May 08, 2024 Version 2

IRIS Software Protocol V.2

DOI
dx.doi.org/10.17504/protocols.io.3byl497wjgo5/v2

Leonardo Zilli[1], Erica Andreose[1], Salvatore Di Marzo[1]

[1]University of Bologna

Leonardo Zilli
University of Bologna

OPEN ACCESS

DOI: dx.doi.org/10.17504/protocols.io.3byl497wjgo5/v2

Protocol Citation: Leonardo Zilli, Erica Andreose, Salvatore Di Marzo 2024. IRIS Software Protocol. **protocols.io** **https://dx.doi.org/10.17504/protocols.io.3byl497wjgo5/v2**Version created by **Leonardo Zilli**

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

Created: April 12, 2024

Last Modified: May 08, 2024

Protocol Integer ID: 99398

Funders Acknowledgement:
University of Bologna
Grant ID:
https://ror.org/01111rn36

Abstract

This project aims to investigates the discoverability of University of Bologna's scholarly output within OpenCitations Meta, a platform that stores and delivers bibliographic metadata for alle the publications involved in OpenCitations Index which contains recording citations allowing the user to visualize all of the citation links between a document and another. Specifically, this project aims to analyze the coverage of University of Bologna's publications within the OpenCitations Meta, particularly those deposited in the IRIS institutional repository. Various publication types are observed to discern patterns in representation. Utilizing OpenCitations data, the citation impact of these IRIS publications is quantified, encompassing both the citations they receive and those they provide. Additionally, the study delves into the nature of citations, distinguishing between those involving publications within IRIS and those referencing external sources. The study unveils the citation impact of IRIS publications, clarifying their influence within the scholarly community. By dissecting citations, the research also delineates the interplay between internal and external citations, providing a deeper understanding of the institution's scholarly ecosystem. The methodology used will be all compliant to the core values of Open Science, the data and the software used will all be made available to allow the user to replicate, reproduce and validate the conclusions drawn by the end results of the project, that will hold significance for researchers and academic institutions, facilitating informed decision-making and fostering a deeper understanding of scholarly communication dynamics.

Guidelines

To let the user with complete reproducibility of the protocol, links to the data used are provided here.
Download link for the UNIBO IRIS bibliographic data dump, dated 14 March 2024:
https://amsacta.unibo.it/id/eprint/7608/
OpenCitations Meta can be accessed and queried through a REST API: https://w3id.org/oc/meta/api/v1

The code used for this research, along with the data produced to answer to the research questions can be found in the **github repository**.

Safety warnings

It is advised / required to run the code on a machine that has at least 16gb of RAM memory available.

Before start

Before starting, we suggest to make sure you have Python3.x installed on your computer, in addition, in order to correctly execute the provided scripts, you must install the required libraries:
requirements.txt

IRIS Dataset Preparation

1

Dataset

UNIBO IRIS bibliographic data dump, dated 14 March 2024 NAME

https://amsacta.unibo.it/7608/1/iris-data-2024-03-14.zip LINK

After downloading the IRIS dataset, the contents appear divided into 7 different csv files, descriptive as the following:
- "ODS_L1_IR_ITEM_CON_PERSON.csv": information about the people involved in the publications (authors, editors, etc.)
- "ODS_L1_IR_ITEM_DESCRIPTION.csv": the string containing the name of the authors and other related metadata of publications
- "ODS_L1_IR_ITEM_IDENTIFIER.csv": the identifiers (including DOIs) of publications
- "ODS_L1_IR_ITEM_LANGUAGE.csv": the language in which the publication has been written (when applicable)
- "ODS_L1_IR_ITEM_MASTER_ALL.csv": basic metadata information of publications (title and date of publication)
- "ODS_L1_IR_ITEM_PUBLISHER.csv": the publishers of publications
- "ODS_L1_IR_ITEM_RELATION.csv": additional metadata related to the context of publications (publication venue, editors, etc.)

These files are connected among themselves through a unique item ID tied to each entry, after we gathered the files we converted them into dataframes with the pandas[1] library.

[1](**https://pandas.pydata.org/**)

2 2 of the csv files are converted into Dataframes

```
df_iris_master = pd.read_csv('./data/iris-data-2024-03-
14/ODS_L1_IR_ITEM_MASTER_ALL.csv')
df_iris_identifier = pd.read_csv('./data/iris-data-2024-03-
14/ODS_L1_IR_ITEM_IDENTIFIER.csv')
```

3 The DataFrames are filtered to remove rows that do not have neither DOI or ISBN ids.

```
df_iris_identifier_filtered =
df_iris_identifier[(df_iris_identifier['IDE_DOI'].notna()) |
(df_iris_identifier['IDE_ISBN'].notna())][['ITEM_ID', 'IDE_DOI',
'IDE_ISBN']]
```

Expected result

	ITEM_ID	IDE_DOI	IDE_ISBN
3	60479	nan	8883125150
8	82956	nan	88.387.3686.3; 88.387.3687.1
9	70006	10.1441/13328	nan
15	59231	10.1111/j.1468-3083.2005.01282.x	nan
30	73478	10.2110/palo.2005.p05-020r	nan

4 The df_iris_identifier DataFrame is joined with the master dataframe to append the title and the date of publication of the publications to the df_iris_identifier_filtered

```
df = df_iris_identifier_filtered.merge(df_iris_master,
on='ITEM_ID')
```

OpenCitations Meta Dump preparation

5 OpenCitations Meta has been queried through the use of the OpenCitations Meta april 2024 dump.

CITATION

OpenCitations (2024). OpenCitations Meta CSV dataset of all bibliographic metadata. figshare..
LINK
https://doi.org/10.6084/m9.figshare.21747461.v8

After downloading the zip file, the csv files contained in the dataset are turned into parquet file for easier handling:

```
openalex_zip = ZipFile('./data/csv_openalex-2024-04-06.zip')

for file in tqdm(openalex_zip.namelist()):
    if file.endswith('.csv'):
        with openalex_zip.open(file) as csv_file:
            with tempfile.NamedTemporaryFile() as tf:
                tf.write(csv_file.read())
                tf.seek(0)

Path("./data/openalex_parquet").mkdir(parents=True, exist_ok=True)
                lf = (
                    pl.scan_csv(tf.name)
                    .select(['id', 'title', 'author', 'type'])

.sink_parquet('./data/openalex_parquet/{}.parquet'.format(file.spl
it('/')[-1].replace(".csv", "")))
                )
```

6 The parquet files are then read by using polars' data streaming capabilities:

```
parquet_files = glob.glob('./data/openalex_parquet/*.parquet')

meta_lf = (
    pl.scan_parquet(parquet_files)
    )
```

ID sanification

7 The DOIs and ISBNs from the IRIS dataset are quite dirty and need to be cleaned.

```
dois = df['IDE_DOI'].dropna().unique().tolist()

#filter and normalize the dois
doi_rule = re.compile(r'10\.\d{4,}\/[^,\s;]*')
not_doi = []
filtered_dois = []

for doi in dois:
    match = doi_rule.search(doi)
    if match:
        filtered_dois.append('doi:' + match.group())
    else:
        not_doi.append(doi)
```

```
isbns = df['IDE_ISBN'].dropna().unique().tolist()

#filter and normalize the isbns
isbn_rule = re.compile(r'(ISBN[-]*(1[03])*[ ]*(: ){0,1})*(([0-9Xx]
[- ]*){13}|([0-9Xx][- ]*){10})') # ??? results to check
not_isbn = []
filtered_isbns = []

for isbn in isbns:
    if isbn_rule.search(isbn) is not None:
        filtered_isbns.append('isbn:' + isbn.replace('-',
'').replace(' ', ''))
    else:
        not_isbn.append(isbn)
```

The two identifiers are then merged into a single list:

```
dois_isbns = filtered_dois + filtered_isbns
```

RQ 1. What is the coverage of the publications available in IRIS (strictly concerning research conducted within the University of Bologna) in OpenCitations Meta?

8

```
rq1_query = (
    pl.scan_parquet(parquet_files, low_memory=True)
    .select(['id', 'type'])
    .with_columns(
        (pl.col('id').str.extract(r"((?:doi|isbn):[^\s]+)"))
    )
    .select(['omid', 'id', 'type'])
    .drop_nulls('id')
    .filter(
        pl.col("id").is_in(dois_isbns)
    )
    .select(pl.len()).collect()
)

print(rq1_query.item())
```

OpenCitations Index Dataset Querying

9

Dataset	
OpenCitations Index	NAME
https://opencitations.net/index/sparql	LINK

The OpenCitations Index Sparql endpoint allowed us after querying it, to download a csv file of the results of our queries which we then transformed into dataframes.

These are the queries that gave as a result the files:

```
# 1 Sparql query
```

```
# 2 Sparql query
```

```
# 3 Sparql query
```

9.1 We then use these csv files to create the pandas DataFrames that will be used to perform the analyses.

```
index_df = pd.read_csv(f'index.csv')
```

Research Question answering

9.2
- *What is the coverage of the publications available in IRIS (strictly concerning research conducted within the University of Bologna) in OpenCitations Meta?*

```
filter the IRIS dataframe for publications conducted at UNIBO
check the intersection of the DOIs from the iris_df with the
meta_df
```

9.3
- *Which are the types of publications that are better covered in OpenCitations Meta?*

```
checking the type of the resulting data of the first question
```

9.4
- *What is the amount of citations (according to OpenCitations Index) coming from the IRIS publications that are involved in OpenCitations Meta (as citing entity and as cited entity)?*

```
checking the number of entities coming from IRIS, stored in
OpenCitations Meta, that cite or are cited inside of OpenCitations
Index
```

9.5 ▪ *How many of these citations come from and go to publications that are not included in IRIS?*

```
check the results of the third question to find which citations
cite entries not available in IRIS, and which entries in IRIS cite
entries not available in IRIS
```

9.6 ▪ *How many of these citations involve publications in IRIS as both citing and cited entities?*

```
check the results of the third quesion to find how many citations
in OpenCitations Index involve the publications in IRIS that are
both acting as citing and cited
```

Data Visualization

10 The visualizations of the results obtained are computed with the _ _ library.

Note

What is the coverage of the publications available in IRIS (strictly concerning research conducted within the University of Bologna) in OpenCitations Meta?

-Insert graph

Note

Which are the types of publications that are better covered in OpenCitations Meta?

-Insert graph

Note

What is the amount of citations (according to OpenCitations Index) coming from the IRIS publications that are involved in OpenCitations Meta (as citing entity and as cited entity)?

-Insert graph

Note

How many of these citations come from and go to publications that are not included in IRIS?

-Insert graph

Note

How many of these citations involve publications in IRIS as both citing and cited entities?

-Insert graph

Citations

Step 5

OpenCitations . OpenCitations Meta CSV dataset of all bibliographic metadata. figshare. **https://doi.org/10.6084/m9.figshare.21747461.v8**