Version 2
Apr 28, 2021

Investigating Invalid DOIs in COCI V.2

Nooshin Shahidzadeh[1], Alessia Cioffi[1], Arianna Moretti[1], Sara Coppini[1]
[1]University of Bologna

In Development dx.doi.org/10.17504/protocols.io.buhjnt4n

Open Science 2020/2021

Sara Coppini

ABSTRACT

A preliminary note

This protocol illustrates the workflow adopted within a scholarly research that operates within the OpenCitations environment, which is an independent infrastructure organization for open scholarship dedicated to the publication of open bibliographic and citation data by the use of Semantic Web (Linked Data) technologies. COCI is the OpenCitations Index of Crossref open DOI-to-DOI citations.

Purpose

The purpose of this research is to find the publishers responsible for the missing citations in COCI by sending incorrect metadata to Crossref, the publishers to whom such invalid citations point to and the number of previously invalid citations which are currently valid. The ultimate aim would be of contributing to the resolution of this type of problem in order to insert the citations now valid in COCI, and correct those still invalid always in order to increase the number of open citations available and indexed in the OpenCitations project.

Study design/methodology

In the beginning, we use an already generated CSV file, containing the valid citing DOIs and the invalid cited DOIs, which is available from Peroni, S. (2021). Citations to invalid DOI-identified entities obtained from processing DOI-to-DOI citations to add in COCI (1.0). Zenodo. https://doi.org/10.5281/ZENODO.4625300. These citations to invalid DOIs have been retrieved while processing Crossref data for adding open citations in COCI, but they have not been added in COCI since they point to a non-resolvable cited document.

Two REST API services can be of help: the DOI REST API to check if the invalid cited DOI is now valid; and the Crossref REST API to retrieve the publisher from the prefix of the DOI, both for the cited publications and the citing ones.

Findings

In addition to collecting the names of the publishers involved in these missing citations, either as the publisher of the citing article or as the publisher of the cited article, which was sufficient to answer our research questions, we have decided to collect additional information that can help us to get a better picture of the situation. As regards the JSON file, we found for each individual publisher 1) the number of incorrect given citations metadata sent, and 2) the number of invalid citations received.

On the other hand, as required by the initial research questions, we also extracted the total number of invalid citations that have since been corrected.

Originality/value

The results of this research may point us to publishers who generally send out incorrect citation metadata and, inversely, those who generally receive invalid citations. These findings can first of all raise awareness of the accuracy of certain publishing houses in managing their metadata (or lack thereof). Moreover, finding these trends and showcasing the labor of the corrections may lead to increasingly valid citations if the proper measures are taken.

Research limitations/implications

Based on the available data for the COCI, there may be a slight bias in our sample, causing some publishers to be incorrectly represented.

Minimal bibliography on related projects

Heibi, I., Peroni, S. & Shotton, D. Software review: COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations. *Scientometrics* 121, 1213–1228 (2019). https://doi.org/10.1007/s11192-019-03217-6.

Silvio Peroni, David Shotton (2020). OpenCitations, an infrastructure organization for open scholarship. Quantitative Science Studies, 1(1): 428-444. https://doi.org/10.1162/qss_a_00023.

Peroni, S. (2021). Citations to invalid DOI-identified entities obtained from processing DOI-to-DOI citations to add in COCI (1.0). Zenodo. https://doi.org/10.5281/ZENODO.4625300.

Citation: Nooshin Shahidzadeh, Alessia Cioffi, Arianna Moretti, Sara Coppini (04/28/2021). Investigating Invalid DOIs in COCI. https://dx.doi.org/10.17504/protocols.io.buhjnt4n

DOI

dx.doi.org/10.17504/protocols.io.buhjnt4n

PROTOCOL CITATION

Nooshin Shahidzadeh, Alessia Cioffi, Arianna Moretti, Sara Coppini 2021. Investigating Invalid DOIs in COCI. **protocols.io**
https://dx.doi.org/10.17504/protocols.io.buhjnt4n
Version created by Nooshin Shahidzadeh

WHAT'S NEW

The protocol has been updated with suggestions and comments provided in two reviews received: 1) Deniz Tural. (2021). Review of: "Investigating Invalid DOIs in COCI v1 (protocols.io.bt5xnq7n)". Qeios. doi:10.32388/WHWOI8. 2) Arcangelo Massari. (2021). Review of: "Investigating Invalid DOIs in COCI". Qeios. doi:10.32388/X2DX81.

CREATED

Apr 26, 2021

LAST MODIFIED

Apr 28, 2021

PROTOCOL INTEGER ID

49419

Input material

1 Given the academic background of the project, we got our input material from Peroni, S. (2021). Citations to invalid DOI-identified entities obtained from processing DOI-to-DOI citations to add in COCI (1.0). Zenodo. https://doi.org/10.5281/ZENODO.4625300. As explained in the description of the resource, this dataset contains a two-column CSV file, where the first column ("Valid_citing_DOI") contains the DOI of a citing entity retrieved in Crossref, while the second column ("Invalid_cited_DOI") contains the invalid DOI of a cited entity identified by looking at the field "reference" in the JSON document returned by querying the Crossref API with the citing DOI.
These citations to invalid DOIs have been retrieved while processing Crossref data for adding open citations in COCI, but they have not been added in COCI since they point to a non-resolvable cited document.

Reading the CSV Data

2 First we read the CSV of the invalid DOIs, containing all the invalid cited DOIs in one column and their valid citing counterparts in another.

Citations to invalid DOI-identified entities

Creating the Output JSON File

3 We create a JSON file for the output data that strictly answer the research questions and data for the visualizations of our findings. We opted for JSON file format to store all the data answering the research questions, because this file format allows us to store more heterogeneous information in a more complex and structured way. Indeed, we stored information not directly requested by the research questions, such as the number of missing citations for each single publisher, to provide a more precise answer. For what concerns the visualizations, the JSON file format allows us to

separate different kinds of data and save also data that is relevant for the graphic representations of our findings but not strictly related to the research questions (e.g. which publishers were involved in citations that were originally invalid and then validated with the DOI API request).

3.1 We add a key called "responsible_publishers" that will include a dictionary containing all publisher data for citing articles.

3.2 We add a key called "receiving_publishers" that will include a dictionary containing all publisher data for cited articles.

3.3 We add a key called "total_number_of_corrected_dois" that will hold only a number.

Processing Each Line in the CSV File and Extracting the Needed Information

4 We use the DOI REST API (https://doi.org/api/handles/[doi]) to search for each of the invalid cited DOIs in the CSV. If we receive the response code "1", we know that the citation data is now valid, we now have two cases.
Step 4 includes a Step case.
If the invalid cited DOI is still invalid
If the invalid cited DOI is now valid

Extracting Publisher Data for Citing and Cited Articles

step case

If the invalid cited DOI is still invalid

In this case we will have to extract the publisher data for both the citing and the cited articles.

5 We search through the REST API for Crossref for the id of the publisher, which is the prefix of the DOI (i.e., all the characters preceding the backslash) using: https://api.crossref.org/prefixes/
The request returns a JSON file with information about the member code of the publisher, the name of the publisher, and the prefix.

5.1 We check if the publisher ID exists in the "responsible_publishers" dictionary as a key. If it does not, we create a new item in the dictionary: the key is the ID of the publisher, and the value is a dictionary containing the name of the publisher and the number of invalid citation data it has sent to Crossref (which at the time of creation would be 1). If, on the other hand, it does exist, we just increment the number inside the inner dictionary of that publisher by 1.

6 We assume the first 6 characters of the invalid cited DOI to be the ID of the publisher to which the invalid citation is pointing.

7 We check the validity of this ID through the Crossref REST API. If this ID exists and belongs to a publisher, we add it to the "receiving_publisher" dictionary as a key, the value of which would be a dictionary containing the number of the citations received and the name of the publisher. (Similar to step 4.1 above)
Step 7 includes a Step case.
If the first 6 characters are not a valid ID
If the first 6 characters are a valid ID

step case

If the first 6 characters are not a valid ID

Citation: Nooshin Shahidzadeh, Alessia Cioffi, Arianna Moretti, Sara Coppini (04/28/2021). Investigating Invalid DOIs in COCI. https://dx.doi.org/10.17504/protocols.io.buhjnt4n

In this case we will try to find the publisher name through other ways.

8 We search through the REST API of Crossref for the reference list of the citing article and see if there is any additional data (for example in the "unstructured" field) given in the XML that could lead us to the cited publisher.

8.1 If additional data was found, then we use the "query" attribute of the Crossref REST API to search for the newly found string. If any publication was found, we add the "publisher" field of that publication to the "receiving_publishers" dictionary, in the form explained above.

Output

9 We return the completed JSON file after processing all the lines. (CSV addition possible later.)

Citation: Nooshin Shahidzadeh, Alessia Cioffi, Arianna Moretti, Sara Coppini (04/28/2021). Investigating Invalid DOIs in COCI. https://dx.doi.org/10.17504/protocols.io.buhjnt4n