Open Access Research Outputs Receive More Diverse Citations

Chun-Kai (Karl) Huang, Cameron Neylon, Lucy Montgomery, Richard Hosking, James P. Diprose, Rebecca N. Handcock, Katie Wilson

ABSTRACT

The goal of open access is to allow more people to read and use research outputs. An observed association between highly cited research outputs and open access has been claimed as evidence of increased usage of the research, but this remains controversial. A higher citation count also does not necessarily imply wider usage such as citations by authors from more places. A knowledge gap exists in our understanding of who gets to use open access research outputs and where users are located. Here we address this gap by examining the association between an output's open access status and the diversity of research outputs that cite it. By analysing large-scale bibliographic data from 2010 to 2019, we found a robust association between open access and increased diversity of citation sources by institutions, countries, subregions, regions, and fields of research, across outputs with both high and medium-low citation counts. Open access through disciplinary or institutional repositories showed a stronger effect than open access via publisher platforms. This study adds a new perspective to our understanding of how citations can be used to explore the effects of open access. It also provides new evidence at global scale of the benefits of open access as a mechanism for widening the use of research and increasing the diversity of the communities that benefit from it.

1. INTRODUCTION

The purpose of research is for it to be used, either applied to solve problems and address issues, or more narrowly to provide insight, capacity, and inspiration for further research. The open access (OA) movement is founded on the goals of putting research in the hands of more people and making it more usable (e.g., the Budapest OA Initiative) (Chan et al., 2002). A seismic shift in access models for scholarly outputs (i.e., from subscription-based models to OA models) has occurred over the past decade with accessible outputs (i.e. can be read or downloaded without payment) rising from approximately 27% of global outputs published in 2011 to over 49% of all outputs published in 2020 being accessible in some form (Neylon & Huang, 2022).

It remains challenging to conclusively demonstrate the benefits of this shift in access models for scholarly outputs. Case studies and qualitative research approaches have helped to shed light on complex relationships between access models, use and impact. Studies have sought quantitative evidence of enhanced usage via a variety of methods. Some have observed associations between increased citation counts and OA, providing the most global evidence of enhanced article usage (Piwowar et al., 2018; Archambault et al., 2014; Bautista-Puig et al., 2020). However, there are several confounding factors that weaken claims of a causal link between OA and enhanced use of research outputs (Basson et al., 2021; Dorta-González et al. 2017). A set of narrowly defined randomised control trials finds no effect (Davis, 2011), and there is an argument that access to academic resources and prestige may well be associated with both the choice to make an output OA and the likelihood of higher citations (Lewis, 2018; Sotudeh, 2020; Hua et al., 2016; Zhang & Watson, 2017).

In addition, we feel that the focus on citation counts fails to address the core goals of OA, specifically that a wider range of research users has more access (Dahler-Larsen, 2018; Linkov et al., 2021; Neylon et al., 2021). We need a different approach to quantify the impact of OA focusing on widening the diversity of users who are able to access scholarly content. Recent advances in data availability and processing mean that we are now able to identify the affiliations of citing authors at scale and hence quantitatively assess the institutional and geographic diversity of citing authors globally. Similarly, we can analyse the fields of research across citing outputs. We refer to these measures under one umbrella term: *citation diversity*.

1.1. Relevant research

There is limited amount of scholarly literature that investigates the relationships between OA and its potential impact on widening the geographic and interdisciplinary dissemination and use of research. Two most closely related works are Young & Brandes (2020) and Neylon et al. (2021). The former reported that OA articles received more interdisciplinary diverse citations than non-OA articles, although only data from two journals were studied. Neylon et al. (2021) showed that OA books garnered more diverse usage (via geographic locations of downloads) as compared to closed books. This study showed not only that OA books are cited and downloaded more than their closed counterparts, but also that they are downloaded by a wider audience. A few other studies (though less concerned with OA) explored the diversity of references and co-authorships. Linkov et al. (2021) proposed the Linguistic Diversity Index as a scientometric measure of the linguistic diversity of sources cited in articles. This index is aimed at encouraging the use of sources from more diverse cultural groups, placing higher importance on rarely represented cultural groups. Naik et al. (2023) showed that the geographic diversity (by air transport network) in co-authorships as having a positive impact on citation counts, albeit at varying levels of strength across different subject areas.

1.2. Contribution

The objective of the current article is to explore the relationships between OA and citation diversity. We do this by examining the geographic locations of author affiliations, and the fields of research, of citing outputs. We use the diversity of these citing outputs as a proxy for wider dissemination of research. Through this our goal is to define the impact of OA on the wider use of research. The study extends previous work and adds to the literature in the following ways:

- The study extends the concept of citation diversity to consider the geographic locations of author affiliations of citing outputs in addition to the fields of research.
- The study draws on publicly available datasets that include 19 million research outputs and 420 million citation links worldwide, making it the largest study of this type to date.
- The large-scale data also enables the study to explore the robustness of the results by comparing results across time, different measures of diversity, various groupings of citation-affiliation links, citation counts, and examining their dependencies.
- The study also takes a first exploration in examining whether there are differences across geographic regions in terms of how OA influences citation diversity (e.g., where increased citations come from).

The rest of the article is structured as follows. In Section 2, we provide details of the data and methods used for this study. Section 3 includes the main results from the analysis, with summarised discussions on the robustness of the results provided in Section 4. We provide detailed discussions of the results in Section 5, including implications for further research. Section 6 concludes the study. Additional information and results are provided in Supplementary material.

2. METHODS

We quantify citation diversity using two different standard measures of diversity that are less sensitive to citation counts. This helps us to address the issues of access to resources and prestige that are potential confounders (Lewis, 2018; Sotudeh, 2020; Hua et al., 2016; Davis, 2011; Zhang & Watson, 2017) in analyses based simply on citation counts which remain with more sophisticated measures such as citation velocity, as shown in previous research (Hutchins et al., 2016; Seppänen et al., 2022).

For our analysis we extracted all research outputs with publication years from 2010 to 2019 (see Section 2.1. for details). For each of the 19 million outputs, we extracted citation counts (from the total of 420 million citation links), metadata of their citing outputs and citing author affiliations, and calculated the Shannon Entropy (or Shannon Index) and the Gini-Simpson Index (or Gini's Diversity Index) as measures of citation diversity. Higher scores for these indices are indicators of more citation diversity. We consider citation diversity based on five different ways of grouping citation links: by institutions, countries, subregions, regions, and fields of research (i.e., citing actors). Fig. 1 demonstrates how citation diversity assessed using these indices is different from traditional citation counts. Two outputs can have a very different diversity of citing actors despite having equal citation counts. For instance, an article that is cited from a wider range of institutions but has the same number of citations will have a greater citation diversity.

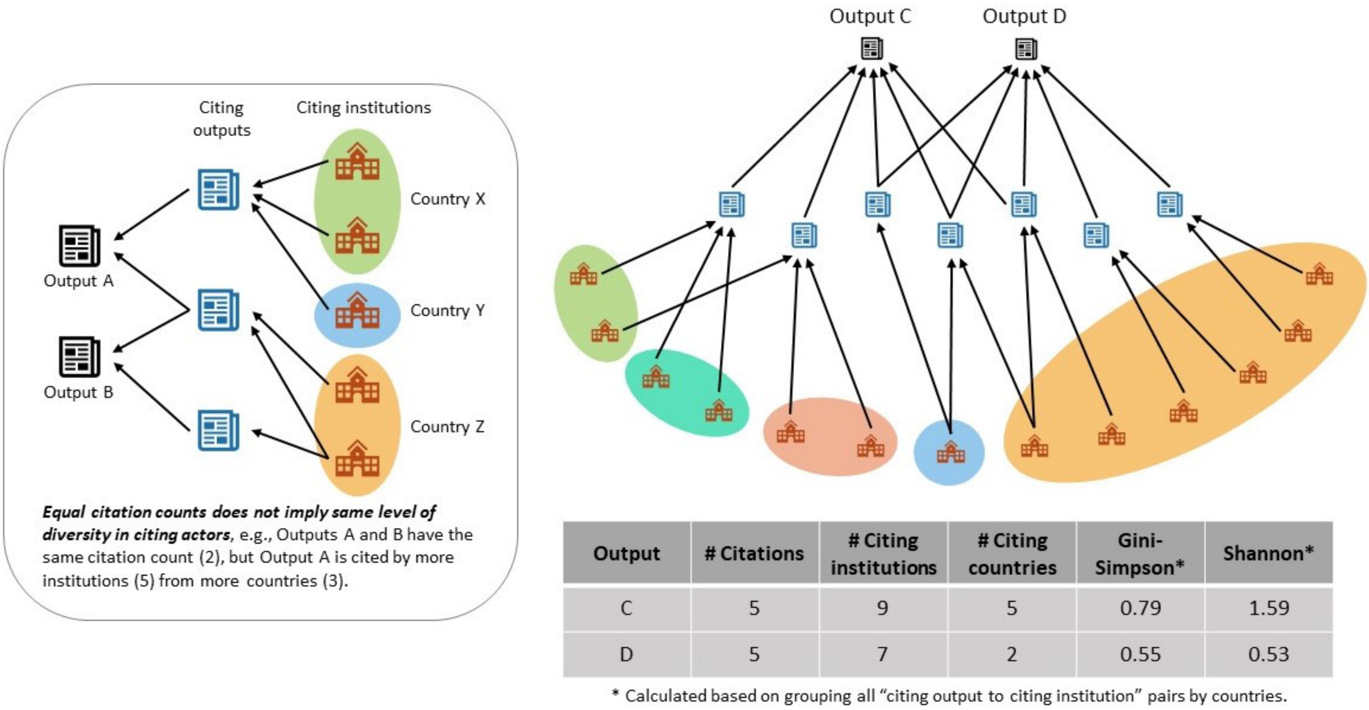
2.1. Input data

COKI Academic Observatory

The COKI Academic Observatory (https://github.com/The-Academic-Observatory) is a large-scale relational database tracking open knowledge performance of research institutions worldwide. It is designed to be an open source, community-driven and community accessible resource, built around re-usable cloud architecture and transparent assumptions. It is developed by the COKI project, based at Curtin University. The project aims to create the world's leading data infrastructure on scholarly communication, OA, diversity, inclusion, and more. COKI's goal is to ensure that the tools and data used to evaluate scholarly outputs and research institutions support more open and productive practices, so that universities can change the stories they tell about themselves, and to put open knowledge at the centre of these narratives.

To analyse citation diversity, we used the data workflows and datasets developed by the Curtin Open Knowledge Initiative (COKI) for analysis of open knowledge performance. The COKI Academic Observatory data collection pipeline (Hosking et al., 2022) is used to create the Academic Observatory dataset which is used to analyse citation counts, affiliations and diversity. This pipeline integrates data from Crossref Metadata (DOIs, publication dates), Unpaywall (OA status), Microsoft Academic Graph (MAG) (institutional affiliations, citation links, fields of research; since this study was completed this has been replaced with OpenAlex), Research Organization Registry (ROR) (institutional information) to generate the "DOI Table" - an enriched metadata source on research outputs.

Fig. 1: Illustrations of citation diversity compared to citation count.



Equal citation counts does not imply same level of diversity in citing actors, e.g., Outputs A and B have the same citation count (2), but Output A is cited by more institutions (5) from more countries (3).

Output	# Citations	# Citing institutions	# Citing countries	Gini-Simpson*	Shannon*
C	5	9	5	0.79	1.59
D	5	7	2	0.55	0.53

* Calculated based on grouping all "citing output to citing institution" pairs by countries.

Illustrative examples to demonstrate differences between citation counts, number of citing actors, and diversity measures. Outputs with equal citation counts do not necessarily have the same level of diversity in citing actors. Citing outputs are affiliated to institutions and these institutional-links can be grouped by their locations. These provide the basis for calculating diversity measures. Only country level diversity scores are provided in the figure. See Section 2.3. for details of calculating the Shannon Entropy (or Shannon Index) and the Gini-Simpson Index (or Gini's Diversity Index). Left: Output A and Output B both have two citations. However, Output A is cited by institutions from two different countries, while all citing institutions for Output B are from Country Z. Hence, Output A has a higher level of citation diversity by country. Right: Output C and Output D both have five citations. However, Output C has both more citing institutions, and these institutions are from more countries. This implies Output C has a higher level of citation diversity by country, i.e., higher score in diversity measures.

These datasets are updated on a regular cycle with MAG updated fortnightly (before it was retired) and Crossref Metadata updated monthly. The specific instances of the tables used directly are:

- `academic-observatory.observatory.doi20220730`
- `academic-observatory.mag.PaperReferences20211206`

We filter all DOIs to those that also have "PaperIDs" from MAG and to publication dates from 2010 to 2019 (both inclusive). The date range is selected based on our confidence in data quality and also considerations given to the fact that most new outputs would have had little time to attract citations. We use the final data extraction of MAG (11 December 2021) for analysis.

The full data in the time range includes 37 million outputs with 424 million citation links. However, only outputs with two or more citations are applicable (non-trivial) in the calculations of citation diversity measures. This resulted in the final data of 19 million outputs and 420 million citation links between these outputs.

2.2. Terminology

For the ease of reference, we include a list of terms and variables used in this study, along with definitions and descriptions of the underlying data.

Term/Variable	Definition	Data description
Citing output (or paper)	If Output A references Output B, then Output A is a citing output for Output B.	Citing outputs are identified using the table of reference links in MAG. Only outputs with Crossref DOIs are included.
Citing actor	An actor upon which a citing output can be affiliated or characterised, e.g., author, author affiliation, country of author affiliation, field of research, etc.	Once a citing output is identified (as above), the authors, author affiliations, and fields of research are identified using MAG. Then geographic locations of author affiliations (e.g., country, subregion, region) are identified using ROR.
Citing author	An author of a citing output.	See data description for "Citing actor".
Citing institution	An institution that is included as an author affiliation in a citing output.	See data description for "Citing actor".
Citing country	The country of a citing institution.	See data description for "Citing actor".
Citing subregion	The subregion of a citing institution.	See data description for "Citing actor".
Citing region	The region of a citing institution.	See data description for "Citing actor".
Citation count	The number of citations (citing outputs) that an output received.	The is calculated by counting the number of citing outputs. See data description of "Citing outputs".
Fields of research	The field of research (or subject discipline) assigned to a given output.	We use data as defined by MAG's "Level 0 fields" under the MAG hierarchy of "concepts" assigned to outputs. Level 0 is the highest-level parent concepts in the MAG concepts hierarchy, i.e., most broad terms.
Citation diversity	A term describing how diverse an output's citations are in terms of their links to citing actors.	There are many ways to quantify diversity of the output-to-citing actor links. The Gini-Simpson Index and the Shannon Entropy are used for this study. These are calculated using MAG's citation links and fields of research assignments, and ROR's organisational metadata.
Gini-Simpson Index (Gini's Diversity Index)	A diversity measure that quantifies the probability that two randomly selected output-to-citing actor links belong to the same citing actor group.	For a given output, the Gini-Simpson Index is calculated by first placing all output-to-citing actor links into bins according to different groups of citing actors (e.g., countries). Then proportions of these links in each bin (country) is calculated. Finally, the Gini-Simpson Index formula is applied.
Shannon Entropy (Shannon Index)	A diversity measure that quantifies the level of uncertainty in predicting the citing actor group assignment of a randomly selected output-to-citing actor link.	This is calculated following the same process as Gini-Simpson Index except with the formula replaced by the Shannon Entropy formula.
Open access outputs (OPEN)	Outputs that are freely accessible via either publisher platforms or	An output's OA status is determined as per Unpaywall data.

	OA repositories.	
Gold outputs (GOLD)	Outputs that are freely accessible via publisher platforms with open licences.	An output's GOLD status is determined as per Unpaywall data.
Green outputs (GREEN)	Outputs that are freely accessible via disciplinary and institutional repositories.	An output's GREEN status is determined as per Unpaywall data.
Closed outputs (CLOSED)	Outputs with no OA copies available from publisher platforms nor open repositories	An output's OA status is determined as per Unpaywall data.
Open access citation advantage (OACA)	This is a statement implying that OA increases the number of citations that an output receives.	NA
Open access citation diversity advantage	This is a statement implying that OA widens the diversity of citations that an output receives.	NA
Percentage (%) change in total citations	Total citations to OA outputs minus total citations to non-OA outputs, then divided by total citations to non-OA outputs, and multiplied by one hundred.	This is calculated for a specified set of target outputs, and citing outputs from a specified citing actor. For example, we may be interested in the citations from subregion Y to subregion X. Hence, outputs from subregion X are split into OPEN and CLOSED outputs. Citing outputs for each of these outputs from subregion Y are identified and counted. Then the percentage change in total citations is calculated.
Percentage (%) ratio in average citations	The average number of citations to OA outputs, divided by the average number of citations to non-OA outputs, and times by one hundred.	This is calculated in a similar way to the above, except the citation count is averaged across the number of target outputs in each of OPEN and CLOSED sets.
Kernel density estimate (KDEs)	This is a non-parametric estimate of the probability density function of a given random variable. In our study we are interested in comparing KDEs between OPEN and CLOSED outputs.	This is created using the `create_distplot` function in the Plotly Figure Factory package in Python. We applied this estimation to 10,000 sampled OPEN outputs and 10,000 sampled CLOSED outputs, respectively, for each combination of diversity measure, citing actor type, and years of publication.

2.3. Analysis methodology

As shown in Fig. 1 our unit of analysis is the affiliation link or field of research associated with an incoming reference to a given output. We calculate the Shannon Entropy and Gini-Simpson Index scores of the set of affiliations associated with citing outputs, with respect to groupings by institutions, countries, subregions, and regions, and also the MAG "Level 0 fields" (aka "fields of research") associated with citing outputs. These two diversity measures provide complementary quantifications of diversity in the citing affiliation/field links associated with individual cited outputs. We note that a "citation link" refers to an output-to-output link via referencing, whereas a "citing affiliation link" or "citing field link" is a further step forward determining the link between an output and an affiliation associated with a citing output, or between an output and the field of research associated with a citing output, respectively. More generally, we refer to these as "output-to-citing actor" links, where the citing actors may be institutions, countries, subregions, regions, or fields of research associated with the citing output.

We define R as the number of groups (e.g., countries, fields of research) and p as the proportion of output-to-citing actor links assigned to a given group. The Shannon Entropy quantifies the level of uncertainty in predicting the group assignment of a randomly selected output-to-citing actor link as:

$$1 - \sum_{i=1}^{R} p_i^2$$

Whereas the Gini-Simpson Index measures the probability that two randomly selected output-to-citing actor links belongs to the same group:

$$-\sum_{i=1}^{R} p_i \ln p_i$$

with $\ln p_i$ as the natural logarithm of p_i.

The analysis is implemented in template SQL queries that are run via an automated reporting framework implemented in Python. The first step is the aggregation of the affiliations associated with incoming citations for each of the 37 million outputs and 424 million citation links in the target time period. The resulting table `citation_diversity_global` is stored in Google's cloud-based BigQuery database. Subsequent analyses and corresponding SQL queries further filter this down to outputs with two or more citations, which corresponds to 19 million outputs with 420 million citation links. The decision to only consider outputs with two or more citations is based on the fact that measuring diversity for outputs with zero citations is nonsensical and outputs with only one citation will trivially be assigned a diversity score of zero. However, these outputs are kept in the table above for validation purposes.

Subsequent analysis steps are implemented in template SQL queries of the cloud-based database with the resulting data downloaded as comma delimited text files (CSVs) suitable for use in the Pandas Python library and stored locally. These local data are then used to generate the tables and graphs in this article. The full process from source data to final outputs is specified in code and automated to support reproducibility and enable detailed critique (Huang et al., 2022).

For this study, we consider four different (but potentially overlapping) categories of outputs: OPEN, GOLD, GREEN, and CLOSED (see Section 2.2 for definitions). Results are compared across these different categories in relation to their impact on citation diversity, where necessary. We also use percentage ratios in average citations and percentage changes in total citations (see previous section for definitions) to examine where increased citations come from and use these to compare the levels of *OA citation diversity advantage* across different subregions and regions.

2.4. Caveats

Statistical Significance

In this study we have avoided using statistical significance as a measure of the likelihood of an effect. There are several reasons for this choice. Firstly, we are predominantly dealing with a population of outputs rather than targeted samples of outputs. This includes all outputs captured by a system that aims to include worldwide research outputs that have Crossref DOIs and MAG PaperIDs. Second, given the large numbers of outputs included in most of our analyses, the resulting p-values are both diminutive and highly associated with sample sizes chosen, making them less useful as a measure of confidence. Third, comparing statistical significance across a large number of groups, where groups also differ widely in distribution, is highly challenging. This would entail considerations for both the effects of multiple comparisons and advanced sampling procedures. On the other hand, downstream distributional analyses of large numbers of outputs are also not practical. Given the above, we have taken the alternative in exploring the consistency of the OA citation diversity advantage across multiple ways of analysing the corpus of outputs. However, where possible, we have included some subsampling analyses to emphasise that this consistency is maintained across comparable but small samples relative to the whole data.

Data Limitations

We acknowledge the following limitations in the data used for our analysis:

- Research outputs included in our analysis are those that are assigned DOIs by Crossref. We acknowledge that there are other DOI registration agencies that assign DOIs to research outputs (e.g., China National

Knowledge Infrastructure - CNKI) and these are not currently indexed in our system. Consequently, there may be limitations in our coverage of certain areas of Asia, Sub-Saharan Africa and other regions. There are also general issues with coverage of certain fields of research where DOIs are not traditionally used in scale (such as in Art, Political Sciences, etc.). In addition, there may be issues of moving windows in terms of assignments of outputs to fields of research, as results of both cultural and methodological changes over time (e.g., Engineering outputs being assigned to Material Science and Computer Science in more recent years).

- The data on fields of research used in our analysis are directly extracted from MAG. MAG used machine learning approaches to classify research outputs into "concepts" and build a hierarchy of these concepts (Wang et al., 2020). We only use the concepts specified in level 0 (most broad or highest parent concepts) of the hierarchy. It is possible that our results based on fields of research may differ if a different set of data on fields of research or subject disciplines is used. We should also note that MAG is now discontinued, and an alternative source will be used in future work (e.g., OpenAlex).
- Our definition of citation diversity is based on the distribution of "output-to-citation actor links" across citation actor groups. This does imply that if a citing output has multiple authors belonging to multiple affiliations, then it will possibly infer multiple output-to-citing actor links. In other words, this citing output may belong to multiple regions (for example). This may have an impact on low-citation outputs with at least one citing output with extraordinarily large number of authors from multiple affiliations. However, our quality checks revealed very low number of such cases (i.e., outliers) and they have no obvious impact on the overall findings. There may be other ways to define or measure citation diversity that incorporate such cases.

3. MAIN RESULTS

3.1. Comparing OA categories

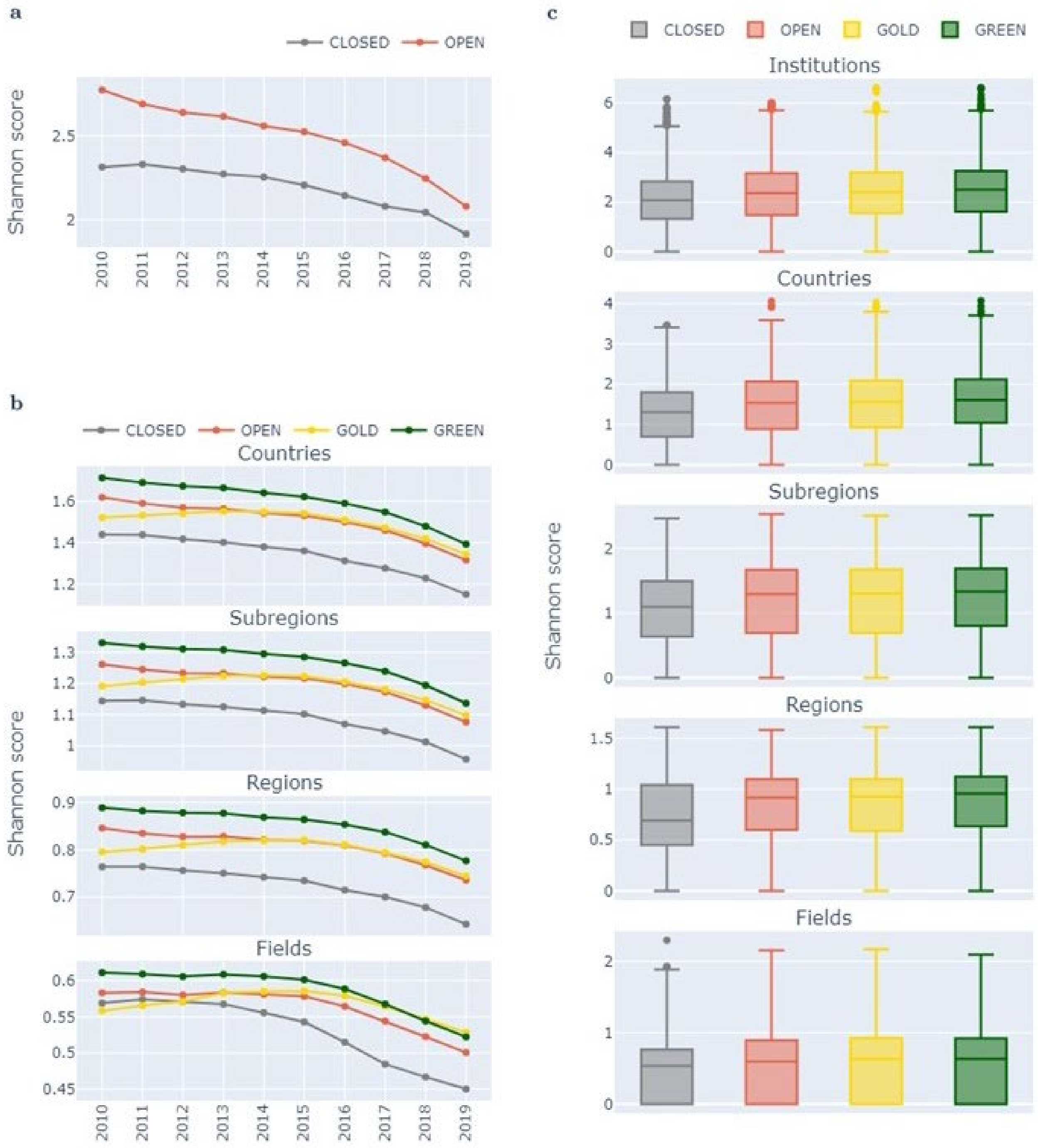
As a first step in our analysis, we confirm the previously observed OA citation advantage, for the first time at a global scale. We observe an association of OA with higher citations at the global scale, consistent with previous literature on OA citation count advantage but with the known caveats described earlier. We see that this association is robust across years of publication, and OA categories (See Robustness of results). Further work on this OA citation count advantage using global datasets could help to reveal what factors are associated with these complex effects. We also characterise the citations by the number of unique citing institutions, countries, subregions, regions, and fields of research. Again, a robust advantage for OA categories is observed (with a few existing exceptions) which offers avenues for further analysis of the causal effects underlying the citation diversity advantage for OA (see Robustness of Results for details).

Turning to our main focus, the diversity of citations, our results showed an enhanced diversity of citing institutions, countries, subregions, and regions for OA research outputs, with this effect being consistently observed across all publication years since 2010 (see Fig. 2a and Fig. 2b), and across almost all fields of research in our study data. There are differences over time, between fields of research and between author's country of affiliation in the scale of the effect, as well as the underlying diversity measures. These are interesting areas for future study. What is striking is how consistent the observed effect is across all these potential groupings. This includes distributional shifts toward higher diversity scores for OA outputs (relative to CLOSED outputs) for all citing actor groups, publication years, and both diversity measures. Fig. 2c demonstrates some of those distributional shifts. Although the shift can be small in some cases, it is consistent across almost all comparisons for various groupings. See Robustness of Results and Supplementary Material for results across all different groupings.

When comparing mechanisms of OA, we see a larger effect in the diversity of citing countries, subregions, regions, and fields of research across all years, and for access provided through repositories (i.e., GREEN outputs) than for OA provided via publisher websites (Fig.2b and Fig. 2c). This effect shows interesting discipline and author-country effects which merit further investigation.

The debate over the citation count advantage is dominated by questions of confounding effects, specifically whether OA is more accessible to researchers from wealthier and more prestigious institutions and/or whether researchers selectively make their best work OA. To address this, we also showed that the citation diversity advantage is present, independent of citation counts (see Robustness of Results). The lack of overall correlation between citation count and citation diversity provides evidence that citation count and citation diversity track different aspects of usage and that there is limited common confounding at the global scale. However, this correlation is higher for outputs with low citation numbers. The cohorts of outputs published in later years have higher proportions of low-citation outputs (i.e., less time to accumulate citations), which may partially explain the downward trends in the median citation diversity scores (Fig. 2a and Fig. 2b). A more in-depth analysis is needed through further research.

Fig. 2: Comparing citation diversity between OA categories.



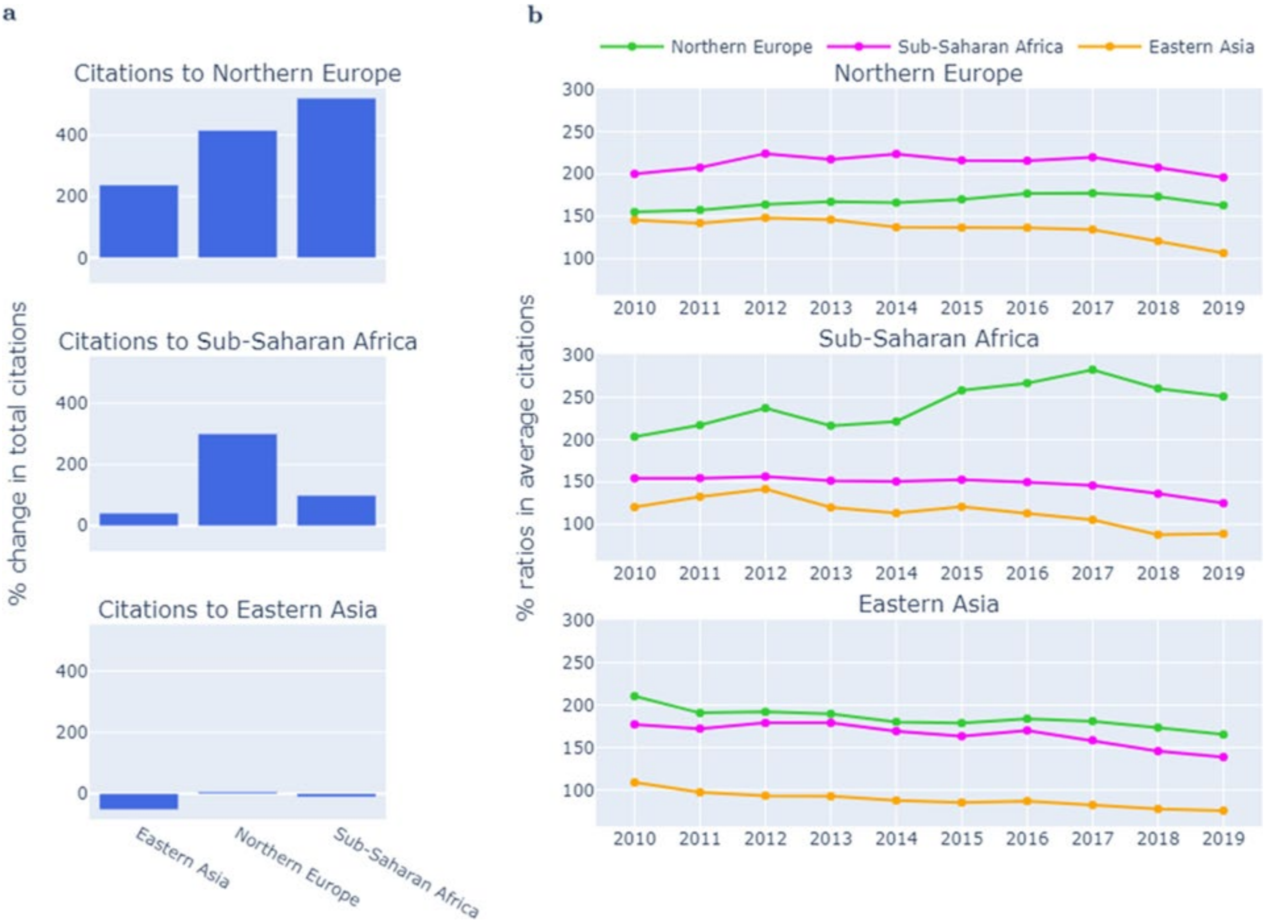
a. The median Shannon scores by citing institutions are compared between OA and CLOSED outputs over a ten-year period. Earlier outputs receive higher scores as a result of having had more time to garner citations (hence more possibility of wider citing affiliations). However, it is consistently observed that OA outputs perform better in the diversity of citing institutions for all years. **b.** The mean Shannon scores are compared across the OA categories, with the scores calculated based on the grouping of citing affiliation links by countries, subregions, and regions, and citing outputs by fields of research. For the first three cases, all OA categories consistently outperform CLOSED outputs. OA outputs also outperform CLOSED outputs for the fields of research in more recent years. This is likely a result of evolving research practices and data quality levels. We also note that the scores should not be compared across different citing actor types as they measure different ways of grouping citing actor links (hence different scales). **c.** Boxplots of Shannon scores for samples of 2017 outputs are provided for various citing actors and compared between various OA categories. Equal size samples are used across OA categories for this comparison.

As an observational cohort study, our analysis is not able to confirm the exact causal links between OA and enhanced citation diversity. However, as a global analysis we can definitively say that within the full cohort in our dataset of 19 million outputs, OA outputs have a greater level of citation diversity. This is demonstrated through both summary statistics and distributional analyses.

3.2. Comparing geographies

To further understand where increased citation diversity comes from and how it compares across geographies, we also examine the geographical distribution of citations. We do this by examining the percentage change in total citations and the percentage ratio in average citations across OPEN and CLOSED outputs for all pairs of subregions and regions. These represent the levels of change in citation from a specific subregion or region when moving from CLOSED to OPEN outputs. As a miniature demonstration, Fig. 3 shows differences between OPEN and CLOSED outputs with respect to citations to and from three selected subregions.

Fig. 3: Changes in citations to and from selected subregions.



a. The three graphs resemble selected citation links to outputs by the subregions: Northern Europe, Sub-Saharan Africa, and Eastern Asia, respectively. Within each graph, the percentage change in total citations (see Robustness of Results) from the three selected subregions (for 2019) are shown. A value above zero indicates a positive effect for OA. While both Northern Europe and Sub-Saharan Africa benefit from OA outputs, there are differences in the results. Eastern Asia is one of the exceptions resulting from less comprehensive coverage by Western bibliographic systems. **b.** An alternative measure is used to track differences in mean citations between OA and CLOSED outputs - percentage ratios (see Methods). The results are provided for all years included in the study. A value above 100 indicates a positive effect for OA.

The top panel in Fig. 3a shows that numbers of citations from all three subregions to outputs affiliated with Northern Europe have increased when moving from CLOSED to OPEN outputs. The increase is greatest for citations from Sub-Saharan Africa (almost five folds), with citations from Northern Europe itself increasing by 400 percent and citations from Eastern Asia increasing by over 200 percent. The middle panel in Fig. 3a similarly shows positive impacts for Sub-Saharan African outputs, albeit at much lower levels. Eastern Asia (Fig. 3a bottom panel) represents an interesting case where the impact of OA on citations seems to be little to negative. This is likely due to local policies and the lack of comprehensive coverage of data from the subregion. Fig. 3b alternatively describes the changes in citations using percentage ratio in average citations. However, the same trends are observed and are consistent over time. In this miniature example we see that outputs affiliated with Northern Europe benefit most from both the highest increased citations to its OA outputs (i.e., highest increased usage by all subregions), and for the highest increased citation of Northern European outputs to Sub-Saharan Africa.

This pattern is also observed for the larger analysis comparing all subregions and regions. Analysing the subregions where the affiliations of citing outputs are located, we see an increase in citations to OA outputs from traditionally under-represented institutions based in subregions with fewer research resources (e.g., as measured in World Bank Statistics on research expenditure) (The World Bank, 2022). This is consistent with greater access to OA being linked to greater use of OA outputs from these subregions, at least as measured by citations (see Robustness of Results). However, the citation diversity advantage also accrues preferentially to traditionally prestigious centres of research.

Overall, we see that traditionally "prestigious" centres of excellence (in terms of wealth and scale, e.g., Northern Europe, North America) benefit most from both increases in citations to their OA outputs (i.e., usage of their outputs by other subregions), and increases in citations from their outputs to OA outputs of other subregions (i.e., their usage of outputs affiliated to other subregions). There are also signals that the level of OA citation diversity advantage is lower overall for outputs with affiliations from traditionally underrepresented subregions or regions (e.g., Sub-Saharan Africa, Northern Africa, Latin America), but show an increase over time from low or negative levels. This may be evidence of increasing visibility over the period of study, which could be linked to OA shifting discovery pathways. However, more work is required to investigate these effects in detail. See Robustness of Results and Supplementary Material for the full set of results.

4. ROBUSTNESS OF RESULTS

To ensure the robustness of our results we include the analyses of our results compared across multiple ways of grouping the data – over time, different diversity measures, citation counts, OA categories, different affiliation groupings by geographic assignments, summary measures, etc. All results are provided in the Supplementary material, with the main findings summarised below.

4.1. A consistent effect across time, measures, and categories

As mentioned earlier, we reproduce the previously described citation count advantage across the whole dataset. We see an association of OA (all categories) with higher citation counts for all years in the analysis. We also note the overall decreasing trend of citation counts due to more recent outputs having fewer citations. These results are presented in Supplementary Figures A.

We then turn to the analysis of the output-to-citing actor links. We start by examining counts of unique citing actors characterised by institutions, countries, subregions, regions, and fields of research. In other words, for each cited output, we count the number of unique citing institutions, countries, subregions, regions, and fields of research, respectively, combining all its citing outputs. The mean and median number of unique citing actors for each OA category are considered and show consistent advantage of OPEN outputs over CLOSED outputs, i.e., OA outputs attract more unique citing actors, for all years included (Supplementary Figures B). Exceptions or less clear patterns for the median count in terms of subregions, regions and fields of study are due to the broader grouping of citing actors and large number of outputs with low citation counts.

To confirm this finding across the distributions of outputs, we also include the distributional summaries (in the form of boxplots) of samples (i.e., 10,000 outputs from each OA category) drawn independently for each OA category and each publication year (Supplementary Figures C). In these boxplots it is observed that OA outputs are characterised by heavier upper tails (and often with the box shifted upward) when compared to the CLOSED category across all publication years and all types of citing actors. Again, we note caveats around small numbers of groups and large numbers of outputs for certain cases in the study dataset. GREEN outputs stand out as the best performing category in terms of the number of unique citing actors (institutions, countries, subregions, regions).

We then introduce citation diversity measures as per the main part of our overall analysis. For both the Shannon and Gini-Simpson measures we see higher mean and median diversity scores for the OPEN outputs (vs. CLOSED outputs) for every year of publication, with respect to citing institutions, countries, subregions and regions. With respect to citing fields of research there is a slight disadvantage for GOLD outputs in 2010-2011 which turns into an advantage by 2012 (Supplementary Figures D). We also examine the distributions of diversity scores for the samples drawn from each category for each year using boxplots (Supplementary Figures E). In addition to increased central tendency for the OA categories, there are also signs in these boxplots of longer upper tails and shorter lower tails - added indications of the OA citation diversity advantage.

To confirm our findings are not confined to specific percentiles of the data, we also study the kernel density estimates (KDEs) and histograms of the diversity scores, for all combinations of diversity measures, citing actors, and years of publication. The KDEs and histograms are compared between OPEN and CLOSED outputs (for 10,000 outputs drawn from each). The results reveal a highly consistent finding of the OA citation diversity advantage. For all data analysed in these figures, OA outputs result in a distributional shift towards higher diversity scores, lower proportions of outputs with low diversity scores, and increased proportions of outputs that score highly for diversity (Supplementary Figures F).

The OA citation diversity advantage holds for both access via the publishers (i.e., GOLD outputs) as well as for access via other repository platforms (i.e., GREEN outputs) with the latter showing a larger effect. One possible confounding effect is the dominance of Pubmed Central and Europe Pubmed Central as important repositories and the higher average citation counts of biomedical research articles. To address this we examine the citation diversity effect by fields of research of the cited articles and note that the OA citation diversity advantage is highly consistent across all “MAG Level 0” fields for GREEN outputs (Supplementary Figures G). There is substantial variation for GOLD outputs and overall OA performances. We also note large differences in the OA effect between selected fields of research. But for the majority of fields where our dataset has good coverage, the OA citation diversity advantage is clearly seen, including for disciplines distinct from biomedical sciences showing that the effect is robust across natural, biological and clinical sciences, and in several areas of social sciences.

4.2. Relationships between citation diversity and citation count

A criticism of claims for an OA citation advantage is that researchers focus on ensuring that their best work is the most accessible and/or that the advantage is primarily a function of the prestige of the authors and their institutions. One of our goals with the diversity analysis was to use indicators that are less dependent on citation counts as a means of reducing this potentially confounding effect.

With the exception of extreme cases where the citing articles have very many authors, articles with very low citation counts will be limited in the values that the diversity measures can take on. We therefore examined the diversity advantage as a function of citation counts to ensure that the effect was robust to this issue.

We undertake this analysis both at the level of the whole corpus and with a set of consistent sized samples to address the differences in the numbers of OPEN and CLOSED outputs over time. Again, the OA citation diversity advantage is robust across all citation count bins for all years of publication for diversity measures based on citations from different institutions, countries, subregions and regions (with some caveats on the last due to the small number of regions).

First, we revisit how unique numbers of citing actors are counted. To confirm that our earlier observations are robust for outputs that attract different levels of citations, we split outputs from the same year into 14 bins depending on their citation counts (roughly keeping bins similar in population size) and compared the distributions of counts of unique citing affiliations across OPEN and CLOSED outputs for samples drawn (i.e., 2000 OPEN vs 2000 CLOSED outputs) from each citation bin (Supplementary Figures H). Boxplots are presented for OPEN vs CLOSED outputs for each citation group for all years and all types of citing actors. We find that OPEN outputs perform no worse, and in fact better in most cases, than CLOSED outputs in attracting unique numbers of citing actors.

Similarly, we construct the comparison of diversity scores across citation bins for all years and both diversity measures (Supplementary Figures I). It is clear from these results that there is consistency in the OA citation diversity advantage across citation bins for almost all cases considered. The main exceptions are in the earlier years for the fields of research plots. However, these plots indicate a switch from negative to positive effects in more recent years, consistent with our earlier observations for mean and median diversity scores. To further explore the potential relationship between the diversity scores and citation counts, we also calculate the quartiles of diversity scores for the complete data for each year. These are presented as line charts (Supplementary Figures J). These results show a weak relationship between diversity scores and citation counts, but only for low citation count, which is not unexpected given the increasing likelihood of more citing affiliations links. The strength of this weak relationship further weakens for outputs with substantial citations.

In summary we find the OA citation diversity advantage to be not completely driven by the large number of low-citation outputs, nor is it simply an effect of highly cited outputs. Rather, the OA citation diversity advantage is a consistent effect that is seen across the cohort of outputs.

4.3. Citations between subregions and regions

Further to observing an OA citation diversity advantage, it is also important to understand where the increased citation diversity originates. In particular, we need to be able to track how a subregion or region benefits from making its outputs OA (e.g., more citations from others) and also how they benefit from OA outputs of other subregions or regions (e.g., more access to outputs of others). To aid such an analysis, we filter the data down to individual subregions and regions. Then, for a given subregion or region, we determine the numbers of citations to its OPEN and CLOSED outputs from each of the other subregions or regions, respectively. Average citation ratios (i.e., the average number of citations to OA outputs, divided by the average number of citations to non-OA outputs, and times by one hundred) and percentage change in total citations (i.e., total citations to OA outputs minus total citations to non-OA outputs, then divided by total citations to non-OA outputs, and multiplied by one hundred) are calculated for each citing subregion or region. A value above one hundred in the former indicates an OA advantage and a value above 0 for the latter indicates an OA advantage. The results are presented in Supplementary Figures K to N.

For most subregions and regions, we observe an OA advantage for citations coming from other subregions and regions. In particular, there are increased citations to OPEN outputs affiliated to institutions from subregions that are traditionally underrepresented in the literature or have fewer resources, e.g., North Africa, Sub-Saharan Africa, and Latin America and the Caribbean. This is consistent with the increased output usage through greater access from these subregions and regions. However, we also note that the OA citation diversity advantage accrues preferentially to traditionally “prestigious” centres of research in terms of wealth and scale of research outputs. For example, Northern Europe seems to benefit most from both increased citations from other subregions (i.e., high OA advantage is seen for almost all citing subregions to Northern Europe), and for its increased usage of outputs from other subregions (i.e., it is the subregion that is consistently one of the top citing subregions in terms of OA advantage for outputs by other subregions). A similar pattern is observed for North America. There are also signs of changing trends in terms of percentage changes in total citations, where the OA advantage has either increased or shifted from negative to positive in more recent years, for selected subregions or regions.

5. DISCUSSIONS

This article proposes new ways of understanding and evaluating citations in relation to the wider dissemination of research – citation diversity via institutions, countries, subregions, regions, and fields of research. The main purpose of these measures and the corresponding data analyses is to understand the impact of OA on the diversity of users of research outputs. We are also interested in how the level of this impact compares across different geographic regions.

Most previous literature has focused on the OA citation count advantage – i.e., OPEN outputs have higher citation counts than CLOSED outputs. As mentioned in the Introduction, there are many debates as to whether there is a real OA citation advantage. Some confounding factors (Tennant et al., 2016) include author self-selection (i.e., authors choose to make their best articles OA), discipline biases (i.e., potentially significant differences across disciplines), and access to resources and prestige (i.e., well-known authors with more resources are more likely to make their work OA). These imply that the focus on citation counting is not able to paint the full picture of the benefits of OA. An OPEN output may receive more citations, but these citations may continue to come from the same groups of researchers. Conversely, an OPEN output may not have received more citations, but the citations may come from a broader set of research users. Hence, we argue that a shift to understanding the diversity of citations provides a stronger and more meaningful evidence of the benefits of OA in reaching wider audiences.

As the main result, we find that OA is associated with higher citation diversity, i.e., OPEN outputs receive more diverse citations as compared to CLOSED outputs. We refer to this phenomenon as OA citation diversity advantage. We find this advantage to be remarkably consistent across the many ways in which we have analysed the data (bar the very few extreme cases), which addresses concerns of confounding factors mentioned above. GREEN is the best performing OA category in terms of providing the highest citation diversity scores overall. Though we do recognise it is difficult to completely split out the effects of GOLD and GREEN outputs.

We also find that there are differences across subregions and regions in terms of how much they benefit from OA citation diversity advantage. In particular, historically wealthier and larger centres of research seem to benefit more from this effect – having more of others citing their OA work and also citing more of others' OA work. Whether this is a true pattern of "the rich get richer" and what that potentially means for advancing OA advocacy and policy making will be an important area for further research.

The current article extends and generalises from the works of Young & Brandes (2020) and Neylon et al. (2021) and opens the door to much further research. An obvious direction is to expand on the measures of citation diversity both in more complex measures (such as accounting for multiple author affiliation links) and introducing new characterisations of citation diversity (such as language diversity of citing outputs; see Linkov, 2021 and Diprose et al., 2022). Our fields of research data are drawn from MAG which is now discontinued. It would be interesting to examine how our results may change if a different subject classification system (e.g., Web of Science subject classification) or database (e.g., OpenAlex) is used. It would also be interesting to explore how citation diversity relates to the diversity in author collaboration (Naik et al., 2023). Improving data coverage of historically underrepresented geographies, disciplines and non-traditional outputs also continues to be a challenge.

6. CONCLUSIONS

The Budapest OA Initiative (Chan et al., 2002), now over 20 years old, notes that OA makes possible

> "...the world-wide electronic distribution of the peer-reviewed journal literature and completely free and unrestricted access to it by all scientists, scholars, teachers, students, and other curious minds."

providing a public good which will

> "…accelerate research, enrich education, share the learning of the rich with the poor and the poor with the rich, make this literature as useful as it can be, and lay the foundation for uniting humanity in a common intellectual conversation and quest for knowledge".

Efforts to demonstrate the success of this endeavour remain as controversial as the choice of paths towards achieving OA. The use of citations to capture the use and value of research will always be limited, but data on other forms of usage for scholarly publishing remain challenging and incomplete. By shifting attention from counting citations to assessing the diversity of citing outputs we have demonstrated that existing data can be repurposed to analyse different goals. In doing so we have demonstrated that even for the narrow form of usage that citation from research outputs represents, OA outputs are being used by a wider diversity of citing outputs, whether we analyse those citing outputs by institution, country, subregion, region, or fields of research.

More broadly, citation diversity measures offer a new view over existing data, providing potential insights that are not offered by simple citation counts. As a potential insight into where the benefits of OA are being seen and a guide to improving our policy implementation of OA for wider access this approach offers many opportunities in addressing (Chan et al., 2002)

"…the task of removing the barriers to open access and building a future in which research and education in every part of the world are that much more free to flourish".

DATA AVAILABILITY

The processed data (as CSV files) used for the analysis and for generating figures are shared on Zenodo (https://doi.org/10.5281/zenodo.7081118) and GitHub (https://github.com/Curtin-Open-Knowledge-Initiative/citation-diversity).

CODE AVAILABILITY

The SQL queries used to generate all data, together with codes used to produce figures, to perform the analysis, and to generate the final text documents are shared via Zenodo (https://doi.org/10.5281/zenodo.7081118) and GitHub (https://github.com/Curtin-Open-Knowledge-Initiative/citation-diversity).

REFERENCES

Archambault, E., Amyot, D., Deschamps, P., Nicol, A., Provencher, F., Rebout, L., & Roberge, G. Proportion of open access papers published in peer-reviewed journals at the European and world level—1996–2013. RTD-B6-PP-2011-2: Study to develop a set of indicators to measure open access. Report. Science-Metrix (2014). Retrieved August 19, 2022 from https://science-metrix.com/sites/default/files/science-metrix/publications/d_1.8_sm_ec_dg-rtd_proportion_oa_1996-2013_v11p.pdf

Basson, I., Blanckenberg, J. P., & Prozesky, H. Do open access journal articles experience a citation advantage? Results and methodological reflections of an application of multiple measures to an analysis by WoS subject areas. *Scientometrics* 126, 459-484 (2021). https://doi.org/10.1007/s11192-020-03734-9

Bautista-Puig, N., Lopez-Illescas, C., de Moya-Anegon, F., Guerrero-Bote, V., & Moed, H. F. Do journals flipping to gold open access show an OA citation or publication advantage? *Scientometrics* 124, 2551–2575 (2020). https://doi.org/10.1007/s11192-020-03546-x

Chan, L., et al. Read the Declaration - Budapest Open Access Initiative (2002). Retrieved September, 6, 2022 from https://www.budapestopenaccessinitiative.org/read/

Dahler-Larsen, P. Making citations of publications in languages other than English visible: On the feasibility of a PLOTE-index. *Research Evaluation* 27(3), 212-221 (2018) https://doi.org/10.1093/reseval/rvy010

Davis, P. M. Open access, readership, citations: a randomized controlled trial of scientific journal publishing. *The FASEB Journal* 25(7), 2129-2134 (2011). https://doi.org/10.1096/fj.11-183988

Diprose, J., Neylon, C., & Kramer, B. Language Diversity in Scholarly Publishing. Curtin Open Knowledge Initiative News. Retrieved September 5, 2023 from https://openknowledge.community/language-diversity/

Dorta-González, P., González-Betancor, S.M. & Dorta-González, M.I. Reconsidering the gold open access citation advantage postulate in a multidisciplinary context: an analysis of the subject categories in the Web of Science database 2009–2014. *Scientometrics* 112, 877–901 (2017). https://doi.org/10.1007/s11192-017-2422-y

Hosking, R., Diprose, J. P., Roelofs, A., Chien, T-Y., Handcock, R. N., Kramer, B., Napier, K., Montgomery, L., & Neylon, C. Academic Observatory Workflows [Software]. Zenodo (2022). https://doi.org/10.5281/zenodo.6366694

Hua, F., Sun, H., Walsh, T., Worthington, H., & Glenny, A. Open access to journal articles in dentistry: Prevalence and citation impact. *Journal of Dentistry* 47, 41-48 (2016). https://doi.org/10.1016/j.jdent.2016.02.005

Huang, C-K., & Neylon, C. Curtin-Open-Knowledge-Initiative/citation-diversity: Codes and Data for Open Access Research Outputs Receive More Diverse Citations [Software]. Zenodo (2022). https://doi.org/10.5281/zenodo.7081118

Hutchins, B. I., Yuan, X., Anderson, J. M., & Santangelo, G. M. Relative Citation Ratio (RCR): A New Metric That Uses Citation Rates to Measure Influence at the Article Level. *PLoS Biology* 14(9), e1002541 (2016). https://doi.org/10.1371/journal.pbio.1002541

Lewis, C. L. The open access citation advantage: Does it exist and what does it mean for libraries? *Information Technology and Libraries* 37(3), 50-65 (2018). https://doi.org/10.6017/ital.v37i3.10604

Linkov, V., O'Doherty, K., Choi, E., & Han, G. Linguistic Diversity Index: A Scientometric Measure to Enhance the Relevance of Small and Minority Group Languages. *SAGE Open* 11(2), 1-9 (2021). https://doi.org/10.1177/21582440211009191

Naik, C., Sugimoto, C. R., Larivière, V., Leng, C., & Guo, W. Impact of geographic diversity on citation of collaborative research. *Quantitative Science Studies* 4 (2): 442–465. (2023) https://doi.org/10.1162/qss_a_00248

Neylon, C., & Huang, C-K. The Global State of Open Access 2021. Zenodo (2022). https://doi.org/10.5281/zenodo.7059176

Neylon, C., Ozaygen, A., Montgomery, L., Huang, C-K., Pyne, R., Lucraft, M., & Emery, C. More Readers in More Places: The Benefits of Open Access for Scholarly Books. *Insights* 34 (1): 27 (2021). http://doi.org/10.1629/uksg.558

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). https://doi.org/10.7717/peerj.4375

Seppänen, J-T., Värri, H, & Ylönen, I. Co-citation Percentile Rank and JYUcite: a new network-standardized output-level citation influence metric and its implementation using Dimensions API. *Scientometrics* 127, 3523-3541. (2022). https://doi.org/10.1007/s11192-022-04393-8

Sotudeh, H. Does open access citation advantage depend on paper topics? *Journal of Information Science* 46(5), 696-709. (2020). https://doi.org/10.1177/0165551519865489

Tennant, J. P., Waldner, F., Jacques, D. C., Masuzzo, P., Collister, L. B., & Hartgerink, C. H. The academic, economic and societal impacts of Open Access: an evidence-based review. *F1000 Research*, 5: 632. (2016). https://doi.org/10.12688%2Ff1000research.8460.3

The World Bank. Research and development expenditure (% of GDP). World Bank Group (2022). Retrieved September 6, 2022 from https://data.worldbank.org/indicator/GB.XPD.RSDV.GD.ZS

Wang, K., Shen, Z., Huang, C., Wu, C., Dong, Y., & Kanakia, A. Microsoft Academic Graph: When experts are not enough. *Quantitative Science Studies* 1(1), 396-413. (2020). https://doi.org/10.1162/qss_a_00021

Young, J. S., & Brandes, P. M. Green and gold open access citation and interdisciplinary advantage: A bibliometric study of two science journals. *The Journal of Academic Librarianship* 46(2), 102105. (2020). https://doi.org/10.1016/j.acalib.2019.102105

Zhang, L., & Watson, E. M. Measuring the Impact of Gold and Green Open Access. *The Journal of Academic Librarianship* 43(4), 337-345. (2017). https://doi.org/10.1016/j.acalib.2017.06.004