within that topic). The data includes only the first 25 results but note that there are authors with much more results.



- On a dataset of 67,381 authors, 65% are disambiguated.
- Not all authors have ORCID. When searching for an author by their name among the multiple results, not all of them have an ORCID. Note that ORCID was introduced in 2012.

Sources

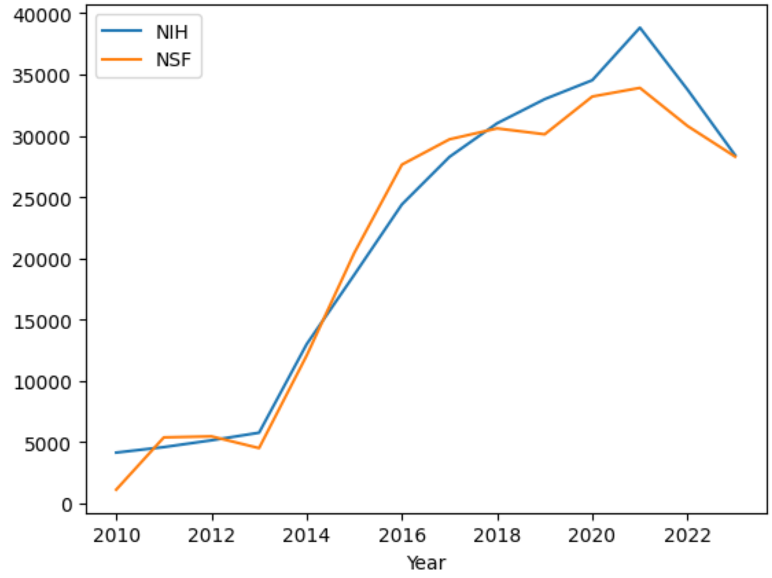
- If a conference has proceeding publications, they can be found as *source* on OA. Searching for a different version of the name (e.g. “Algorithms and Data Structure Symposium”, “Symposium on Algorithm and Data Structure” or “Algorithms and Data

Structure") can result in different results not always overlapping. It is also possible to search for acronyms.

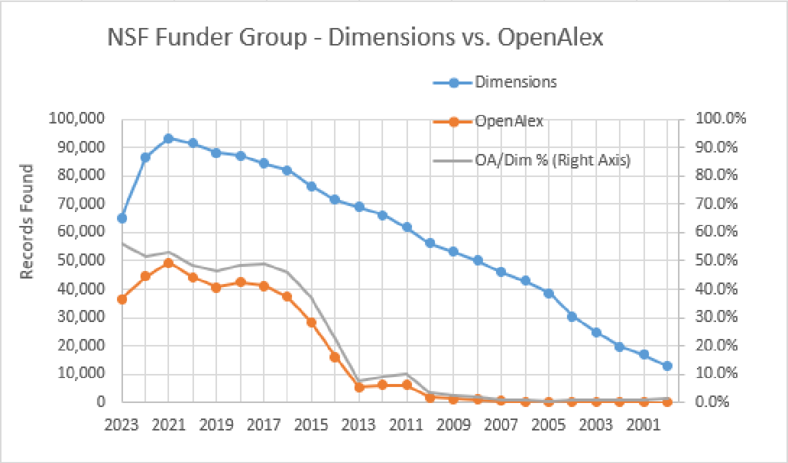
- **Different databases register/disambiguate sources in different ways.** OA can have multiple sources for the same yearly conference or just one.

Funding

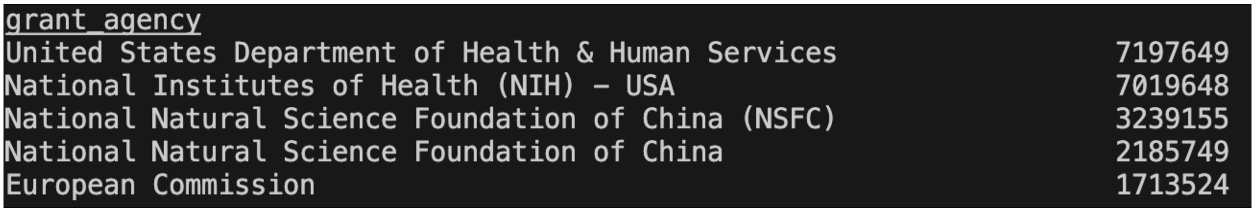
- Openalex Funding info is from Crossref
- Unique number of Funders = 32437
- Unique number of Funders with more than 100 associated works = 22025
- Number of Funded works = 8989818
- Number of works with grant id = 6872829
- Plot was made to compare funding info in WoS with OpenAlex. It shows the number of works associated with NIH and NSF in journals indexed by WoS.The plot with WoS data is currently still missing.



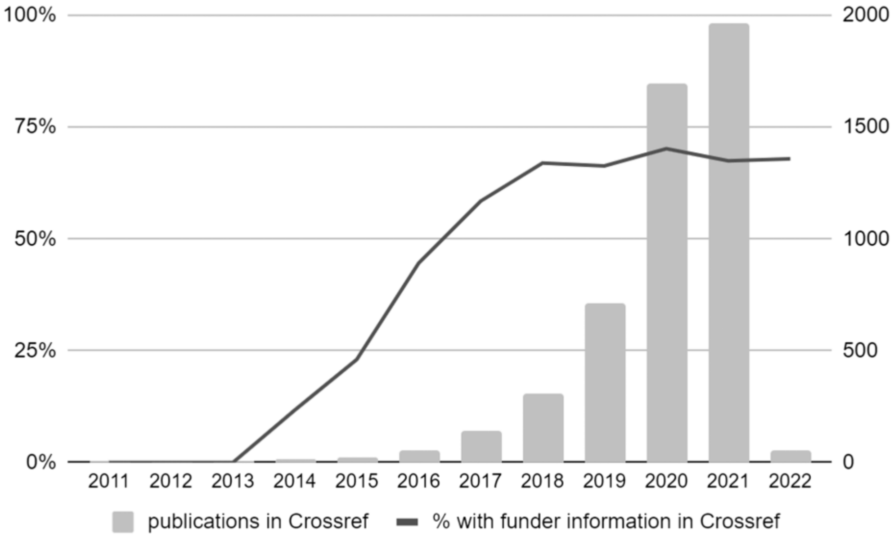
  - The interpretation of the plot above is complicated.
    Different databases register/disambiguate institutions in different ways.
    For something like NSF it requires a complicated API query in Openalex to match (chaining together 62 funding ids) with NSF funding info in *Dimensions*. This makes cross-database comparisons difficult. See this blog post
    From the blog post: "*OpenAlex appears to disambiguate and re-unify funding acknowledgments up to the top-level NSF group, while Dimensions seems to preserve whatever specific funder entity was acknowledged. This means that just searching for the top-level NSF returns 2.5x more records in OpenAlex than Dimensions.*"



- Following is a screen snippet counting total number of funded works in WoS



```
grant_agency
United States Department of Health & Human Services        7197649
National Institutes of Health (NIH) - USA                   7019648
National Natural Science Foundation of China (NSFC)         3239155
National Natural Science Foundation of China                2185749
European Commission                                         1713524
```

- This paper is based on a sample of research outputs of a Dutch Funding organization. It compares the coverage of those outputs in different bibliographic databases.
  - Only 67% of the sample contains funding information in Crossref
  - Other databases infer additional information from the acknowledgments
  - 

  - From the paper, it seems Dimensions or WoS is better for funding.

Corresponding author emails on Web of Science

Data collection:

- *OpenAlex data set*: Collected all OpenAlex publication records that meet the following criteria:
    - Published between January 1, 2022 and December 1, 2022.
    - Are of the OpenAlex type “article”
      According to the documentation, most publications are classified as “article,” but there are other types, such as “book”, “review,” etc. For a full list of types, see here.
    - Has references listed
    - Is written in English
- *Web of Science data set*:
    - Scraped data for the same period (Jan 1 - Dec 1, 2022) from Web of Science.
    - Looked up all the missing DOIs in a second scraping round.
    - Extracted all the records for which Web of Science lists one or more corresponding email addresses.

The table below shows:
(1) The field
(2) The number of publications collected through OpenAlex.
(3) The number of publications, including at least one corresponding author email, that could be retrieved from Web of Science.
(4) The percentage of publications that were found in the *OpenAlex data set* but missing from the *Web of Science data set*.

(1) Field	(2) # OpenAlex	(3) # Web of Science	(4) % Missing (WoS)
Political Science	50,724	24,918	50.88%
Economics	38,511	26,479	31.24%
Psychology	200,569	141,580	29.41%
Biology	292,754	256,265	12.46%
Computer Science	545,545	342,132	37.29%
Chemistry	300,042	261,845	12.73%
Sociology	55,040	32,924	40.18%
Physics	99,676	78,192	21.55%
Environmental Science	121,884	93,957	22.91%

Manual check of missing data
For each of the 9 fields, I got a random sample of 1000 missing DOIs and manually checked them on the Web of Science platform.

The table below shows:
(1) The field
(2) The number of WoS matches for the sample of 1000 OpenAlex DOIs.
(3) The number of WoS matches that contain a corresponding author email address.

Interpretation:

- Of the missing data, 89% of OpenAlex DOIs are not available on WoS.
- Of those missing matches we could find on WoS, approx. 9% do not list an email address.

(1) Field	(2) # Web of Science	(3) # Contains email address
Political Science	99	0
Economics	65	0
Psychology	60	1
Biology	95	0
Computer Science	92	0
Chemistry	87	0
Sociology	127	0
Physics	290	2
Environmental Science	63	0

Summary

- OpenAlex vs Web of Science: On average, 10.86% of the publications we collected through OpenAlex are missing on Web of Science.
- Corresponding author email addresses: On average, 26.18% of the publications we collected through OpenAlex are **either entirely missing** on Web of Science **or do not list an email address** on Web of Science.

Snapshot dataset

- Need quite a lot of space +1TB

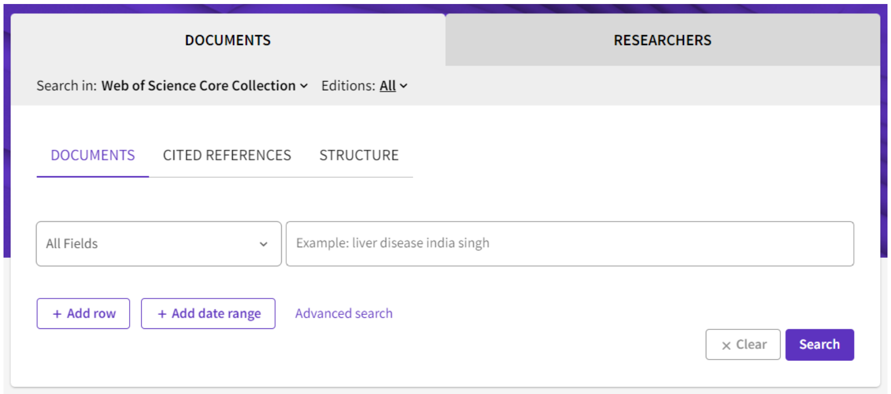
- Uses AWS buckets to store data. There is no cost associated with the download. If using your own machine or the SODAS server, you can use AWS CLI. On the ITU HPC (and other systems without admin privileges) you need to download via Python scripts.
- The full dataset (gzipped JSON line files) then needs to be converted to CSV. OA provides a script to do this. However**,** that **script does not parse all the data**. Funding info is excluded, but you can change the script to include any field that is available via the API. The script could also **benefit from changing the file type from CSV to parquet or feather.** Why?
  - More memory efficient. Don’t necessarily need to load all data.
  - With a library like Polars you might be able to load the required info in memory. So there is no need for distributed computing.

Small guide to scraping Web of Science

These are some high-level instructions explaining the general idea of collecting corresponding author emails from Web of Science.

There are two search options:

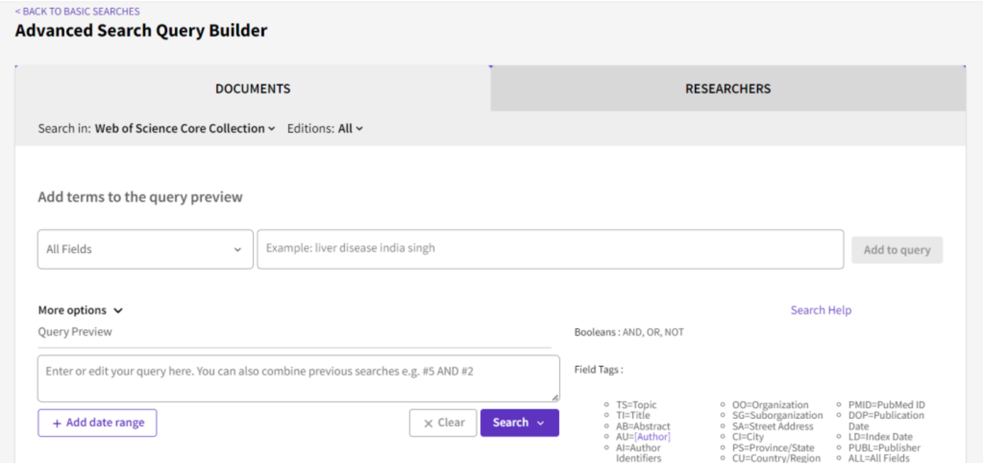
1. Define a search query through the Web of Science basic search. For instance, you can search for publications in a specific date range, field, etc. See via this link.



Search in: Web of Science Core Collection

Editions: All

2. Look up a list of DOIs in bulk through the Advanced Search option on Web of Science. You can enter slightly more elaborate queries to search the WoS database. Through the Advanced Search, you can build queries that look up lists of DOIs, which is much more efficient than looking them up one by one. See here via this link.



Advanced Search Query Builder

Search in: Web of Science Core Collection Editions: All

Add terms to the query preview

Example: liver disease india singh

Query Preview

Enter or edit your query here. You can also combine previous searches e.g. #5 AND #2

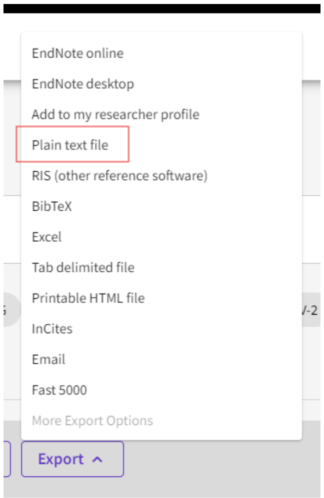
Booleans : AND, OR, NOT

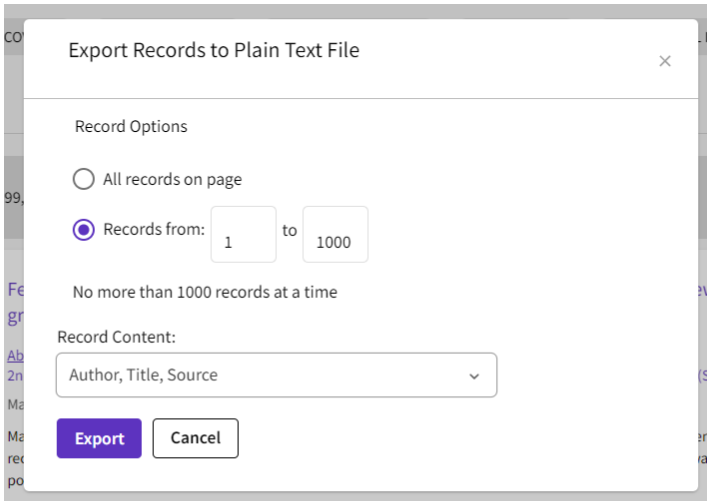
Field Tags :

- TS=Topic
- TI=Title
- AB=Abstract
- AU=[Author]
- AI=Author Identifiers
- OO=Organization
- SG=Suborganization
- SA=Street Address
- CI=City
- PS=Province/State
- CU=Country/Region
- PMID=PubMed ID
- DOP=Publication Date
- LD=Index Date
- PUBL=Publisher
- ALL=All Fields

For the scraping process:

- An automated scraper should navigate to the search interface, define the search options (Basic Search), or insert a bulk search query (Advanced Search).
- On the results page, there is the *Export* option.
- If you select *Plain text file*, you can download up to 1,000 results in a (semi-)structured format. It's fairly efficient to collect data this way, and parsing the data is also relatively easy since these text files all follow the same structure.





Export Records to Plain Text File

Record Options

○ All records on page

● Records from: 1 to 1000

No more than 1000 records at a time

Record Content: Author, Title, Source

Some notes:

- Based on said experience, we think it is more efficient to use a selenium scraper that navigates to the Web of Science website and performs the relevant searches. However, it should also be possible to actually *scrape* the HTML of the result records.
- For the selenium scraper, use the browser-based option (i.e., do not run this as a 'headless' scraper). Every now and then, seemingly random errors occur, and it's

(almost) impossible to fix them without seeing what WoS looks like in the browser window.