Evaluating the Availability of Open Citation Data

Chifumi Nishioka[1][0000−0002−1853−3038] and
Michael Färber[2][0000−0001−5458−8645]

[1] Kyoto University Library, Kyoto 606-8501, Japan
nishioka.chifumi.2c@kyoto-u.ac.jp
[2] Institute AIFB, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
michael.faerber@kit.edu

Abstract. Citation data of scientific publications are essential for different purposes, such as evaluating research and building digital library collections. In this paper, we analyze to which extent citation data of publications are openly available, using the intersection of the Crossref metadata and unpaywall snapshot as publication dataset and the COCI dataset as open citation data. We reveal that for 24.2% of the publications, the citation data is openly available, while for 16.6%, the citation data is closed. We find that the percentage of publications with open citation data has increased over the years. We observe that whether publications are published with open access has no influence on whether their citations are openly available. However, publications published in journals from the Directory of Open Access Journals (DOAJ) tend to have more citation data openly available than publications from other journals.

Keywords: Open citation · Open access · Open data.

1 Introduction

Citation data[3] of scholarly publications has been used for various purposes, such as evaluating research impact [12] and building digital library collections [11]. A variety of citation databases have been used so far. However, the providers often charge high subscription fees and the datasets are mainly oriented towards human readability instead of machine readability for allowing data reuse [9]. To resolve these challenges, several organizations have started to make citation data openly available in a machine readable way. One of the most notable projects for open citation data is the *Initiative for Open Citations*[4] (I4OC), which was established in 2017 to promote the unrestricted availability of citation data [9]. The I4OC has made citation data open by encouraging scholarly publishers to publish reference lists of publications which they already deposit to Crossref. It

[3] In this paper, a citation refers to the directional link from a citing bibliographic entity to a cited bibliographic entity.
[4] https://i4oc.org/, last accessed on 04/24/02018

has achieved initial success, with many major scholarly publishers opening their reference lists in this way.

Note that the meaning of "open" in the context of citation data is different from the one in the context of publications (e. g., open access). In the context of publications, "open" indicates that the publications are freely accessible and reusable without restrictions. For citation data, "open" in addition means that the data is structured (i .e., expressed in a machine-readable format) and separate (i .e., available without the need to access a source publication) [8].

Several recent studies (e.g., [4,3]) compare scholarly datasets containing also citation data. However, they do not distinguish between open citation data and closed citation data (i.e., citation data that are intentionally closed by publishers). Heibi et al. [6] present statistics of open citations, focusing on types and publishers. They investigate the state of open citations based on the COCI (*OpenCitations Index of Crossref open DOI-to-DOI references*) dataset [5] (see also Section 2), focusing on each citation. In contrast, this paper looks into the extent to which publications make their citation data open. Specifically, we focus on outgoing citations of publications.

In this paper, we investigate the following questions: (1) What percentage of publications have made citation data open, and how does the percentage vary according to the publication types and publication year? (2) Do open access (OA) publications make citation data more open compared to their toll-access counterparts?

To answer those questions, we conduct an analysis on 100 million publications that are contained in both the Crossref metadata dataset and the unpaywall snapshot. As open citation data, we use the COCI dataset [5]. The citation data of the COCI dataset are originally from publishers. Thus, they are of high quality. We label each publication as "open citation publication," "closed citation publication," or "others" (i.e., without any reference or with references not registered in Crossref) and calculate statistical key figures of various kinds concerning open citation data. The paper contributes to understand the current state of open citation data and their challenges (e.g., for which kinds of publications citation data are not available).

The paper is organized as follows: Section 2 describes the datasets used for our analysis. We report the result of our analysis in Section 3 before concluding the paper in Section 4.

2 Dataset

Fig. 1 summarizes how the publication dataset to be analyzed is generated and how the state of open citation of each publication is identified.

As metadata of publications, we use the crossref metadata dataset as of September 5, 2018.[5] The dataset has been built by querying the Crossref API[6]

[5] `https://archive.org/download/crossref_doi_dump_201809`, last accessed on 04/23/2019

[6] `https://github.com/greenelab/Crossref`, last accessed on 04/23/2019

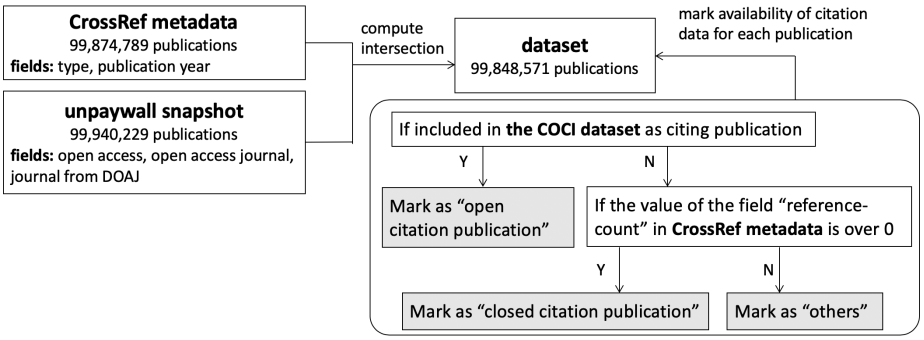

Fig. 1. Pipeline of the dataset generation.

and contains metadata of 99,874,789 publications that are assigned Crossref DOIs. To analyze the influence of OA on the availability of citation data, we use a snapshot of unpaywall[7] from September 24, 2018, together with the Crossref metadata dataset. Unpaywall is a browser extension, which finds legal OA versions of scholarly publications. The snapshot of the database used by the browser extension is publicly available. It covers 99,940,229 publications with Crossref DOIs and contains information regarding the state of OA for each publication. In this paper, we use 99,848,571 publications that are included in both the Crossref metadata dataset and unpaywall snapshot as basis of our analysis.

We mark each publication as (1) "open citation publication", (2) "closed citation publication", or (3) "others". We use the OpenCitations COCI dataset [5] generated on November 12, 2018 [7] as open citation data. Note that the COCI datasets are generated based on the citation data that are deposited to Crossref by publishers. Thus, they are of high quality compared to other datasets.

(1) If a publication is included in the COCI dataset as a *citing publication*, we mark it as open citation publication (i.e., publication that makes citation data open). We identify 24,178,446 of 99,848,571 publications (24.22%) as open citation publication.
(2) Crossref also allows publishers to keep their citation data closed. Thus, there are publications whose citation data are included in Crossref metadata but not publicly available. Publications which are not included in the COCI dataset as citing publications, but have a Crossref's field `references-count` greater than zero, are marked as "closed citation publications". 16,589,545 publications (16.62%) are judged as closed citation publications.
(3) We mark publications that are neither open citation publication nor closed citation publication as "others." 59,080,580 publications (59.15%) are classified as such. "others" include both publications without any reference, such as editorial notes, and publications whose reference data are not registered at Crossref. Although they have different meanings, both the COCI dataset

[7] `https://unpaywall.org/products/snapshot`, last accessed on 04/23/2019

Table 1. State of open citation depending on publication types.

	all	open cit. pub. (%)	closed cit. pub. (%)	others (%)
journal-article	73,397,619	20,120,816 (27.41)	15,258,852 (20.79)	38,017,951 (51.80)
book-chapter	11,726,076	1,048,255 (8.94)	898,513 (7.66)	9,779,308 (83.40)
proceedings-article	5,398,085	2,798,106 (51.84)	342,295 (6.34)	2,257,684 (41.82)
component	3,380,129	9 (0.00)	2 (0.00)	3,380,118 (100.00)
dataset	1,703,078	67,250 (3.95)	2,996 (0.18)	1,632,832 (95.88)

and Crossref metadata dataset do not have a way to express no-value information [1], which indicates that a publication has no reference.

3 Result

In this section, we first report the state of open citation in each type of publication. Then, we present the result in each publication year. Thereafter, we examine whether open access publications promote open citations.

3.1 Type

Crossref DOIs have been assigned to various types of objects, such as journal articles and book chapters. Table 1 shows the state of open citation for the five object types appearing most frequently (given by the number of Crossref DOIs). We observe that the percentage of open citation publications whose type is "journal article" is 27.41%. This is almost the same value as the percentage for all types (24.22% as described in Section 2). We observe a high percentage of open citation publications and a low percentage of closed citation publications in proceedings articles (51.84% and 6.32%, respectively). Regarding other types, it is reasonable that publications whose type is "component" (e .g., figures, supplemental materials) have no reference.

3.2 Publication Year

Fig. 2 shows the acceptance of open citations per year. The bars in Fig. 2 represent the number of total publications, open citation publications, and closed citation publications per publication year. The polygonal lines provide the percentage of open citation publications and closed citation publications per publication year. Please note that the number of publications in 2018 is low because we use the dataset captured in September, 2018. We see that the total number of publications as well as the number of open citation publications has increased over decades. In addition, the percentage of open citation publications has gradually increased. While the percentage has been around 18% before 1998, it reached 36% in 2011. Although older publications are read less [10], making the citation data of older publications available is important to trace the evolution of scholarly knowledge. Regarding closed citation publications, we see that the percentage decreases slightly over the years.

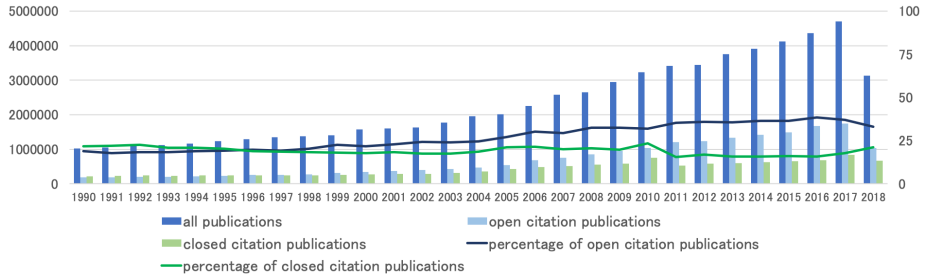

Fig. 2. State of open citation along with publication year.

Table 2. State of open citation depending on whether a publication is open access (OA).

	all	open cit. pub. (%)	closed cit. pub. (%)	others (%)
OA	24,961,752	7,077,997 (28.36)	2,391,627 (9.58)	15,492,128 (62.06)
non-OA	74,886,819	17,100,449 (22.84)	14,197,918 (18.96)	43,588,452 (58.21)

3.3 Open Access

This section examines the influence of open *access* on the state of open citation.

Open Access (OA) We first investigate whether there is a difference in the state of open citation between open access (OA) publications and non-open access (non-OA) publications. Table 2 shows the number of total publications, open citation publications, and closed publications as well as their percentages. Referring to the state of OA, Table 2 indicates that 25.00% of the publications are OA in some way (e.g., gold OA, green OA), which is consistent with the result given by Piwowar et al. [10]. OA publications are more likely to have their citation information open than non-OA publications. However, the difference is small. One possible reason is that if a publication is hybrid OA or green OA, publishers may not make citation data open.

Open Access (OA) Journal We examine whether there is a difference in the state of open citation between publications in an OA journal and those in non-OA journals. Table 3 presents the number of all publications, open citation publications, and closed publications as well as their percentages. Unlike the result in Table 2, publications in OA journals make citation data more open. Compared to publications in non-OA journals, publications in OA journals are categorized as "others" more frequently. The reason might be that many small publishers publish OA journals but do not have enough resources to organize citation data.

Table 3. State of open citation depending on whether a journal of a publication is an open access (OA) journal.

	all	open cit. pub. (%)	closed cit. pub. (%)	others (%)
OA journal	5,670,103	1,613,260 (28.45)	280,547 (4.95)	3,776,296 (66.60)
non-OA journal	94,178,468	22,565,186 (23.96)	16,308,998 (17.32)	55,304,284 (58.72)

Table 4. State of open citation depending on whether a journal of a publication is included in DOAJ.

	all	open cit. pub. (%)	closed cit. pub. (%)	others (%)
in DOAJ	3,227,017	1,580,857 (48.99)	273,392 (8.47)	1,372,768 (42.54)
not in DOAJ	96,621,554	22,597,589 (23.39)	16,316,153 (16.89)	57,707,812 (59.72)

DOAJ DOAJ (Directory of Open Access Journals) is an online platform that hosts a curated list of OA journals. The project defines OA journals as scholarly journals which make all their content available for free. These journals meet high quality standards, notably by exercising peer review or editorial quality control[8]. As of July 3, 2019, over 13,000 journals have been registered at DOAJ. We verify whether there is a difference in the state of open citation between publications from a journal in DOAJ and those from a journal not in DOAJ. The results are shown in Table 4. The percentage of open citation publications for publications in DOAJ is significantly higher than that for publications not in DOAJ. Since DOAJ lists journals that meet quality standards, citation data of publications are properly organized by publishers and deposited to Crossref. However, a certain amount of citation data (8.47%) is closed.

4 Conclusion

In this paper, we analyzed the current state of open citation based on the COCI dataset. We found out that for 24.22% of the publications with assigned Crossref DOIs, the citations were open. As described by Di Iorio [2], such a value is not sufficient for evaluating researchers. 16.62% of the publications are identified as closed citation publication and 59.15% are "others". "Others" include both publications without any reference and publications whose reference data are not registered to Crossref. Although they have different meanings, the existing datasets do not have a way to express no-value information. We found that the percentage of open citation publications has gradually increased over the years. We did not observe a difference regarding the availability of open citation data between OA publications and the non-OA publications. However, publications published in a journal listed in the Directory of Open Access Journals (DOAJ) were more likely to open the citation data than those published in other journals.

[8] `https://doaj.org/publishers`, last accessed on 05/14/2019

References

1. Darari, F., Prasojo, R.E., Nutt, W.: Expressing no-value information in RDF. In: Proceedings of the International Semantic Web Conference (ISWC) Posters and Demonstrations Track. CEUR Workshop Proceedings (2015)
2. Di Iorio, A., Peroni, S., Poggi, F.: Open data to evaluate academic researchers: an experiment with the Italian Scientific Habilitation. arXiv preprint arXiv:1902.03287 (2019)
3. Färber, M.: The Microsoft Academic Knowledge Graph: A Linked Data Source with 8 Billion Triples of Scholarly Data. In: Proceedings of the 18th International Semantic Web Conference. ISWC'19 (2019)
4. Haunschild, R., Hug, S.E., Brändle, M.P., Bornmann, L.: The number of linked references of publications in Microsoft Academic in comparison with the Web of Science. Scientometrics **114**(1), 367–370 (2018)
5. Heibi, I., Peroni, S., Shotton, D.: Coci, the opencitations index of crossref open doi-to-doi citations. arXiv preprint arXiv:1904.06052 (2019)
6. Heibi, I., Peroni, S., Shotton, D.: Crowdsourcing open citations with croci-an analysis of the current status of open citations, and a proposal. arXiv preprint arXiv:1902.02534 (2019)
7. OpenCitations: COCI CSV dataset of all the citation data (Version 3) (2018), https://doi.org/10.6084/m9.figshare.6741422.v3
8. Peroni, S., Shotton, D.: Open Citation: Definition (2018). https://doi.org/10.6084/m9.figshare.6683855.v1
9. Peroni, S., Shotton, D., Vitali, F.: One year of the OpenCitations Corpus. In: International Semantic Web Conference. pp. 184–192. Springer (2017)
10. Piwowar, H., Priem, J., Larivière, V., Alperin, J.P., Matthias, L., Norlander, B., Farley, A., West, J., Haustein, S.: The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. PeerJ **6**, e4375 (2018)
11. Smith, L.C.: Citation analysis. Library Trends **30**(1), 83–106 (1981)
12. Todeschini, R., Baccini, A.: Handbook of bibliometric indicators: quantitative tools for studying and evaluating research. John Wiley & Sons (2016)