Retrieving SSH Journals Citation Information from three datasets (COCI, META and ERIH-PLUS) - Workflow V.2

Marta Soricetti[1], Sara Vellone[1], Olga Pagnotta[1], Lorenzo Paolini[1]

[1]University of Bologna

Sara Vellone

VERSION 2

MAY 10, 2023

OPEN ACCESS

DOI: dx.doi.org/10.17504/protocols.io.n92ldpeenl5b/v2

Protocol Citation: Marta Soricetti, Sara Vellone, Olga Pagnotta, Lorenzo Paolini 2023. Retrieving SSH Journals Citation Information from three datasets (COCI, META and ERIH-PLUS) - Workflow. **protocols.io** https://dx.doi.org/10.17504/protocols.io.n92ldpeenl5b/v2Version created by Olga Pagnotta

Protocol status: Working
We use this protocol and it's working

Created: Apr 22, 2023

Last Modified: May 10, 2023

PROTOCOL integer ID: 80913

Keywords: OpenScience, Citation, OC-COCI, OC-Meta, ERIH-PLUS, journals

ABSTRACT

Purpose

The main purpose of our research is to answer to three different question and find out:

- by looking at citations data contained in COCI, the number of citations included in Meta which refer to publication in SSH (Social Sciences and Humanities) journals indicated in ERIH-PLUS
- the disciplines citing the most VS the disciplines cited the most
- the citations from/to publication contained in Meta which are not included in SSH journals

We want to create a connection between these three different datasets for having an overall view of the citations present in each of them.

Methodology

For this reason, we approach the problem from a computational point of view. We started by extracting only the relevant data by operating a first processing of COCI, ERIH-PLUS and META's datasets. Then we built a python software able to analyze the CSV format data, querying them in order to retrieve the info needed and to present the results in a clear and understandable way, which will be available in CSV format.

Findings

The findings show that the majority of citations come from and go to Psychology publications, and a good amount of citations according to the three datasets is involved in SSH publications.

Originality/Value

The research conducted by us has the purpose to add information to existing resources with the aim of facilitating their use and allowing the researchers to have a clearer view of the data contained in each dataset. Further development can be made, for example analyzing other disciplines, to have the same overview as the one created by us but related to other fields.

GUIDELINES

The required Python version for running the current software is Python 3.10.
The library CSVManager needs to be manually installed by downloading the corresponding folder from the linked Github repository.

1 We started to analyse the datasets using pandas:

- COCI[1]: COCI dump
- Meta[2]: csv dataset of Open Citations Meta
- ERIH-PLUS: list of approved journals

Dataset

COCI NAME

https://opencitations.net/download#coci LINK

Dataset

META NAME

https://opencitations.net/download#meta LINK

Dataset

ERIH-PLUS (approved journals) NAME

https://kanalregister.hkdir.no/publiseringskanaler/erihplus/periodical/listApproved LINK

[1] v19 (released on January 2023)

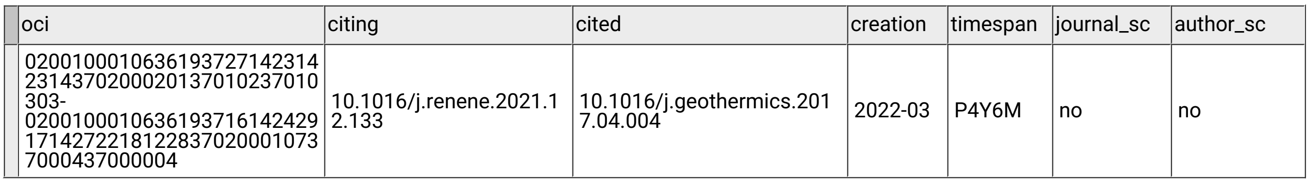
[2] v3 (released on February 2023)

1.1 *COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations*

COCI is an RDF dataset containing details of all the citations that are specified by the open references to DOI-identified works present in Crossref.
It includes, in this latest version: 1,463,920,523 citations and 77,045,952 bibliographic resources.

The following are the names of the column of the dataset:

- **[field "oci"]**the Open Citation Identifier (OCI) for the citation;
- **[field "citing"]**the DOI of the citing entity;
- **[field "cited"]**the DOI of the cited entity;
- **[field "creation"]**the creation date of the citation (i.e. the publication date of the citing entity);
- **[field "timespan"]**the time span of the citation (i.e. the interval between the publication date of the cited entity and the publication date of the citing entity);
- **[field "journal_sc"]**it records whether the citation is a journal self-citations (i.e. the citing and the cited entities are published in the same journal);
- **[field "author_sc"]**it records whether the citation is an author self-citation (i.e. the citing and the cited entities have at least one author in common).



oci	citing	cited	creation	timespan	journal_sc	author_sc
02001000106361937271423142314370200020137010237010303-02001000106361937161424291714272218122837020001073700043700004	10.1016/j.renene.2021.12.133	10.1016/j.geothermics.2017.04.004	2022-03	P4Y6M	no	no

1.2 *META*

OpenCitations Meta stores and delivers bibliographic metadata for all publications involved in the OpenCitations Indexes.
It includes, in this latest version: 90,102,757 bibliographic entities; 282,247,615 authors and 2,367,265 editors; 644,830 publication venues

and 18,397 publishers.

The following are the names of the column of the dataset:

- **[field "id"]**the IDs for the document described within the line;
- **[field "title"]**the document's title;
- **[field "author"]**the authors of the document;
- **[field "pub_date"]**the date of publication;
- **[field "venue"]**information about the venue, i.e. the bibliographical resource to which the document belongs;
- **[field "volume"]**the volume sequence identifier (e.g. a number) to which the entity belongs;
- **[field "issue"]**the issue sequence identifier (e.g. a number) to which the entity belongs;
- **[field "page"]**the page range of the resource described in the row;
- **[field "type"]**the type of resource described in the row;
- **[field "publisher"]**the entity responsible for making the resource available;
- **[field "editor"]**the editors of the document.

id	title	author	issue	volume	venue	page	pub_date	type	publisher	editor
"meta:br/060209 doi:10.4230/lipics.approx/random.2020.19"	"Distributed Testing Of Graph Isomorphism In The CONGEST Model"	"Levi, Reut [meta:ra/0610110096 orcid:0000-0003-3167-1766]; Medina, Moti [meta:ra/0612046435 orcid:0000-0002-5572-3754]"	""	""	" [meta:br/060182 issn:1868-8969]"	""	"2020"	"report"	"Schloss Dagstuhl - Leibniz-Zentrum Für Informatik [meta:ra/0605251]"	"Byrka, Jarosław [meta:ra/069044096 orcid:0000-0002-3387-0913]; Raghu Meka [meta:ra/0605252]"

1.3 *ERIH-PLUS (list of approved journals)*

ERIH PLUS is an academic journal index for the HSS (Humanities and Social Sciences) society in Europe.
This version (22/4/23) contains 11120 records.

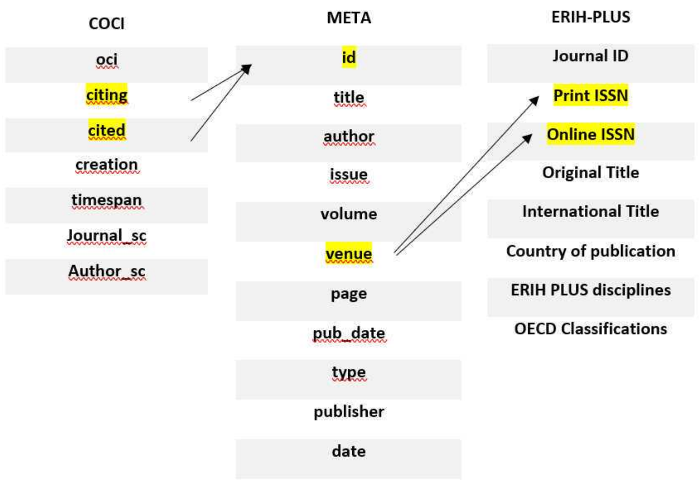
The following are the names of the column of the dataset:

- **Journal ID**
- **Print ISSN**
- **Online ISSN**
- **Original Title**
- **International Title**
- **Country of Publication**
- **ERIH PLUS Disciplines**
- **OECD Classifications**
- **[Last Updated],** which contains data about the last update of each journal in the dataset

Journal ID	Print ISSN	Online ISSN	Original Title	International Title	Country of Publication	ERIH PLUS Disciplines	OECD Classifications	[Last Updated]
488138	1392-4095	2351-6526	Acta Historica Universitatis Klaipedensis	Acta Historica Universitatis Klaipedensis	Lithuania	History	History and Archaeology	02/02/2023 17:14:12

Processing of Input Data

2 We tried to define a mapping of the datasets to understand which are the information that the three datasets have in common. By looking at the data and the column names, we have identified the following:



Mapping of the three datasets and relevant columns

Taking as a starting point META, we have identified the COCI columns "citing" and "cited" as corresponding to DOIs, which are contained also in the "id" column of META.
For what concerns EIRH-PLUS, we have noticed that "print ISSN" and "online ISSN" correspond to the "venue" column of META.

We have processed the data by filtering and cleaning them, keeping only the relevant information for the research purpose.

Preprocessing

For META and COCI we have reused some methods of the OpenCitations Preprocessing software

Software

Preprocess NAME

OpenCitations DEVELOPER

adapting it to our needs.

In particular, we have created two classes: MetaPreProcessing and CociPreProcessing. In these Classes, we have reused a method, splitted_to_file, taken from OpenCitations Preprocess, which we have adapted to the specific class in which we have used it. The general structure of the method is the following: it takes in input n, the integer number of preprocessed entities which passed the preprocessor filter; a data structure containing the preprocessed entities to store in the output file and the string of the file extension.

In order to retrieve the information from the zipped datasets, we have created a method for each of the above mentioned classes, split_input, which is the main function for reading, filtering and cleaning the data and storing the processed version of the data in the output files. In particular, it reads the content of the .zip and adds to a list the names of the files inside the original dataset folder: if the file is a CSV, it opens it and reads its content to process it and write the new output file.
The library used is *zipfile,* in particular the methods .ZipFile, .namelist, .open.

2.1 META

We have decided to use META as central object for the filtering and cleaning processes, and thus we will start from it in order to answer the research questions.

However, we have performed some filtering also on this dataset, cleaning up unnecessary information.

In the class MetaPreProcessing we manage the preprocessing of the META dump.

For the columns "id" and "venue" of the original files we have decided to keep only the DOIs and the ISSNs, removing thus all the other identifiers specified for each entity in META, obtaining the following structure:

id	title	author	issue	volume	venue	page	pub_date	type	publisher	editor
"doi:10.4230/lipics.approx/random.2020.19"	"Distributed Testing Of Graph Isomorphism In The CONGEST Model"	"Levi, Reut [meta:ra/0610110096 orcid:0000-0003-3167-1766]; Medina, Moti [meta:ra/0612046435 orcid:0000-0002-5572-3754]"	""	""	"issn:1868-8969"	""	"2020"	"report"	"Schloss Dagstuhl - Leibniz-Zentrum Für Informatik [meta:ra/0605251]"	"Byrka, Jarosław [meta:ra/069044096 orcid:0000-0002-3387-0913]; Raghu Meka [meta:ra/0605252]"

the data in bold are those we have cleaned

2.2 ERIH-PLUS

Since this dataset was made by just one csv file, we have created a simpler script for filtering, cleaning and storing the new dataset.
It takes as input the path of the original csv and first unifies the Print ISSN and the Online ISSN in one string value, stored in the column "venue_id" and then it considers only the column ERIH PLUS Discipline and created the following dataset:

venue_id	ERIH_disciplines
"issn:1392-4095 issn:2351-6526"	History

2.3 COCI

In the class CociPreProcessing we manage the preprocessing of the COCI dump.

After the preprocessing, we will keep only the citing and cited columns, having added the prefix "doi:" to the two values.

citing	cited
"doi:10.1016/j.renene.2021.12.133"	"doi:10.1016/j.geothermics.2017.04.004"

Further processing of Data

3 The previous steps allowed us to create three cleaned datasets, each composed by different CSV files: META_preprocessed (8438 CSVs), COCI_preprocessed (122 CSVs), ERIH_preprocessed (1 CSV). To answer the research questions we have performed some further operations upon META and ERIH: we have merged META_preprocessed dataset together with ERIH_preprocessed, obtaining thus 7622 new CSVs having as columns all the columns of META with the addition of the ERIH-PLUS disciplines.
We have then created two sub-datasets of ERIH_META, one containing all the DOIs contained in SSH journals, and the other one with all the DOIs not contained in SSH journals, erih_meta_with_disciplines and erih_meta_without_disciplines.

3.1 *ERIH_META*

The merged dataset obtained with META_preprocessed and ERIH_preprocessed has been obtained by creating a new class, Erih_Meta, which has four internal methods. The method get_all_files is responsible for reading the input files and creating a list containing them. The method splitted_to_files has the same structure as the one described in 2. Processing of Input Data. The method find_erih_venue, which returns the corresponding ERIH-PLUS disciplines given an input ISSN list, composed of ISSN taken from META_preprocessed column "venue". Within this method we have used the class CSVManager of OpenCitations

Software

OCMeta, CSVManager NAME

https://github.com/opencitations/oc_meta SOURCE LINK

and in particular the method get_value, adapting the whole class to the structure of ERIH_preprocessed dataset.

The method find_erih_venue is also used in the erih_meta method to fill in the column "erih_disciplines" of ERIH_META. If for a given ISSN the corresponding discipline is found, it is added to the column; if not, the method writes an empty string in the erih_disciplines column. In addition to that, the method erih_meta writes the new CSVs, having the following structure:

"id"	"title"	"author"	"issue"	"volume"	"venue"	"page"	"pub_date"	"type"	"publisher"	"editor"	"erih_disciplines"
"doi:10.1207/s15327078in0502_6"	"An Emerging Consensus: Younger And Cohen Revisited"	"Younger, Barbara A. [meta:ra/062105338915]; Hollich, George [meta:ra/062105338916]; Furrer, Stephanie D. [meta:ra/062105338917]"	"2"	"5.0"	"issn:1525-0008 issn:1532-7078"	"209-216"	"2004-03-01"	"journal article"	"Wiley [meta:ra/0610116001 crossref:311]"	""	"Psychology"

3.2 *erih_with_disciplines*

Starting from ERIH_META we have created a subset with two columns: "id" (the DOIs) and "erih_disciplines". It contains the publications belonging to SSH journals, according to ERIH_PLUS.

"id"	"erih_disciplines"
"doi:10.12759/hsr.46.2021.1.181-205"	"History, Interdisciplinary research in the Humanities, Interdisciplinary research in the Social Sciences, Sociology"

3.3 *erih_without_disciplines*

Starting from ERIH_META we have created a subset with one column: "id" (the DOIs). It contains the publications NOT belonging to SSH journals.

"id"
"doi:10.4230/lipics.approx/random.2020.19"

Answering to the research questions

4 Our research questions:

1. How many citations (according to COCI) involve, either as citing or cited entities, publications in SSH journals (according to ERIH-PLUS) included in OpenCitations META?
2. What are the disciplines that cites the most and those cited the most?
3. How many citations start from and go to publications in OpenCitations META that are not included in SSH journals?

4.1 ***The first question***

We have created a Class CountCitations, which has four methods: get_all_files, splitted_to_file, create_citations_map, count_citations. The first method is responsible for reading the input files and creating a list containing them; the second method has the same structure and function of the one described in 2. Processing of Input Data. The method create_citations_maps is responsible of the iteration in COCI_preprocessed and populates two dictionaries, which are used to create new CSV files, with two columns: "citing" and "cited". In the citing column all the COCI_preprocessed citing DOIs which are present in erih_meta_with_disciplines are inserted, while in the cited column are inserted the cited DOIs of COCI_preprocessed, if present in erih_meta_with_disciplines.

For searching the values in erih_meta_with_disciplines we use the class CSVManager, adapted to the structure of erih_meta_with_disciplines dataset, and in particular the method get_value.

The method count_citations is responsible of counting the number of rows in each CSVs produced in the previous step.

4.2 ***The second question***

We have created the class CountDisciplines, with five methods: get_all_files, splitted_to_file, create_disciplines_map, split_disciplines, create_count_dictionaries. The first two methods have the same structure and function of the first question. The method create_disciplines_map creates new CSVs with four columns: "id", "citing", "cited", "disciplines". It starts by iterating in COCI_preprocessed and if the DOI is found in erih_meta_with_disciplines, it is inserted in the output CSV in the "id" column, the corresponding discipline is written in the "disciplines" column, and either the column "cited" or "citing" is filled with True, depending on the nature of the DOI in COCI, and the opposite column is filled with False (e.g., if it is citing, the "citing" column will be filled with True and the "cited" with False). The method split_disciplines is in charge of splitting in different rows of the output CSVs the disciplines, if they are more than one. In this case, also the other columns are repeated. The method create_count_dictionaries starts from the previous step's output CSV and creates two dictionaries, dict_cited and dict_citing, having as keys the disciplines retrieved from the column "disciplines" and basing on whether they have the column "citing" or "cited" as True, they are inserted in the corresponding dictionary. The value is the number of times that discipline appears in the CSV.

4.3 ***The third question***

We have created the class CountCitationsNoSSH, with four methods: get_all_files, splitted_to_file, create_citations_map, count_citations. The first two methods have the same structure and function of the first question. The method create_citations_maps is responsible of the iteration in COCI_preprocessed and populates two dictionaries, which are used to create new CSV files, with two columns: "citing" and "cited". In the citing column all the COCI_preprocessed citing DOIs which are present in erih_meta_without_disciplines are inserted, while in the cited column are inserted the cited DOIs of COCI_preprocessed, if present in erih_meta_without_disciplines.

For searching the values in erih_meta_without_disciplines we use the class CSVManager, adapted to the structure of erih_meta_without_disciplines dataset, and in particular the method get_value.

The method count_citations is responsible of counting the number of rows in each CSVs produced in the previous step.

Results

5 The answers to the three questions gives significant insights on the current situations within publications in SSH Journals.

5.1 ***The answer to the first question***

How many citations (according to COCI) involve, either as citing or cited entities, publications in SSH journals (according to ERIH-PLUS) included in OpenCitations META?

The citations which involve publications in SSH journals according to the three datasets are 272179152.

5.2 ***The answer to the second question***

What are the disciplines that cites the most and those cited the most?

The discipline citing the most is Psychology with 55344192 citations, while the discipline cited the most is still Psychology with 84921140.

5.3 ***The answer to the third question***

How many citations start from and go to publications in OpenCitations **META** that are not included in SSH journals?

We are still processing the result of the third question.