Research on ERIH PLUS approved SSH journals present in OpenCitations Meta database V.2

Ali Ghasempouri[1], maddalena.ghiotto[1], sebastiano.giacomini[1]

[1]unibo

Ali Ghasempouri

VERSION 2

MAY 09, 2023

OPEN ACCESS

DOI:
dx.doi.org/10.17504/protocols.io.5jyl8jo1rg2w/v2

Protocol Citation: Ali Ghasempouri, maddalena.ghiotto, sebastiano.giacomini 2023. Research on ERIH PLUS approved SSH journals present in OpenCitations Meta database . **protocols.io** https://dx.doi.org/10.17504/protocols.io.5jyl8jo1rg2w/v2Version created by maddalena.ghiotto

Protocol status: In development
We are still developing and optimizing this protocol

Created: May 03, 2023

Last Modified: May 09, 2023

DISCLAIMER

DISCLAIMER – FOR INFORMATIONAL PURPOSES ONLY; USE AT YOUR OWN RISK

The protocol content here is for informational purposes only and does not constitute legal, medical, clinical, or safety advice, or otherwise; content added to protocols.io is not peer reviewed and may not have undergone a formal approval of any kind. Information presented in this protocol should not substitute for independent professional judgment, advice, diagnosis, or treatment. Any action you take or refrain from taking using or relying upon the information presented here is strictly at your own risk. You agree that neither the Company nor any of the authors, contributors, administrators, or anyone else associated with protocols.io, can be held responsible for your use of the information contained in or linked to this protocol or any of our Sites/Apps and Services.

PROTOCOL integer ID:
81345

ABSTRACT

In this study, we present a comprehensive workflow to assess the coverage of publications in Social Science and Humanities (SSH) journals indexed in ERIH-PLUS and their Open Access status according to the Directory of Open Access Journals (DOAJ). The workflow utilizes three data sources: ERIH-PLUS, OpenCitations Meta, and DOAJ.

The application of this workflow results in a dataset containing detailed information on SSH publications, including their disciplines, countries of origin, and Open Access status. Each step of the methodology enriches the dataset with new variables and insights. The output of this workflow includes discipline and country rankings, as well as visualizations to effectively communicate the findings. By following this step-by-step approach, researchers can better understand the landscape of SSH publications, identify trends in disciplines and countries, and evaluate the prevalence of Open Access in the field.

The software implemented by means of this methodology and related documentation can be found in our github repository that will eventually be published on Zenodo.

Retrieve OpenCitation Meta publication and Journals that a...

1 Starting from the ERIH-PLUS index of Social Science and Humanities approved journals dataset ERIHPLUSapprovedJournals.csv (downloaded 27/04/2023) we want to retrieve all the publications belonging to one of those journals, included in OpenCitations Meta database (https://opencitations.net/meta#:~:text=For%20each%20publication%2C%20the%20metadata,and%20PubMed%20Identifiers%20(PMIDs).)

1.1 In order to fulfill this task, we download the OpenCitations Meta data dump.
To process the files we run the file *main.py* : this script processes CSV files in parallel.
This is how it works:

```
with concurrent.futures.ProcessPoolExecutor(max_workers=4) as
executor:
```

This line creates a ProcessPoolExecutor with 4 worker processes. The with statement is used to ensure that the executor is properly closed once the processing is done. The number 4 represents the maximum number of worker processes that will be created to execute the tasks concurrently. You can adjust this number based on the resources available on your system.

```
results = executor.map(process_file_wrapper, [(f, erih_plus_df)
for f in batch_files])
```

The executor.map() function is used to apply the process_file_wrapper function to a list of input arguments. In this case, the input arguments are tuples, each containing a file from the batch_files and the erih_plus_df DataFrame. The executor.map() function returns an iterator that yields the results of applying the process_file_wrapper function to each tuple in the list.

```
all_results.extend(results)
```

This line extends the all_results list with the results obtained from processing the current batch of files. The extend() method is used to add multiple items to the list at once.

1.2 Processing the ERIH-PLUS Journals dataset and matching venues identifiers with OpenCitations Meta dump's venues.

For this step, the function *process_meta_csv* () is called. It takes in input as parameters
1. the chunk of the csv that is being processed
2. and the ERIH-PLUS dataset as a dataframe.

Note

Input: ERIH-PLUS approved journal's dataset
Structured as follow:

	Journal ID	Print ISSN	Online ISSN	Original Title	International Title	Country of Publication	E
	486254	1989-3477	NaN	@tic.revista d'innovació educativa	@tic.revista d'innovació educativa	Spain	

OpenCitations Meta data about **venues** (issn that we need to decide how to retrieve)

Note

output: A dataset mapping OpenCitations Meta venue data (OMID and ISSN) to ERIH-PLUS venue data (Journal ID and ISSN).
This dataset will have the following structure:

	OC_omid	issn	EP_id	Open_Access	E
	meta:br/060167	[4522-4592, 5687-3452]	503890	True	
	meta:br/060167	isbn:242352513	NaN	NaN	
	meta:br/060167	issn:4522-4592	503890	4522-4592	

Note that our research question is about the **coverage of publication** so we will eventually need to query the number of publication to OpenCitations Meta database/ retreive the number of publications each journal has from ERIH-PLUS

1.3 Processing DOAJ data dump and adding Open Access information to **final df** (output dataframe of step 1.1.)
For this step, the function *process_doaj_file()* is called.
This function takes in input
1. **final df**
2. the path for the DOAJ data dump
and returns a Dataframe **merged_data** structured as follows:

	OC_omid	issn	EP_id	Publications_in_venue	Open_Access
	meta:br/060167	[4522-4592, 5687-3452]	503890	56	True
	meta:br/060167	[2423-5251, 8763-2891]	876390	12	Unknown

This is how the function works:
Fetch Open Access information from DOAJ

Create a dictionary of Open Access ISSNs

- Using the fetched data from DOAJ, create a dictionary with ISSNs as keys and Open Access status (True/Unknown) as values.

Merge Open Access information with the main dataframe

- Using the dictionary created before, create a new column in the main dataframe indicating

the Open Access status for each journal. we can use the map() function in pandas to achieve this.

- Example code:

```
merged_data['Open Access']
=merged_data['OC_ISSN'].map(open_access_dict)
```

Fill missing Open Access information with 'Unknown'

- Some ISSNs may not have a corresponding entry in the Open Access dictionary. In this case, we cannot assume whether these journals are Open Access. Fill the missing values in the 'Open Access' column with 'Unknown'.

Retrieve data about countries and disciplines

2 Our second and third research question are
1. What are the disciplines that have more publications?
2. What are countries providing the largest number of publications and journals?

These information are both present in the ERIH-PLUS dataset.

2.1 **Disciplines:**
As we can see every journal can have multiple disciplines:

In order to count the disciplines and understand which one has the highest number of publications we will need to disassemble them and map them individually with the venue.

Note

Output:

	ERIH-PLUS Disciplines	OMID of the journal
	Cultural Studies	...
	Art and Art History	identifier
	Art and Art History	identifier
	Cultural Studies	identifier
	Human Geography and Urban Studies	
	Human Geography and Urban Studies	

2.2 **Countries:**

For the countries we will do the same.

We take into consideration that some journals have no specified country. To tackle this issue we can try to see if it is present in the DOAJ dataset.

2.3 Checking wether missing countries in erih-plus are present in doaj dataset, where **Country of Publisher** contains the same country information as ERIH-PLUS **Country of Publication.**
This is done by a python function that does the following:

1. Selects all ISSN in ERIH-PLUS that have no country
2. checks if the ISSN exists in DOAJ dataset
3. if yes, retrieves the country information in the Country of Publisher column
4. add that information to ERIH-PLUS filtered dataset.

	Country	Venue_OMID
	Spain	...
	Italy	identifier
	Unknown	identifier

Create our final Datasets

3 By merging mergeing **the disciplines** dataset and the **countries** dataset with **merged_data**, by means of the **unique OMID,** we add to the dataset the publication count for each venue.

The output datasets will look like this:

	ERIH-PLUS_Discipline	OMID	Publication_Count
	Cultural Studies	...	12
	Art and Art History	identifier	12
	Art and Art History	identifier	65
	Cultural Studies	identifier	
	Human Geography and Urban Studies		
	Human Geography and Urban Studies		

Disciplines_df

	Country	OMID	Publication_count
	Spain	identifier	12
	Italy	identifier	65

Countries_df

Final counts

4 1. To find the countries that have the highest number of publication we will sum the values in Publication_count column for every country.

2. This information will be saved in a dictionary that has the Country name as key and the total count of publication for each country as value.

3. To find the disciplines with the highest number of publication we will follow the same process but it will need to be done with both of the disciplines classifications.

4. To answer to our research question "How many of the SSH journals are available in Open Access according to the data in DOAJ?"

We will simply count the rows that have a "yes" value in the Open Access column.

Visualize results

5 For each visualization, we use Python libraries like Matplotlib, Seaborn, Plotly to create the charts and maps.

Analyzing the coverage of publications in SSH journals

- Visualization: Bar chart showing the number of publications for each journal in the ERIH-PLUS list. The x-axis represents the journals, and the y-axis represents the number of publications.

Identifying the disciplines with the most publications

- Visualization: Bar chart showing the number of publications for each discipline. The x-axis represents the disciplines, and the y-axis represents the number of publications.

Identifying the countries providing the largest number of publications and journals

- Visualization 1: Choropleth map showing the number of publications by country. Each country is colored according to the number of publications, with darker colors representing higher publication counts.
- Visualization 2: Bar chart showing the number of journals for each country. The x-axis represents the countries, and the y-axis represents the number of journals.

Fetching the list of Open Access journals from DOAJ

- Visualization: Pie chart showing the proportion of Open Access SSH journals (according to ERIH-PLUS) based on DOAJ data. The chart will have two segments: Open Access journals and Non-Open Access journals.