

Open Science for Social Sciences and Humanities: Open Access availability and distribution across disciplines and Countries in OpenCitations Meta

Seyedali Ghasempouril[0000–0003–3446–1115], Maddalena Ghiotto1[0009–0009–1309–6340],
Sebastiano Giacomini1[0009–0007–7813–0939]

Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy

Abstract

Purpose: this study aims to investigate the representation and distribution of Social Science and Humanities (SSH) journals within the OpenCitations Meta database, with a particular emphasis on their Open Access (OA) status, as well as their spread across different disciplines and countries. The underlying premise is that open infrastructures play a pivotal role in promoting transparency, reproducibility, and trust in scientific research.

Study Design and Methodology: the study is grounded on the premise that open infrastructures are crucial for ensuring transparency, reproducibility, and fostering trust in scientific research. The research methodology involved the use of secondary data sources, namely the OpenCitations Meta database, the ERIH PLUS bibliographic index, and the DOAJ index. A custom research software was developed in Python to facilitate the processing and analysis of the data.

Findings: the results reveal that 78.1% of SSH journals listed in the European Reference Index for the Humanities (ERIH-PLUS) are included in the OpenCitations Meta database. The discipline of Psychology has the highest number of publications. The United States and the United Kingdom are the leading contributors in terms of the number of publications. However, the study also uncovers that only 38% of the SSH journals in the OpenCitations Meta database are OA.

Originality: this research adds to the existing body of knowledge by providing insights into the representation of SSH in open bibliographic databases and the role of open access in this domain. The study highlights the necessity for advocating OA practices within SSH and the significance of open data for bibliometric studies. It further encourages additional research into the impact of OA on various facets of citation patterns and the factors leading to disparity across disciplinary representation.

Keywords: Open Access, Social Science and Humanities, OpenCitations Meta, ERIH-PLUS, DOAJ, scholarly communication, open science, citation

1 Introduction

Social Science and Humanities (SSH) are known to be among the most difficult domains of application for bibliometric studies, a field concerned with the application of quantitative methods to measure scientific publications and their impact in the field of scholarly research. Major research in the domain includes the understanding of scientific citations and the investigation of publication pattern, as well as the use of such measurements in policy and management contexts. Bibliometrics application in the SSH seems to be hindered by their underrepresentation, as bibliometric studies rely in large part in the databases' representativeness of the scientific activity studied.

Throughout the years, the comparative study of databases' disciplines has focused on Web of Science and Scopus, two authoritative but commercial sources, demonstrating in both older and more recent studies (Sivertsen, 2014) that “the use of either WoS or Scopus for research evaluation may introduce biases that favor Natural Sciences and Engineering as well as Biomedical Research to the detriment of Social Sciences and Arts & Humanities” (Mongeon & Paul-Hus, 2016). Nonetheless,

the coverage of SSH Journals seems to be lower compared to two other disciplines even when involving other sources (Visser et al., 2021).

The problem seems to arise from the different publication patterns in SSH publications with respect to other domains. For example, “significant parts of the scholarly production in the SSH are published in national journals, book chapters, and monographs. As a result of this diversity, the challenge of setting criteria for the selection of source items is seen as much greater than for the sciences” (Sivertsen & Larsen, 2012).

Only in recent years, mainly thanks to Open Science initiatives and the shift towards Open Access publishing, new infrastructures have emerged or opened their data, contributing to the advancement of reliable and transparent scientific studies. A significant progress for SSH has been the creation of the European Reference Index for the Humanities (ERIH PLUS, (<https://kanalregister.hkdir.no/publiseringsskanaler/erihplus/>), “a deliberate attempt to go beyond the commercial indexing services such as Web of Science and Scopus by covering more comprehensively all peer-reviewed scholarly journals in the SSH that are publishing at a minimum national level”(Lavik & Sivertsen, 2017).

Another turning point for open bibliometric studies has been the Initiative for Open Citations (I4OC, <https://i4oc.org/>) which promotes the unrestricted availability of scholarly citation data as a crucial requirement for the bibliometrics and scientometrics domain. Thanks to this initiative, many publishers have decided to open their deposited references at Crossref (<https://www.crossref.org/>), an association that provides persistent identifiers assigned to academic publications and publishes metadata related to these publications. In their recently published comparative study, Borrego et al. (2023) assessed the ERIH-PLUS coverage of Crossref and identified it as the most promising resource for bibliographic discovery in the Arts and Humanities.

They also pointed out the need for improvement of its metadata completeness. One of the founders of I4OC is OpenCitations (Peroni & Shotton, 2020), a not-for-profit infrastructure organization for open scholarship dedicated to the publication of open bibliographic and citation data using Semantic Web technologies. OpenCitations aligns fully with open science guidelines and complies with FAIR principles (<https://www.go-fair.org/fair-principles/>). Currently, OpenCitations retrieves and organizes citation data from various sources, including all Crossref open references. However, as of the present day, no study on the coverage of this new resource for SSH journals has been conducted.

Our research aims to address the lack of representation of OpenCitations in the panorama of comparative studies on bibliographic indexes, specifically regarding the coverage of SSH journals. At the same time, it aims to conduct further evaluation by observing the distribution of journals across disciplines and countries, as well as assessing their Open Access status. To do so, we will consider OpenCitations Meta (<https://opencitations.net/meta>) the database that stores all bibliographic metadata for all publications involved in the OpenCitations indexes and compare it against ERIH PLUS as it is the most comprehensive SSH index at the present day, with 11.129 Journals listed.

Our research questions are the following:

- **RQ1:** What is the coverage of publications in Social Science and Humanities (SSH) journals (according to ERIH-PLUS) included in OpenCitations Meta?
- **RQ2:** What are the disciplines that have more publications?
- **RQ3:** What are countries providing the largest number of publications and journals?
- **RQ4:** How many of the SSH journals are available in Open Access according to the data in DOAJ?

The study has its foundation in the belief that open infrastructures are fundamental to ensure transparency, reproducibility, and trust in scientific research, having an essential value for the bibliometrics domain. The present study can contribute to fostering a culture of open science and calling for more incentives and adequate policies where needed.

Moreover, in agreement with Tennant et al. (2016), this study considers OA as of the challenges faced by the “open science” transformation since it has been proven to lead to significant academic and societal advancements, enabling unprecedented data access for scientific studies to non academic scientists, and to machines for data mining, for instance.

Nevertheless, open Access publishing is still damaged by questionable or inaccessible practices, such as Article Processing Charges and Predatory Journals. Severin et al. (2020) research on discipline-specific OA publishing practices found the OA uptake in SSH to be lower than in most other fields. Among the reasons behind this trend, one can find lower levels of general OA awareness among scholars, the notion of prestige related to the choice of venue publications and misunderstandings related to licensing and plagiarism. Therefore, the need for the promotion of Open Access good practices and the dissemination of reliable OA Journals is of extreme importance today.

In the context of our research, to ensure the reliability of the data about OA status, we will refer to the Directory of Open Access Journals DOAJ (<https://doaj.org/>), since it is a trusted source with 19,366 indexed journals at the present day.

The rest of the paper is structured as follows. In Section 2, we present the Materials and Method: in Section 2.1 the Data used and in Section 2.2 the Methodology followed in the research. In Section 3, we present our Results, in Section 4, we provide Discussions including limitations, in Section 5, Conclusions and possible further research are provided.

2. Materials and Methods

The adoption of open science good practices to transparently manage every research outcome, from managing data and developing a methodology to the publication of results, is a central aspect of the present research: The Data Management Plan (DMP) (Ghasempouri et al., 2023a) that was published at the beginning of our work ensured transparency about the handling of the data produced. It describes in detail the choices made to comply with FAIR data principles, making our data findable, accessible, interoperable, and reusable.

The DMP includes two datasets: 1) a research software (Ghasempouri et al., 2023b), and 2) a data catalogue with the output files of the software where our results are stored (Ghasempouri et al., 2023c; Ghasempouri et al., 2023d). The first version of the DMP has undergone peer review, and a second version has been published with necessary modifications. Finally, further improvements have been done, reflecting changes made during our work process, the DMP is currently published in its fourth version. The same practice has been adopted to develop the study’s methodology, which is a core research outcome essential for enabling reproducibility. Our methodology will be described below, but we suggest referring to the published workflow on protocols.io (Ghasempouri et al., 2023e) for a more detailed description. As mentioned, the published workflow has also been modified during the research, incorporating peer reviews and in-process changes that we deemed necessary. The current working version is the fifth.

2.1. Used Data

The research used secondary data retrieved from the following sources: the OpenCitations Meta database, the ERIH PLUS bibliographic index, and the DOAJ index. The OpenCitations Meta (OC Meta) dataset was downloaded as a dump on 24/02/2023 (OpenCitations, 2022). We opted for the dump download because accessing the data through a SPARQL endpoint or REST API would have taken too much time due to the large quantity of data we needed. The dataset represents the largest source in this research, consisting of a zipped folder (8GB) containing over 22,000 CSV files (36GB unzipped), covering more than 90 million bibliographic entities. The ERIH PLUS (EP) dataset “Approved Journals” contains metadata about all SSH Journals included in the Index and is publicly available for download in CSV format as of 27/04/2023 (2MB) (ERIH PLUS, 2023). The metadata for each Journal entry includes a unique ERIH PLUS identifier, Print-ISSN and Online ISSN,

Original and International title, Country of publication, ERIH PLUS disciplines classification, and the OECD classification. For the purpose of this study's disciplinary classification, the ERIH PLUS classification was adopted. The DOAJ dataset was also downloaded in its publicly available dump in CSV format on 28/05/2023 (22.9MB) (DOAJ, 2023). According to the guidelines (<https://doaj.org/apply/guide/>) the DOAJ includes only OA journals whose copyright holders grant usage rights to others using an open license. Additionally, the full text of all content must be available for free and open access without delay. The main metadata needed from the DOAJ dataset was the Journal ISSN and EISSN, as well as the Country of Publisher, to integrate with ERIH PLUS countries.

2.2. Methodology

To answer research questions, a purpose-tailored research software in Python programming language was created. The main library used to explore, manipulate, and combine available data into meaningful new datasets is Pandas (<https://pypi.org/project/pandas/>). Other libraries used include glob, os, tqdm, concurrent.futures, arg-parse, csv.

As a first step, a dataset named `SSH_publications_in_OCMeta_and_Open_Access_status` is created as the result of the parsing and merging of the three datasets. This dataset is intended to indicate how many SSH journals listed in EP are included in the OC Meta database, along with the number of publications (OC Meta) and Open Access (DOAJ) status for each journal. It is created as the output of the method `process_files()` called on an object of the class `PlayaristsProcessor`, initialized with the path to the three considered datasets, a batch size and the number of CPU workers to use for parallel processing. These last two variable are required given the dimensions of the unzipped OpenCitations Meta dump, which is divided in batches and processed in parallel for enabling an efficient parsing with a reasonable execution time. In detail, the `concurrent.futures` module is used to provide a high-level interface for asynchronously executing callable. In this step, a `ProcessPoolExecutor` object is created and a function that performs the filtering of OpenCitations Meta dump according to the journals present in EP is mapped to the object. The ancillary function that performs this step is `process meta csv()`. First, it creates a dictionary out of the EP data with unique Print ISSN and Online ISSN as keys and EP Journal ID as value, then it retrieves the issn identifiers in OC Meta venue column with a string matching technique and it proceeds to verify whether these identifiers are present in the key values of the EP dictionary. The positive matches are stored in a new DataFrame containing the unique OC omid identifier, a list of the issn of each journal, the unique EP identifier and the number of Publications present in OC Meta for each journal. Additionally, it's noteworthy that the `process_files()` function also addresses potential discrepancies between different datasets, particularly between OC Meta and ERIH PLUS. It identifies and records in a file named `duplicate omids.csv` those OC Meta identifiers that seem to correspond to distinct journals but are identified by ERIH PLUS as pertaining to the same journal. Finally, the Open Access status is added by means of the `process doaj file()` ancillary function. This function takes in input the DOAJ dataset as a DataFrame and the newly created DataFrame with the SSH journals in OC Meta; it verifies whether the identifiers in the issn column of this last DataFrame are present in the DOAJ values of Journal ISSN and Journal EISSN columns and extend the result DataFrame by adding the Open Access column. This column's value is set as "True" if a positive match is found and as "Unknown" if not; since there might be Open Access Journals not enlisted in DOAJ. This first DataFrame provides the data needed to answer questions 1 and 4 of this research about the OC Meta coverage of EP journals and their OA status and is exported in csv format with the name `SSH_Publications_in_OCMeta_and_Open_Access_status.csv`.

As a second step, to answer research questions 2 and 3, two other datasets are created: `SSH_Publications_by_Discipline` and `SSH_Publications_and_Journals_by_Country`. Both datasets have three columns, the first, respectively, for disciplines and country classification, the second holding journal counts value and the third holding publication counts value. The count of journals per discipline is not comprehended in our research question, nonetheless, since the process was similar to the one adopted to retrieve the count of journal per country, it was considered worth adding as an

additional result. It is worth mentioning that EP journals specify more than one discipline. Given the lack of access to more granular data about single publications disciplinary classification, our approach was to consider all publications under the same journal as belonging to all disciplines specified for that journal. To retrieve information about disciplines a dictionary is created by the method `create_disciplines_dict()` of the class `DisciplinesProcessor`. This method merges `SSH_Publications_in_OCMeta_and_Open_Access_status` with EP dataset to include disciplines information, then it creates the dictionary by setting each unique discipline in the EP dataset as key and a list of EP unique identifiers of the journals belonging to each discipline as value. The process unfolds approximately in the same way to retrieve information about countries, but in this case it's the `create_countries_dict()` of the class `CountriesProcessor` to be called. This method differs from the one specified above because it needs to parse the DOAJ dataset to retrieve Country of publisher values whenever countries information are missing in ERIH PLUS. To do so, it relies on the ancillary function `retrieve_DOAJ_countries()` that adds missing countries to the dictionary. This function filters `SSH_Publications_in_OCMeta_and_Open_Access_status` to create a sub `DataFrame` with only rows of identifiers that lack country information, then merges it twice with DOAJ, once for Journal ISSN column and once for Journal EISSN, using the "how=left" parameter. Then, it adds the countries in DOAJ to the missing values of the Country column. While performing this step, the function also saves a list of the issn that still remained without country information, for transparency and data completeness. Lastly, it iterates over the updated `DataFrame` and either adds new journals identifiers to the respective country key in the dictionary, if it already exists, or creates a new country key and related journal identifiers as value. Finally, to compute the counts of publications and journals for both disciplines and countries, the method `counts()` of the `CountsProcessor` class is run. This method takes in input either the country or the discipline dictionary created in the previous step and the label to set as the first column value in the final dataset. It iterates over the dictionary keys to filter `SSH_Publications_in_OCMeta_and_Open_Access_status` according to journals in the list specified as value, then it stores the length of the filtered `DataFrame` as the count of the journals. Lastly, it sums all the values in the column `Publications_in_venue` to calculate the count of publications. The final `DataFrame` is exported as csv file according to the export path defined as `CountsProcessor` attribute.

As will be elaborated shortly, a significant disparity is evident between the publication count of the top four largest Mega Journals (MJ) and all other journals. Consequently, the most recent software update has incorporated a feature that enables the exclusion of these primary four journals from the analysis. This enhancement facilitates more comprehensive comparisons. To achieve this, a new parameter named `remove_megajournals` has been added, which takes boolean values and can be specified by the user while launching the program. If set as `True`, when starting to process disciplines and countries, it will remove the first four rows of `SSH_Publications_in_OC_Meta_and_Open_Access_status`. The newly introduced argument is used in the instantiation of the following classes: `DisciplinesProcessor`, `CountriesProcessor`, and `CountsProcessor`.

Moreover, in the second version of the software, a more detailed comparison of US Journals and European Journals has been carried out, in order to better understand how disciplines are distributed in the two countries with the highest number of publications, as well as exploring the correlation between the number of publications and the number of disciplines per journal. The class `Compare_US_EU`, that inherits from `ResultsProcessor`, was created to perform such task, with two internal methods: "`compare_us_eu`" and "`counts_us_eu`".

The first method, takes in input the ERIH dataset and the Countries dictionary, filters the dictionary to retrieve the values (the ERIH identifiers) for European countries (including cases where two European countries are specified as the country of publication of the same journal, while excluding Turkey and Russia) and "United States and stores them in two separate lists called "`eu_ID`" and "`us_ID`". Then, it filters the ERIH dataset, and creates two new temporary `DataFrames`, one only with US Journals and one with European Journals. To do so, the `.isin()` method is used to check whether an ERIH identifier in the dataset is present in the two lists previously created of US and

European identifiers. Subsequently, for both the created DataFrames, the number of disciplines per Journal mentioned in the column "ERIH PLUS Disciplines" are counted and the count is added as a new column "disc_count" value. After this step the two Dataframes are merged with SSH_Publications_in_OCMeta_and_Open_Access_status through the common ERIH identifier to add data about the number of Publications in venue; the intersection between the two is obtained using intersection of keys from both frames and the result DataFrames are named "us_meta" and "eu_meta". On these two Dataframes the following operations are performed: 1) they are filtered to create two datasets, one for US and one for Europe, containing only the columns "EP id", "Publications_in_venue", "Original_Title", "Country_of_Publication", "ERIH_PLUS_Disciplines", "disc_count" and exported as csv file with the name of "us data" and "eu data". 2) they are filtered to create two other datasets, again, one for US and one for Europe, containing only the columns that were in SSH_Publications_in_OCMeta_and_Open_Access_status and they are exported in csv with the name "meta_coverage_us" and "meta_coverage_eu". The creation of the datasets mentioned in 1) is useful to record meaningful information that are excluded from the datasets mentioned in 2). The latter, nonetheless are necessary to have a data structure fit for the computation of the disciplines count described below.

The second method, "counts_us_eu" is similar to counts() but takes either "meta_coverage_us" or "meta_coverage_eu" as parameter too and performs the count on them. The results are also exported as csv.

As far as software testing is concerned, given the complexity and size of the datasets, synthetic data have been created to test the software functioning. Creating diverse data occurrences and possible errors or odd instances to evaluate how the software would react in such situations. Here are the designed requirements: both OCMeta and DOAJ contain publications or articles not included in ERIH PLUS; OCMeta dumps are divided into multiple files: the program must correctly concatenate data from different files; OCMeta files contain different types of entries, not just publications (for example, venues). Also, multiple publications are associated with the same venue. OCMeta publications have a variable number of venue identifiers, which are not associated with a precise number of columns. We designed three test cases: 1) to test the merging of OC Meta and ERIH-PLUS, 2) to test the merging of the SSH_Publications_in_OCMeta_and_Open_Access_status, resulting from the above step with DOAJ dataset, 3) to test the correct assignment of countries and disciplines to different journals. Considering the construction of our software pipeline we can assume that if the software's success in these tests can be a solid proof of its functioning to fulfil its requirements, minor errors could still be present in the handling of data but they shouldn't be compromising for the overall quality of our research. As a final step in the workflow, suitable visualization to make the results' analysis more accessible were produced. These were generated using matplotlib and seaborn.

3. Results

Here we present the most significant results of the research, along with some other supporting data exploration. As mentioned before, the initial results were compiled including Mega Journals (much larger than traditional Journals), but we realized they might have been skewed by the fact that, according to our approach, each publication of a Journal counts as 1 for the discipline(s) of that Journal. In particular, there are four Mega Journals whose number of publications, reasonably covering a vast range of disciplines, is unproportionally higher compared to other Journals.

These are relevant data for the MJ present in our dataset:

ERIH id	Title	Number of Publications	Disciplines	Country
470652	PLOS ONE	268.538	Anthropology Archaeology Demography Environmental Studies History & Philosophy of Science Human Geography and Urban Studies Interdisciplinary research in the Social Sciences Psychology Sociology	US
422484	Nature	207.581	Anthropology Archaeology Demography Economics Environmental Studies Human Geography and Urban Studies Interdisciplinary research in the Social Sciences Media Studies and Communication Science and Technology Studies Social Statistics and Informatics Sociology	UK
341568	PNAS	148.780	Anthropology	US
341841	Science	145.090	Psychology	US

Table 1 Data about Mega Journals present in our dataset.

In the following section, we will present both the results compiled with MJ and without MJ, for the research questions that included publications' count ("What are the disciplines that have more publications" and "What are countries providing the largest number of publications and journals"). By comparing the results, the ranks produced without MJ seem to present a more balanced count, which is likely more representative of the real situation in SSH publications practices.

Nonetheless, the research lack the granular data needed to fairly assess how many SSH publications MJ publish, and what is the ration with publications of other areas of the same Journals. For this reason both of the results are presented to provide grounds for the reader to draw their own interpretation of the data.

3.1. What is the coverage of publications in Social Science and Humanities (SSH) journals (according to ERIH-PLUS) included in OpenCitations Meta?

OpenCitations Meta cover the **78.1%** of ERIH PLUS index, for a total of **8.691 journals**, and **5.496.449 publications** (Figure 1). Comprehensively, the SSH Journals in OpenCitations Meta represent the **7%** of the whole database.

This result shows that, on the one hand, OC Meta covers a significant proportion of ERIH PLUS journals. According to the most recent study, this percentage is almost comparable to Crossrefs' coverage (Borrego et al., 2023) **80%** and significantly higher than Scopus' coverage **49%**, affirming OpenCitations' promising position as a bibliographic database for SSH. Nevertheless, it is worth noting that a higher coverage of journals does not necessarily imply a larger coverage of publications, since the quantity of articles published by a journal may differ significantly, as well as the completeness on the coverage of publications for each journal may still vary among different databases. As it has been found in our research, the counts of publications per journal in OpenCitations Meta SSH journals ranges from 1 to 257538 (PLOS ONE Online Journal, issn:1932-6203.).

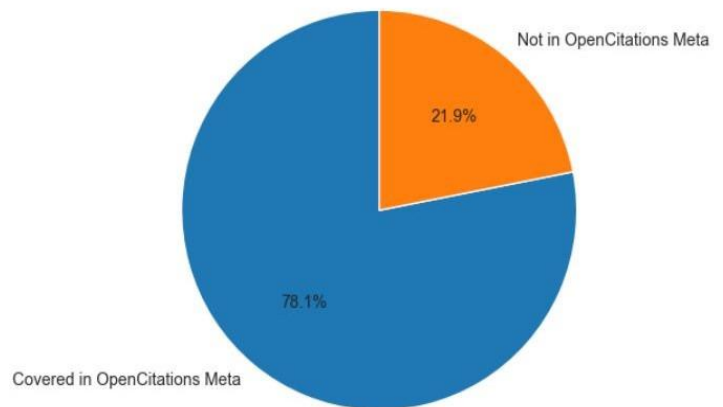


Figure 1 Pie Chart. Percentage of ERIH PLUS Journals covered by OC Meta

3.2. What are the disciplines that have more publications?

Disciplines have been ranked according to their publications number in descending order (Figure 2). A wide gap between Psychology, the discipline with the highest number of publications and the other disciplines can be observed. Although ERIH PLUS doesn't specify which disciplines belong to the Social Sciences and which belong to the Humanities, referring to a reasoned classification provided by Spinaci et al. (2022) it is possible to identify a trend that shows Social Science as covering more publications than Humanities. In fact, among the top 10 disciplines, only Archaeology belongs to the Humanities domain, whereas all the fields placing lower in the rank belong to the Humanities.

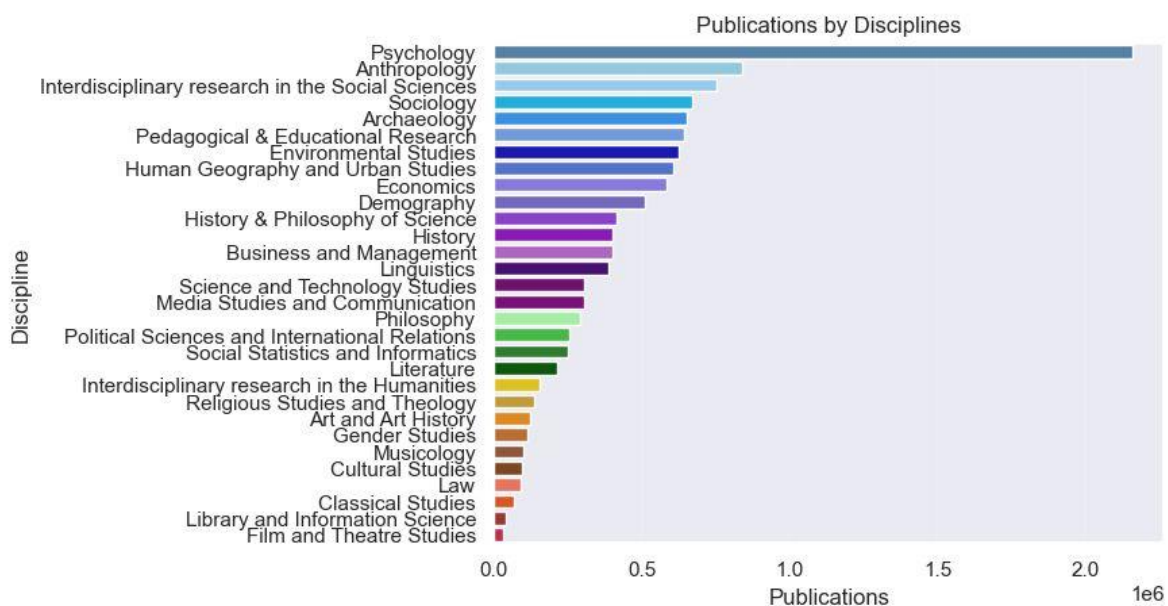


Figure 2 Bar chart. Disciplines ranked according to publications' number.

After running the software without these Mega Journals we could observe a significant change in the ranking of the disciplines with more publications (Figure 3). Most noticeably, a group of disciplines that belonged to these Journals and were all ranking among the disciplines with more publications have migrated to lower positions. In particular we should mention Anthropology (from 2nd to 11th), Interdisciplinary Research in the Social Sciences (from 3rd to 8th), Sociology (from 4th to 12th), Archaeology (from 5th to 13th), Environmental Studies (from 7th to 15th), Human Geography (from 8th to 18th and Demography (from 10th to 29th). On the other hand, we can see some other disciplines not belonging to these MJ, rise higher in the rank. In particular Pedagogical and Educational Research (from 6th to 2nd), History (from 12th to 3rd), Business and Management (from 13th to 4th), Linguistics (from 14th to 5th), Economics (from 9th to 6th), Philosophy (from 17th to 7th), Political Sciences and International Relations (from 18th to 9th) and Literature (from 20th to 10th).

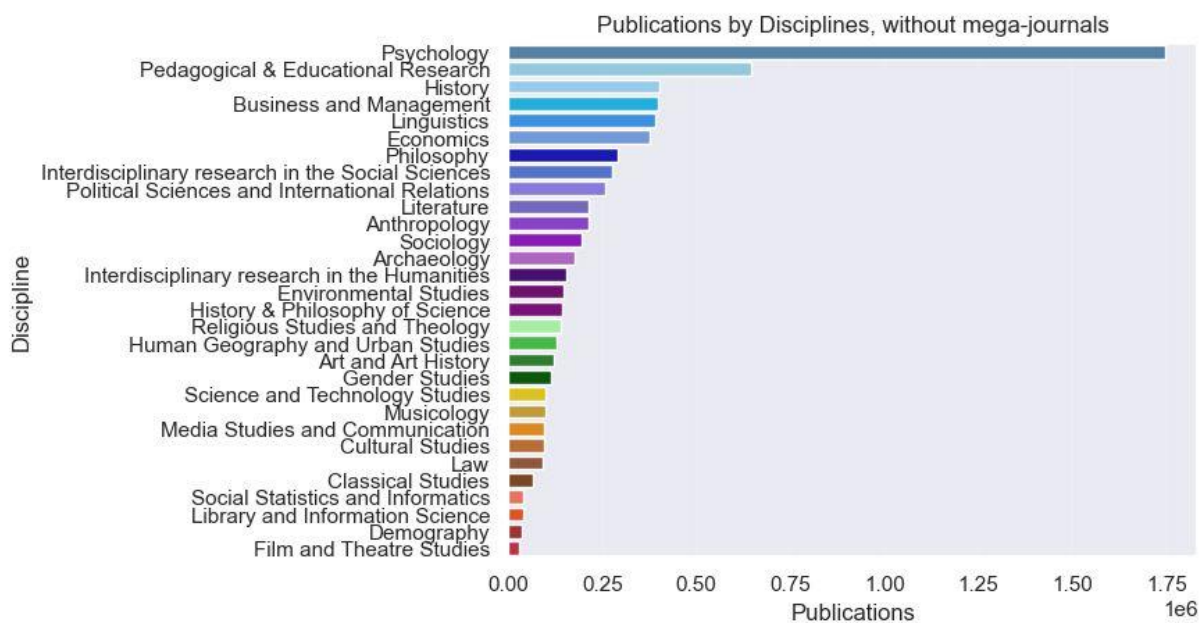


Figure 3 Bar chart. Disciplines ranked according to Publications number, excluding Mega Journals.

Although Psychology still appears to be the most published discipline, the exclusion of MJ from the rank shows a much more balanced and diversified panorama of most published disciplines, with more Humanities disciplines in the higher ranks (History, Philosophy, Linguistics, Literature, Archaeology). To provide some grounds for understanding the exceptional position of Psychology amongst other SSH disciplines, some factors may be of interest. Firstly. Psychology is a field of studies that covers a wide scope of topics, from areas like Clinical Psychology and Developmental Psychology to research in Neuropsychology, Emotion studies and so on. These disciplines are known to employ rigorous experimental methodologies, including controlled experiments, surveys, observational studies, and meta-analyses. This scientific approach contributes to the credibility and reliability of the research conducted in the field. Moreover, the findings from psychological research can have practical implications for improving individual and societal well-being. As a result, not only governments and institutions might be willing to allocate more research fundings in this area, but also publishers might be more motivated to publish psychological research to share valuable findings that can produce a more direct societal impact.

3.3. What are countries providing the largest number of publications and journals?

3.3.1. Publications

The results showed the coverage of **92 countries** in for SSH publications and journals in OpenCitations Meta. Due to the high number of publications the x axis of the bar graph is written in scientific notation ($0.1 = 100.000$, $1 = 1.000.000$). In this case, the difference in the ranks with or without MJ, doesn't affect the results, in fact, all four Mega Journals belong either to the United States or to the United Kingdom. Respectively, these countries hold **2.245.731** and **1.605.994** publications including MJ; **1.683.323** and **1.398.413** publications excluding MJ. In other words, the United States loses 562.408 publications that were coming from its three MJ, against the 207.581 publications loss of the United Kingdom, and as a result, the gap between the two countries shortens significantly.

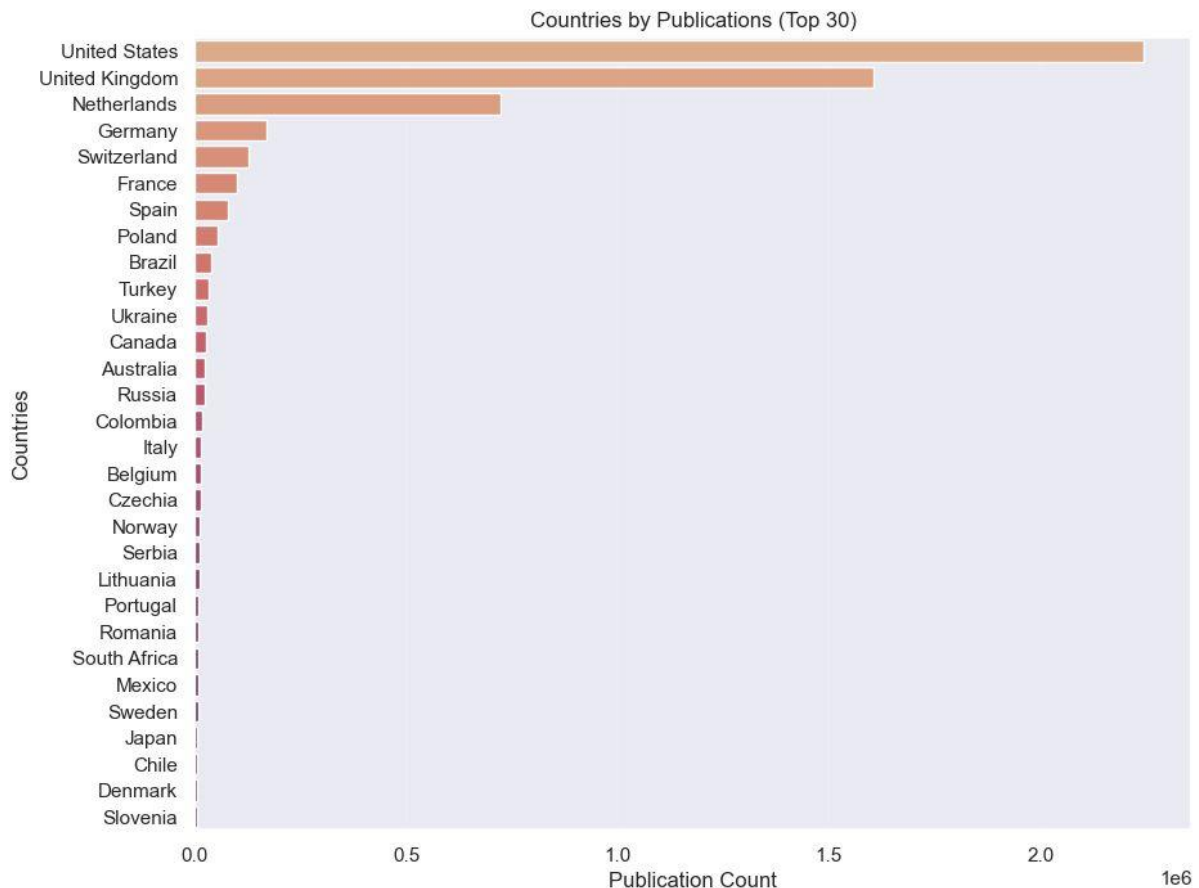


Figure 4 Bar chart. Countries with the highest number of Publications, with Mega Journals.

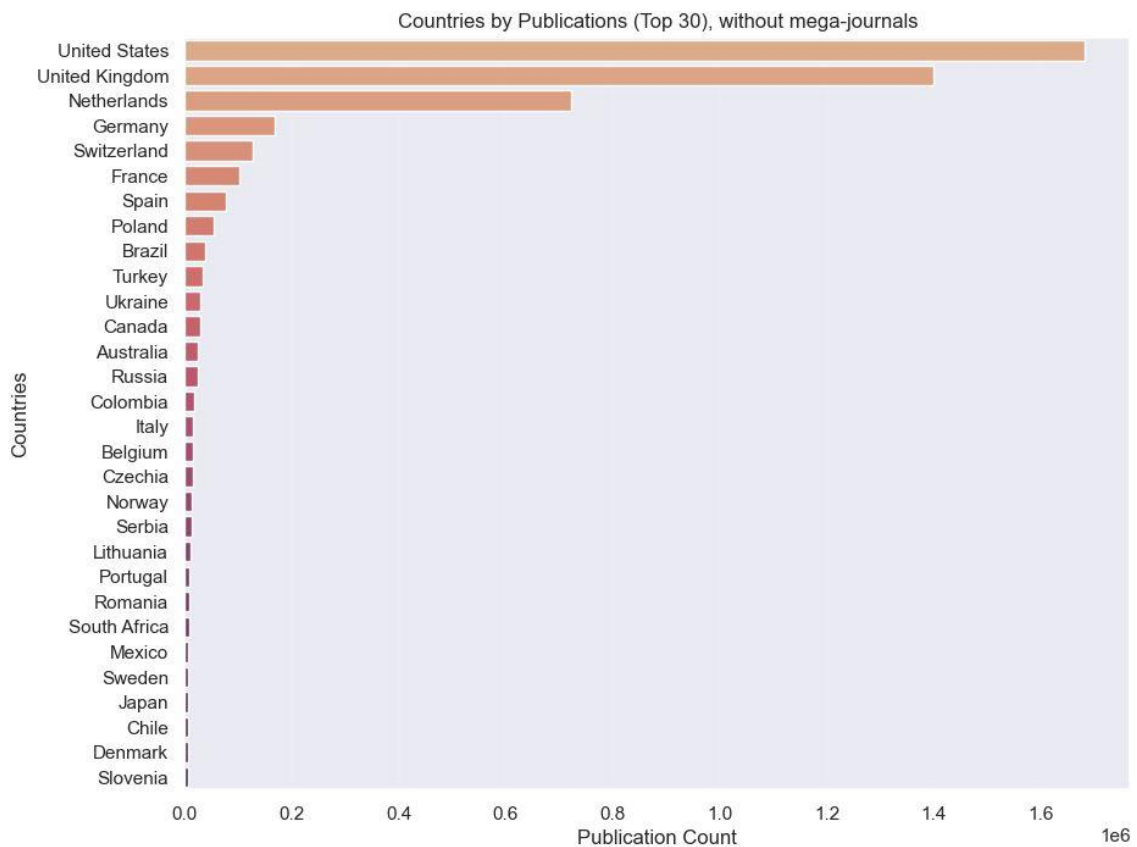


Figure 5 Bar chart. Countries with the highest number of Publications, excluding Mega Journals.

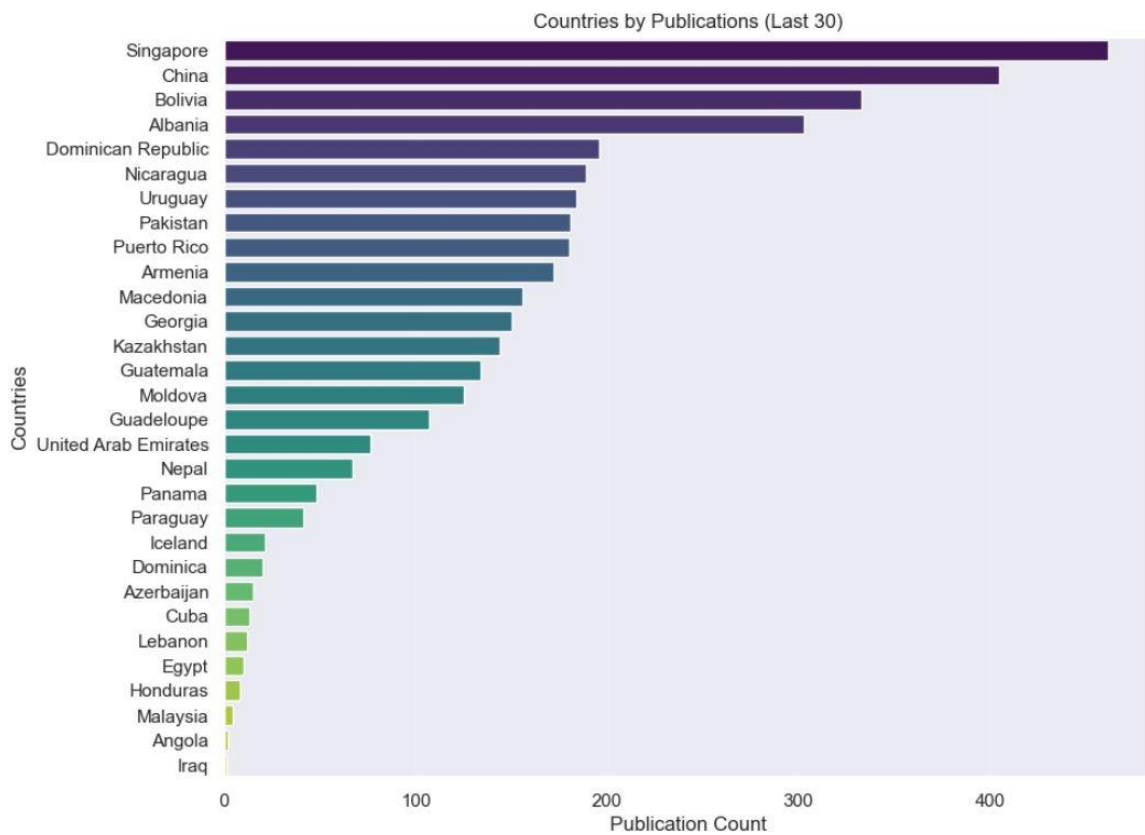


Figure 6 Bar char. Countries with the lowest number of Publications.

As the bar chart shows, beside the United States, all other countries with the highest rankings are European countries. This is not surprising, given that, as ERIH PLUS states: “the main target group of the index are researchers and research within a European framework. To the extend which the index holds journals from other parts of the world, it is because they are assumed to add value to the ERIH PLUS main target group and scope” (ERIH PLUS webpage, About, 2021). Nevertheless, if we consider the geographical distribution of the data, the value for a small European country can’t be possibly compared with the value for wide territorial extensions such as the USA or South Africa. For this reason, while doing the same research on Scopus and Crossref, Borrego et al. (2023) have aggregated European countries into East, North, South and West Europe, highlighting the predominance of European countries’ publication in the database. Accordingly, in the present study, when aggregating the data for European countries, it has been found that Europe has a total number of **3.024.993 (55%)** publications against the **2.245.731 (40%)** of United States. The predominance of European publications is even more stressed when excluding MJ from the count: **2.817.412 (59%)** against **1.683.323 (35%)**. Given these considerations, to provide a more faceted understanding of the distribution of SSH disciplines in Europe, a comparison between Europe and United States, both without and with MJ (Figure 7 and 8).

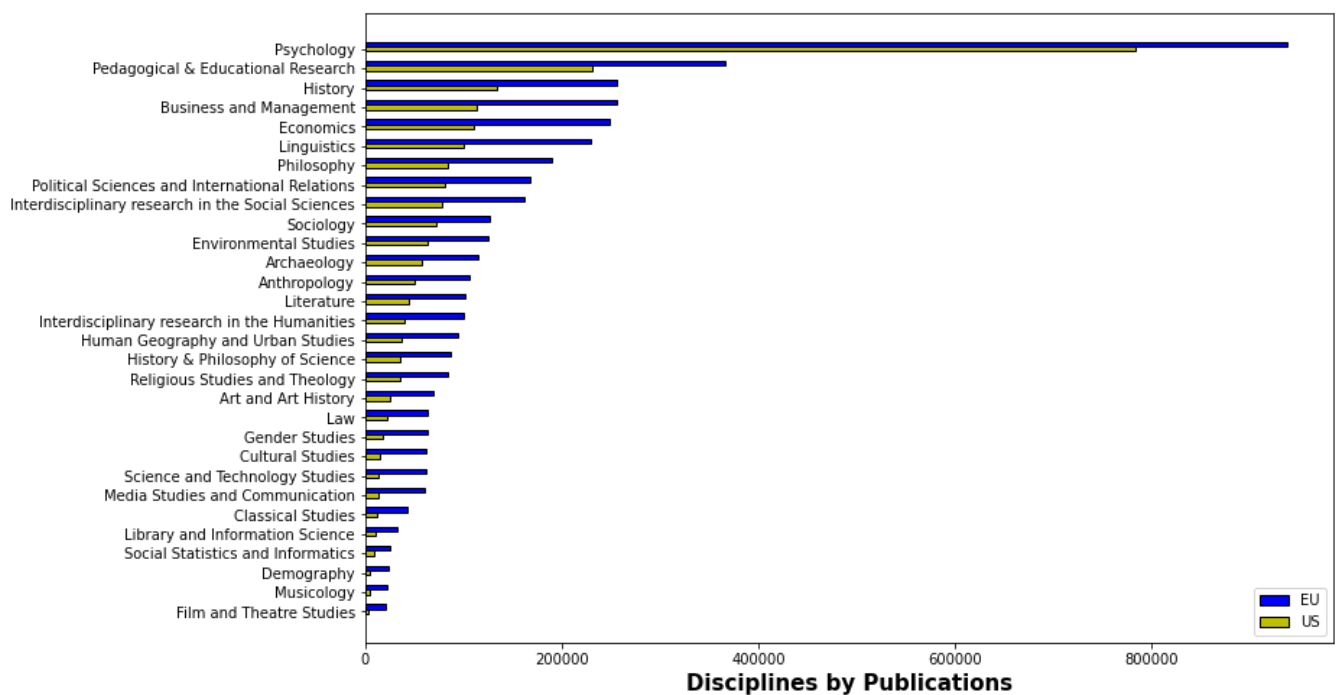


Figure 7 Bar chart. Comparison between US and EU’s disciplinary distribution, without Mega Journals.

As the bar chart shows, there doesn’t seem to be a significant change in the disciplines’ distribution between US and EU, when excluding MJ, rank of European countries mimics more or less the general one. On the contrary, by keeping MJ in the comparison, not only it is possible to visually assess how much the United States influence the rank of specific disciplines (especially Psychology and Economics) but we can also observe a much less linear decrease in publication counts in the US publication, in so far as the disciplines belonging to MJ are all visibly boosted by United States’ publications (Interdisciplinary research in the Social Sciences, Sociology, Environmental Studies, Archaeology, Anthropology, Human Geography and Urban Studies).

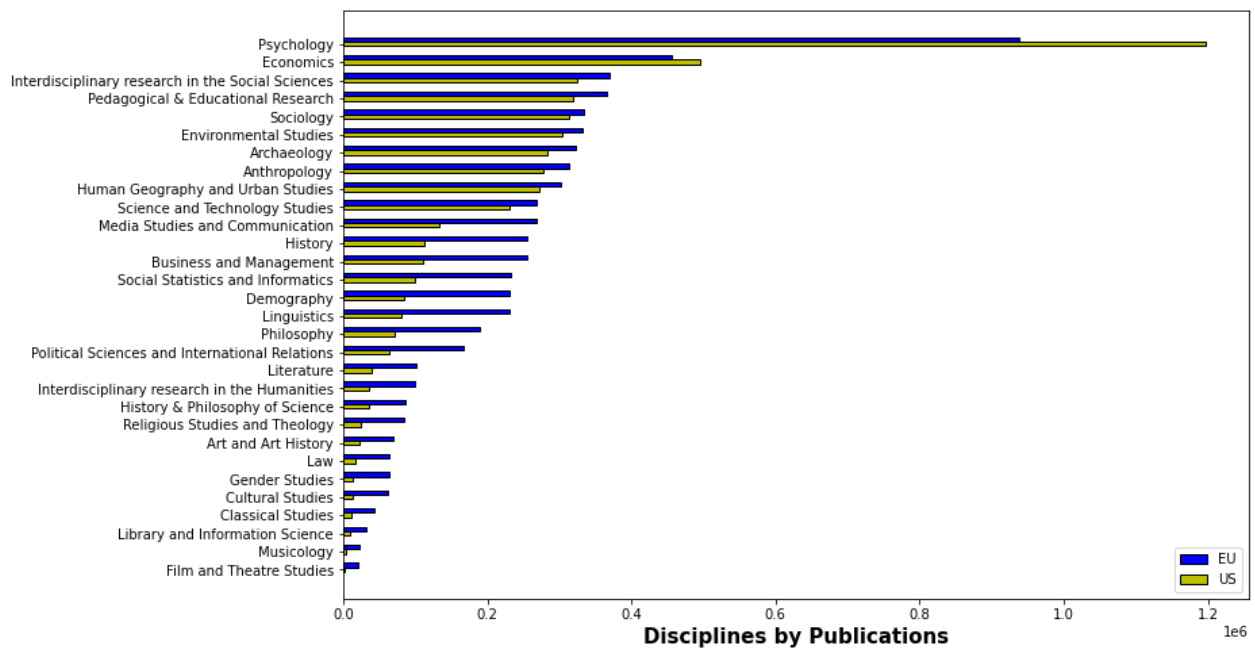


Figure 8 Bar chart. Comparison between US and EU's disciplinary distribution, with Mega Journals.

3.3.2. Journals

Highlighting the distribution of journals over countries is especially relevant in the present study, considering the factor mentioned above of the variation of publications per journal. In Figure 9 and 10, it is particularly remarkable to observe the different position held by Australia and Switzerland. The first is the 13th country for number of publications with 24.225 publications but doesn't even place in the top 30 if considering Journals, as it publishes only 38 Journals; the second is the 5th for number of publications with 127.230 articles, but is at the 30th position considering journals, having only 49. A similar comparison can be profoundly interesting for evaluating different countries publication practices. As we have seen, scientific research in the SSH is carried out on a national level more than in other domain, for this reason these data could be very useful for further comparative research on national publication practices in the SSH.

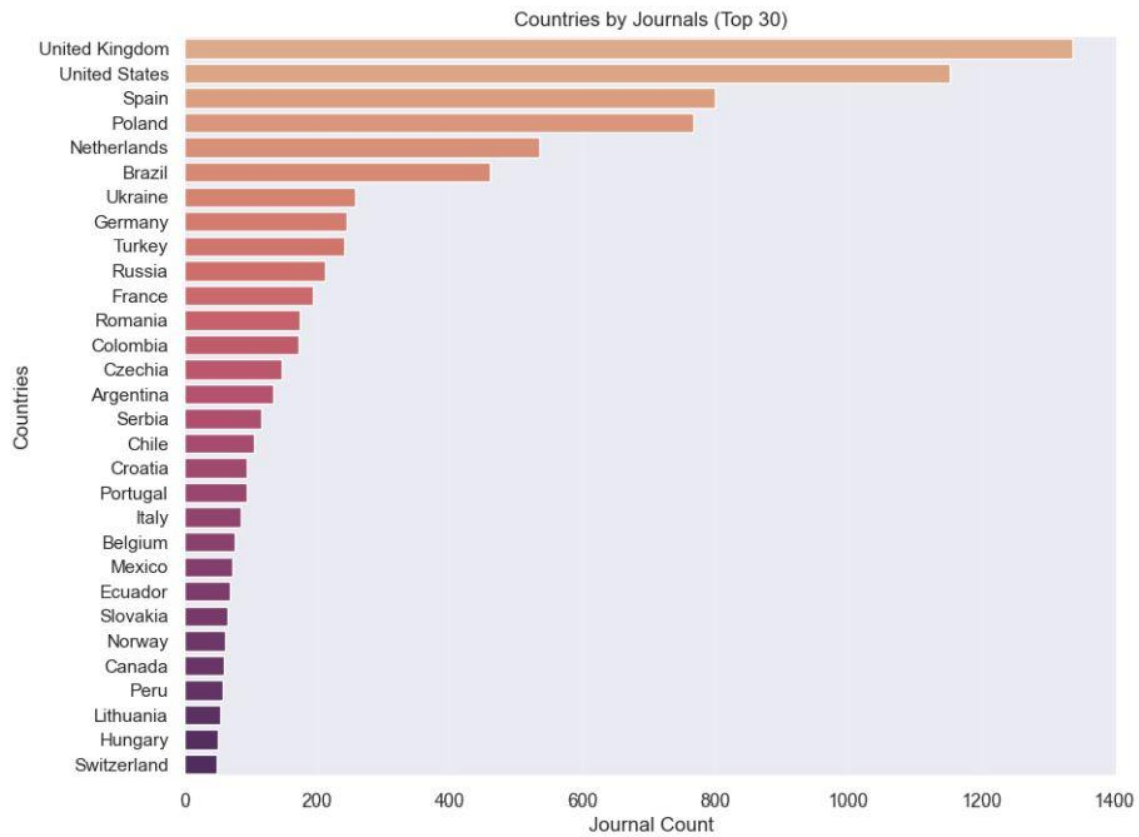


Figure 9 Bar chart. Distribution of journals over countries. Top 30.

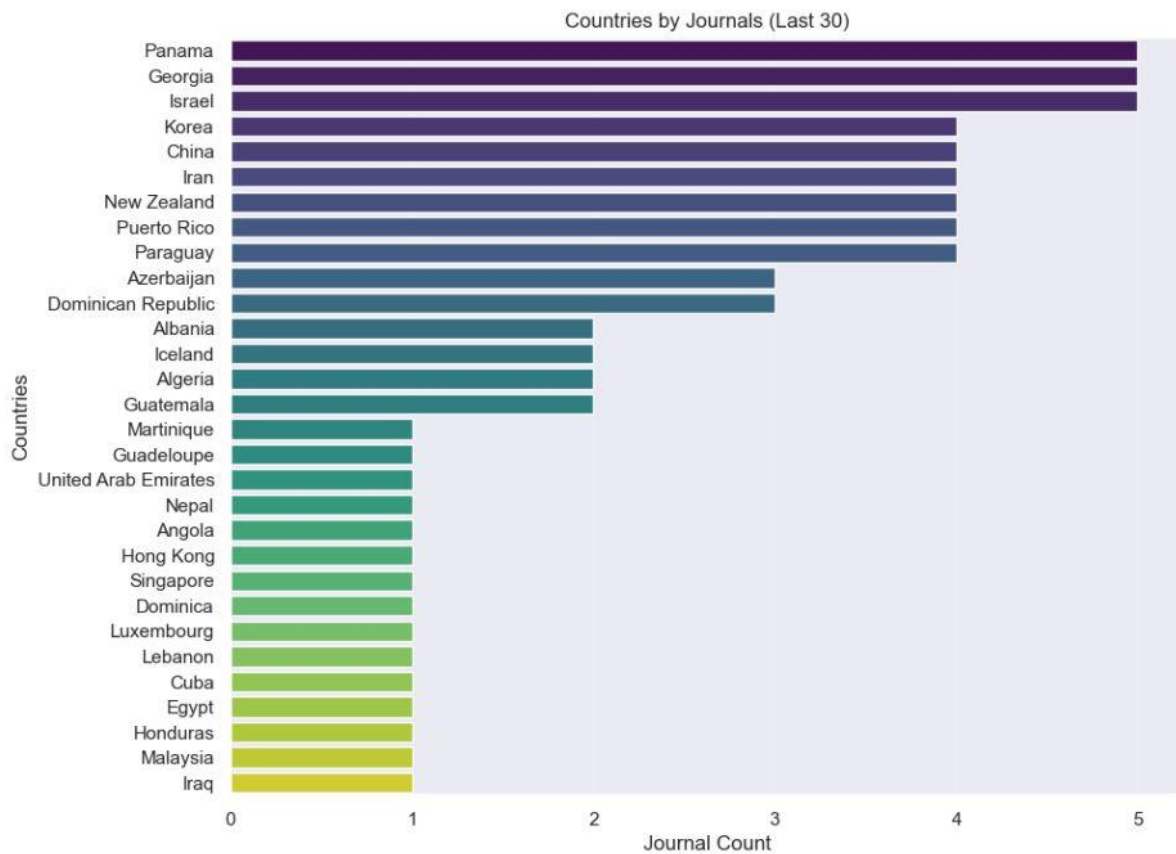


Figure 10 Bar chart. Distribution of journals over countries.

3.4. How many of the SSH journals are available in Open Access according to the data in DOAJ?

Only **38%** of the SSH Journals covered by OpenCitations Meta, according to DOAJ criteria, are OA, for a total of **3.299 Journals** and **743.559 Publications**. This percentage aligns with many of the studies compared by Severin et al. (2020) that found SSH among the slowest domains in the uptake of OA practices.

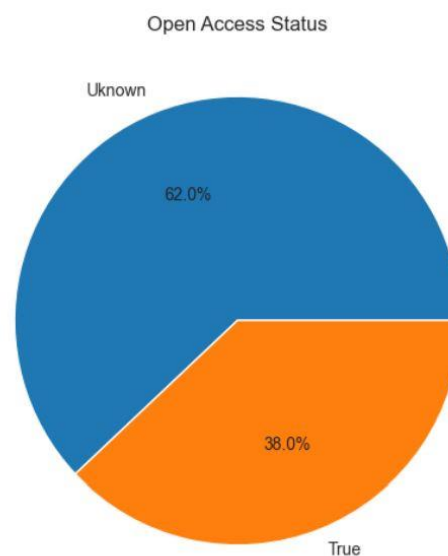


Figure 11 Pie chart. Percentage of SSH Open Access Journals in OC Meta.

4. Discussion

4.1. Limitations

The main limitations of the study are the followings:

- The data used to identify SSH journals are biased towards Europe, making it difficult to generalize the results global scale. This is mainly due to the Eurocentric approach of the ERIH PLUS index considered for this study, to address this, it would be advisable and interesting to reproduce the same study with more balanced indexes, even though at the present day, qualitative, public resources of this kind seem to be lacking. Furthermore, there was a limited presence of OA journals in the field.
- The way publications are connected to disciplines: in ERIH dataset, one or more disciplines are assigned to a journal, not having data about single publications coming from ERIH PLUS, we had to retrieve the number of publications per journal from OpenCitations Meta, and all the journal's disciplines have been assigned to each publication (ex. if a journal with 50 publications was listed among Literature and Linguistics, in our results, both Literature and Linguistics would have 50 publications). This also means that, if a journal, along with SSH, is also publishing non SSH articles, those articles – being included in OC Meta – would add to the count of SSH disciplines. These limitations in our sources might skew the results, as proved by the analysis on Mega Journals. It is worth mentioning that ERIH provides a system, facilitated by Dimensions, to search and download publications' data in batches of 500 publications at time and that single disciplines are, in fact, assigned to every publication in the index. Nonetheless, accessing these information in an easy way, both for structure of data and time consumed, requires an effort that goes beyond the purpose of the present research. Another possibility is to access granular data about disciplines by incorporating resources coming from proprietary database like Dimensions, that also contain research area metadata.

5. Conclusions

To summarize the present research results, a satisfactory coverage of SSH journals in OpenCitations has been observed, with Psychology emerging as the discipline with the highest number of publications. However, the study also uncovers that a considerable percentage of the Journals in the OpenCitations Meta database are not Open Access. Although there is space for improvement, the resource proves to be a good base for bibliometric studies on SSH a at European level, especially in comparison with other bibliographic indexes and considering the consistent and comprehensive provision of metadata and citations data that OpenCitations holds.

To further enhance the usefulness of this study and provide a comprehensive overview of the issue it would be relevant to deepen the knowledge of OA availability and type across different SSH disciplines and countries. Considering that DOAJ provides many interesting information related to OA (when the journal started being OA, if there are Article Processing Charges (APC) etc.), future research could delve deeper into OA trends and the impact of APC, as well as the the general OA impact on various aspects of citation patterns. This could involve studying whether OA articles receive more citations than those behind paywalls, analysing the citation lifecycle of OA versus non-OA articles to determine if OA prolongs the paper's relevancy and citation rate, and examining citation networks to understand if OA journals tend to cluster together or bridge gaps between different research communities. Such inquiries could provide valuable insights into the broader implications of the OA movement on the visibility, influence, and interconnectedness of SSH research, resulting extremely useful to convince different actors involved in scholarly publication in the SSH of the validity of OA publishing. A similar thorough analysis could significantly help to drive political decisions on resources allocation to efficiently support the shift towards accessible scholarly research.

Moreover, the results on disciplinary distribution could be enriched and deepened, for instance by examining what is the distribution of disciplines in each country. It could be extremely useful to reproduce this research having more precise information about disciplinary classification, connecting a discipline to a single publication and not to a journal. The results also provide grounds for further investigation into the factors contributing into disparity across disciplinary representation to help in understanding and addressing them, as well as long-term evaluation of the changes in the SSH domain. Finally, it would be interesting to compare with Colavizza et al. (2022) to look at the coverage of Digital Humanities in OpenCitations Meta, mixing the data from the two studies.

Overall, the highlighted strengths and limitation of the present study prove the importance as well as the need of having accessible open data for bibliometric studies, in order to be able to perform trustworthy, balanced and replicable studies on scholarly research. So far, the open science movement was able to advance the adoption of good practices for transparent and replicable scientific research, nonetheless the space for improvements remains vast and it is of extreme relevance today to call for more awareness to be able to promote a more open research culture.

References

- Borrego, Á., Ardanuy, J., & Arguimbau, L. (2023). *Crossref as a bibliographic discovery tool in the arts and humanities*. *Quantitative Science Studies*, 4(1), 91–104. https://doi.org/10.1162/qss_a_00240
- DOAJ. (2023). *The DOAJ dataset downloaded 28-05-2023 journalcsv__doaj_20230528_0035_utf8* [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.8163248>
- ERIH PLUS. (2023). *ERIH PLUS (EP) dataset "Approved Journals" 27-04-2023*. <https://doi.org/10.5281/zenodo.8163337>
- Ghasempouri, S., Ghiotto, M., Giacomini, S. (2023a). *Open Science for Social Science and Humanities: Open Access availability and distribution across disciplines and Countries in OpenCitations Meta - DATA MANAGEMENT PLAN* (Version 4). Zenodo. <https://doi.org/10.5281/zenodo.8174644>
- Ghasempouri, S., Ghiotto, M., Giacomini, S. (2023b). *open-sci/2022-2023-playarists-code: Open Science for Social Sciences and Humanities: Open Access availability and distribution across disciplines and Countries in OpenCitations Meta - RESEARCH SOFTWARE* (v1.1.0). Zenodo. <https://doi.org/10.5281/zenodo.8253819>
- Ghasempouri, S., Ghiotto, M., Giacomini, S. (2023c). *Open Science for Social Sciences and Humanities: Open Access availability and distribution across disciplines and Countries in OpenCitations Meta - RESULTS DATASET* (with Mega Journals) (Version 0) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.8250858>
- Ghasempouri, S., Ghiotto, M., Giacomini, S. (2023d). *Open Science for Social Sciences and Humanities: Open Access availability and distribution across disciplines and Countries in OpenCitations Meta - RESULTS DATASET* (without Mega Journals) (Version 1) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.8249907>
- Ghasempouri, S., Ghiotto, M., Giacomini, S. (2023e). *Open Science for Social Sciences and Humanities: Open Access availability and distribution across disciplines and Countries in OpenCitations Meta - PROTOCOL*. V.5. <https://dx.doi.org/10.17504/protocols.io.5jyl8jo1rg2w/v5>
- Lavik, G. A. V., Sivertsen, G. (2017). *Erih Plus – Making the Ssh Visible, Searchable and Available*. *Procedia Computer Science*, 106, 61–65. <https://doi.org/10.1016/j.procs.2017.03.035>
- Mongeon, P., & Paul-Hus, A. (2016). *The journal coverage of Web of Science and Scopus: A comparative analysis*. *Scientometrics*, 106(1), 213–228. <https://doi.org/10.1007/s11192-015-1765-5>
- OpenCitations (2022). *OpenCitations Meta CSV dataset of all bibliographic metadata*. figshare. Dataset. <https://doi.org/10.6084/m9.figshare.21747461.v3>
- Peroni, S., & Shotton, D. (2020). *OpenCitations, an infrastructure organization for open scholarship*. *Quantitative Science Studies*, 1(1), 428–444. https://doi.org/10.1162/qss_a_00023
- Severin, A., Egger, M., Eve, M. P., & Hürlimann, D. (2020). *Discipline-specific open access publishing practices and barriers to change: An evidence-based review* (7:1925). F1000Research. <https://doi.org/10.12688/f1000research.17328.2>

Sivertsen, G., & Larsen, B. (2012). *Comprehensive bibliographic coverage of the social sciences and humanities in a citation index: An empirical analysis of the potential*. *Scientometrics*, 91. <https://doi.org/10.1007/s11192-011-0615-3>

Sivertsen, G. (2014). *Scholarly publication patterns in the social sciences and humanities and their coverage in Scopus and Web of Science*.

Spinaci, G., Colavizza, G., & Peroni, S. (2022). *A map of Digital Humanities research across bibliographic data sources*. *Digital Scholarship in the Humanities*, 37(4), 1254–1268. <https://doi.org/10.1093/llc/fqac016>

Tennant, J. P., Waldner, F., Jacques, D. C., Masuzzo, P., Collister, L. B., & Hartgerink, C. H. J. (2016). *The academic, economic and societal impacts of Open Access: An evidence-based review* (5:632). F1000Research. <https://doi.org/10.12688/f1000research.8460.3>

Visser, M., van Eck, N. J., & Waltman, L. (2021). *Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic*. *Quantitative Science Studies*, 2(1), 20–41. https://doi.org/10.1162/qss_a_00112