# first-workflow-Playarists

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

[1]unibo

Ali Ghasempouri: dhdk
maddalena.ghiotto: dhdk
sebastiano.giacomini: dhdk

Ali Ghasempouri

APR 13, 2023

**DOI:**
dx.doi.org/10.17504/protocols.io.5jyl8jo1rg2w/v1

**Protocol Citation:** Ali Ghasempouri, maddalena.ghiotto, sebastiano.giacomini 2023. first-workflow-Playarists. **protocols.io** https://dx.doi.org/10.17504/protocols.io.5jyl8jo1rg2w/v1

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

**Created:** Mar 30, 2023

**Last Modified:** Apr 13, 2023

**PROTOCOL integer ID:** 79752

## 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only and does not constitute legal, medical, clinical, or safety advice, or otherwise; content added to protocols.io is not peer reviewed and may not have undergone a formal approval of any kind.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.

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.

## 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 23/03/2023) we want to retreive all the publications belonging to one of those journals, included in OpenCitation Meta database (https://opencitations.net/meta#:~:text=For%20each%20publication%2C%20the%20metadata,and%20PubMed%20Identifiers%20(PMIDs).)

1.1 We are currently evaluating what would be the most efficient methodology to do that, given the size of OpenCitation Meta data dump (56GB, 8GB zipped).
Current options are:

- Iterating over ERIH-PLUS issn list and request all publications in OpenCitation Meta by means of its SPARQL endpoint. This would include using SPARQLwrapper python library
- Downloading the data dump and performing chunk operations (either reading the csv with pandas setting a chunksize parameter, using os library to iterate over the folder's files, reading directly the zip file using gzip library etc.)

Note that the OC data dump has a row for each entity that is either a publication or a venue. At this moment we don't need publication information, so we would need to cut down the dataset to only have venues information in it.

Note

**Input:** ERI-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 | |

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

Note

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

| OC_omid | OC_issn | EP_id | EP_issn |
|---|---|---|---|
| meta:br/060167 | issn:1865-3804 | 503890 | 1865-3804 |
| meta:br/060167 | issn:4522-4592 | 503890 | 4522-4592 |
| meta:br/060167 | isbn:242352513 | NaN | NaN |

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

## 1.2

Note

**HERE WE NEED TO HAVE A STEP FOR ADDING INFORMATION ABOUT OPEN ACCESS TO THE DATAFRAME WE JUST CREATED, SO THAT THE OMIDS ARE DIRECTLY CONNECTED TO THE INFORMATION ABOUT ACCESSIBILITY OF THE JOURNAL!**

**Adding Open Access information to the dataframe**

**Input:** Dataset mapping OpenCitation venue data to ERIH-PLUS venue data, dataset containing country and discipline information for each journal, list of Open Access journals from DOAJ

**Output:** Updated dataframe with an additional column indicating whether the journal is Open Access or not

**Fetch Open Access information from DOAJ**

- Query the DOAJ API or download the dataset to obtain a list of Open Access journals and their ISSNs.

**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/False) 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:

main_dataframe['Open Access'] = main_dataframe['ISSN'].map(open_access_dict)
main_dataframe with the name of our dataframe and open_access_dict the name of our dictionary containing Open Access information.

**Fill missing Open Access information with 'False'**

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

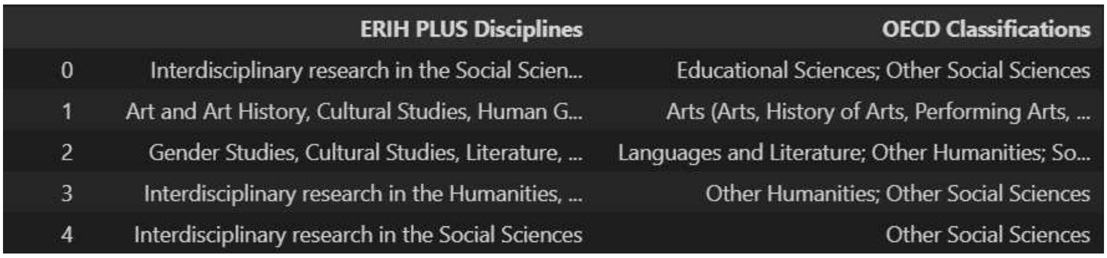
1.3 For the sake of easing the next steps we will create a second version of the ERIH-PLUS dataset, filtering out all journals that are not in OC meta. This will be called **ERIH-PLUS_filtered.** This dataset will also have the OMID in it to be able to connect it with out dataframes that maps issns

## Retrieve data about countries and disciplines

2 Our second and third research question are "what are the disciplines that have more publications? What are countries providing the largest number of publications and journals?" so we need to include in our dataframe information about each journal country and the discipline it belongs to.

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

## 2.1 Disciplines:



| | ERIH PLUS Disciplines | OECD Classifications |
|---|---|---|
| 0 | Interdisciplinary research in the Social Scien... | Educational Sciences; Other Social Sciences |
| 1 | Art and Art History, Cultural Studies, Human G... | Arts (Arts, History of Arts, Performing Arts, ... |
| 2 | Gender Studies, Cultural Studies, Literature, ... | Languages and Literature; Other Humanities; So... |
| 3 | Interdisciplinary research in the Humanities, ... | Other Humanities; Other Social Sciences |
| 4 | Interdisciplinary research in the Social Sciences | Other Social Sciences |

As we can see every journal can have multiple disciplines:
The ERIH-PUS  Disciplines are separated with **,**
OECD Classifications disciplines are separated with **;**

In order to count the disciplines and understand which one has the highest number of publications we will need to disassemble them and map both classifications in order to prevent information loss on the two classifications.
Luckily, we don't have too many disciplines to map, given that ERIH-PLUS has 30 disciplines and OECD has 14.

To create the table we need we will write a python function that takes in input the ERIH-PLUS filtered dataset

Note

**Output:**

| ERIH-PLUS Disciplines | OECD Classification | OMID of the journal |
|---|---|---|
| Cultural Studies | Other Humanities | ... |
| Art and Art History | Arts (Arts, History of Arts, Performing Arts, Music) | identifier |
| Art and Art History | Other Humanities | identifier |
| Cultural Studies | Arts (Arts, History of Arts, Performing Arts, Music) | identifier |
| Human Geography and Urban Studies | Arts (Arts, History of Arts, Performing Arts, Music) | |
| Human Geography and Urban Studies | Other Humanities | |

2.2 **Countries:**

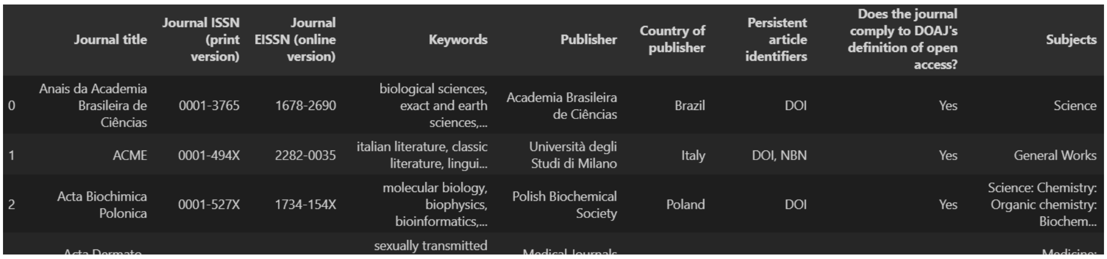
For the countries we will do the same.

This is relatively less complicated since, from a first exploration, there are not multiple countries in the ERIH-PLUS filtered dataset.

We will 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.

DOAJ dataset has the following structure:



| | Journal title | Journal ISSN (print version) | Journal EISSN (online version) | Keywords | Publisher | Country of publisher | Persistent article identifiers | Does the journal comply to DOAJ's definition of open access? | Subjects |
|---|---|---|---|---|---|---|---|---|---|
| 0 | Anais da Academia Brasileira de Ciências | 0001-3765 | 1678-2690 | biological sciences, exact and earth sciences,... | Academia Brasileira de Ciências | Brazil | DOI | Yes | Science |
| 1 | ACME | 0001-494X | 2282-0035 | italian literature, classic literature, lingui... | Università degli Studi di Milano | Italy | DOI, NBN | Yes | General Works |
| 2 | Acta Biochimica Polonica | 0001-527X | 1734-154X | molecular biology, biophysics, bioinformatics,... | Polish Biochemical Society | Poland | DOI | Yes | Science: Chemistry: Organic chemistry: Biochem... |
| | Acta Dermato | | | sexually transmitted | Medical Journals | | | | Medicine: |

where, **Country of Publisher** contains the same country information as ERIH-PLUS **Country of Publication.**

We could write 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.

## Create our final Dataset: name of dataset

3 We will merge our two datasets, **the disciplines one and the countries one**, by means of the **unique OMID,** finally we will also add a column stating wether the journal is open access or not. Here is how we envision the final dataset to look like:

| OMID identifier | ERIH-PLUS Discipline | OECD discipline | Country | Open_access |
|---|---|---|---|---|
| | | | | |

Also DOAJ has a "subject" column with a discipline classification. We need to discuss if we want to include that or not.

## Retrieve Publication information

4 We now need to count:

1. Number of publication for each SSH journal in OCMeta
2. Total number of SSH Publications in OCMeta

We could do this by means of a **sparql query** to OpenCitations endpoint
https://opencitations.net/meta/sparql
for each OMID we will ask for the number of publication and we can save this information adding a column in the final dataset.

| OMID identifier | Publication_count | ERIH-PLUS Discipline | OECD Discipline | Country | Open_access |
|---|---|---|---|---|---|
| | | | | | |

4.1 To find out the coverage of SSH publication we will sum the values in Publication_count column
We still need to find out how to retrieve the total number of publication a Journal has, that has to be found somehow in ERIH-PLUS. Once we have that we can compute a percentage.

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

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.

4.3 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.

4.4 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.