# Open Access Research Outputs Receive More Diverse Citations

CURTIN OPEN KNOWLEDGE INITIATIVE

DATE: 21 SEPTEMBER 2022

AUTHORS: 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.[1,2] A higher citation count also does not necessarily imply wider usage such as citations by authors from more places.[3,4,5] 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.

# Main

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).[6] 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.[7]

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.[8,9,10] However, there are several confounding factors that weaken claims of a causal link between OA and enhanced use of research outputs.[2,11] A set of narrowly defined randomised control trials finds no effect, 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.[1,12,13,14,15]

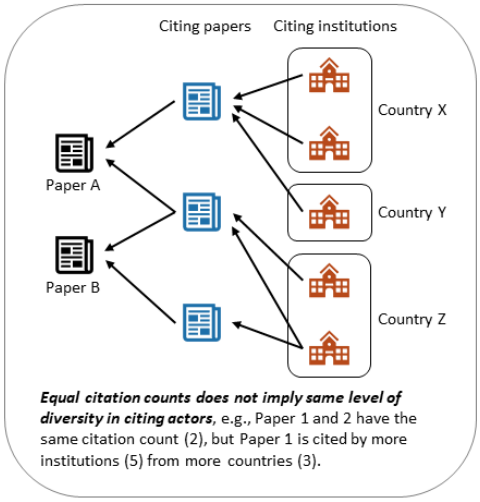
In addition, we feel that the focus on citation counts fails to address the core goals of OA, specifically that a wider range of research users has more access. We need a different approach to quantify the impact of OA focusing on widening the diversity of users who are able to access scholarly content. 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 are able to analyse the fields of research across citing outputs. We refer to these measures under one umbrella term: *citation diversity*. 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[1,12,13,14,15] in analyses based simply on citation counts which remain with more sophisticated measures such as citation velocity, as shown in previous research.[16,17]

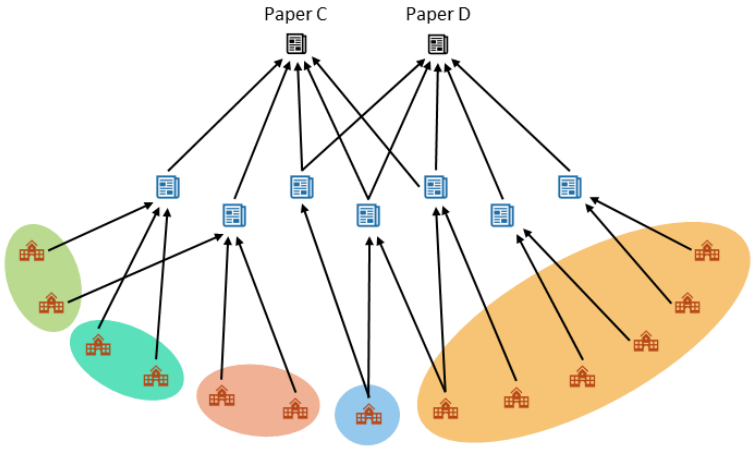
To analyse citation diversity we used the data workflows and datasets developed by the Curtin Open Knowledge Initiative for analysis of open knowledge performance.[18] For the current analysis we used an integrated dataset that combines data from Crossref Metadata, Microsoft Academic Graph (MAG) and the Research Organization Registry (ROR) to provide information on affiliations and geographical locations, fields of research, and publication dates. We used data from Unpaywall to define OA status of individual outputs.

For our analysis we extracted all relevant research outputs with publication years from 2010 to 2019 (see Methods 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 groups). Fig. 1 demonstrates how citation diversity assessed using these indices is different from traditional citation counts. Two outputs having equal citation counts does not imply they have the same level of diversity in citing groups such as citations from a number of different institutions.

---

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




| Paper | # 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 paper 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 Methods for details of calculating the Shannon Entropy (or Shannon Index) and the Gini-Simpson Index (or Gini's Diversity Index).

---

# Comparing OA categories

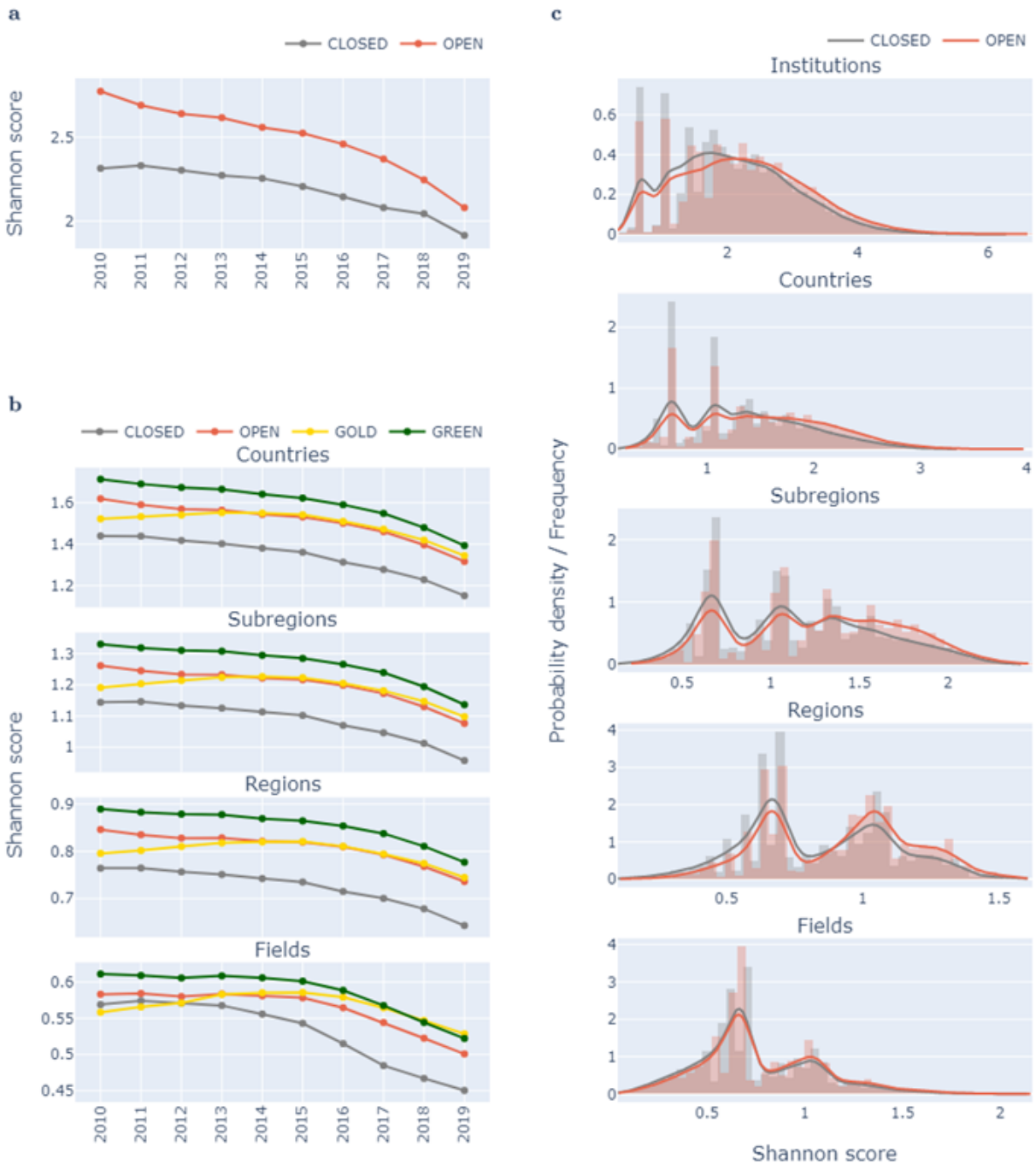
For the purpose of this study, we consider four different (but potentially overlapping) categories of OA:

- OPEN - freely accessible via either publisher platforms or OA repositories;
- GOLD - freely accessible via publisher platforms with open licence;
- GREEN - OA via disciplinary and institutional repositories;
- CLOSED - OA copy not available from publisher platforms nor open repositories.

In our analysis, 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. 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 characterised 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 Methods for details).

**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. **c.** Kernel density estimation (KDE) of Shannon scores are provided for various types of grouping and compared between OA and closed outputs. Clusters at zero are removed for increased readability but are presented in the Supplementary Figures. The first peaks from the left are due to the high number of outputs with low number of citations. A decrease in proportion of lower scores (including the cluster at zero not shown) and a right-shift pattern is observed for OA outputs as compared to their closed counterparts.

In analysing scores for the diversity measures, our results showed an enhanced diversity of citing institutions, countries, subregions, and regions for OA research outputs, with this effect being consistent for all publication years since 2010 (see Fig. 2), and across almost all fields of research in our study data (see Methods). 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 of these potential groupings (see Methods for a comprehensive analysis). This includes distributional shifts toward higher diversity scores for OA outputs (relative to closed outputs) for all citing groups, publication years, and both diversity measures.

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 OA) than for OA provided via publisher websites. This effect shows interesting discipline and author-country effects which merit further investigation.

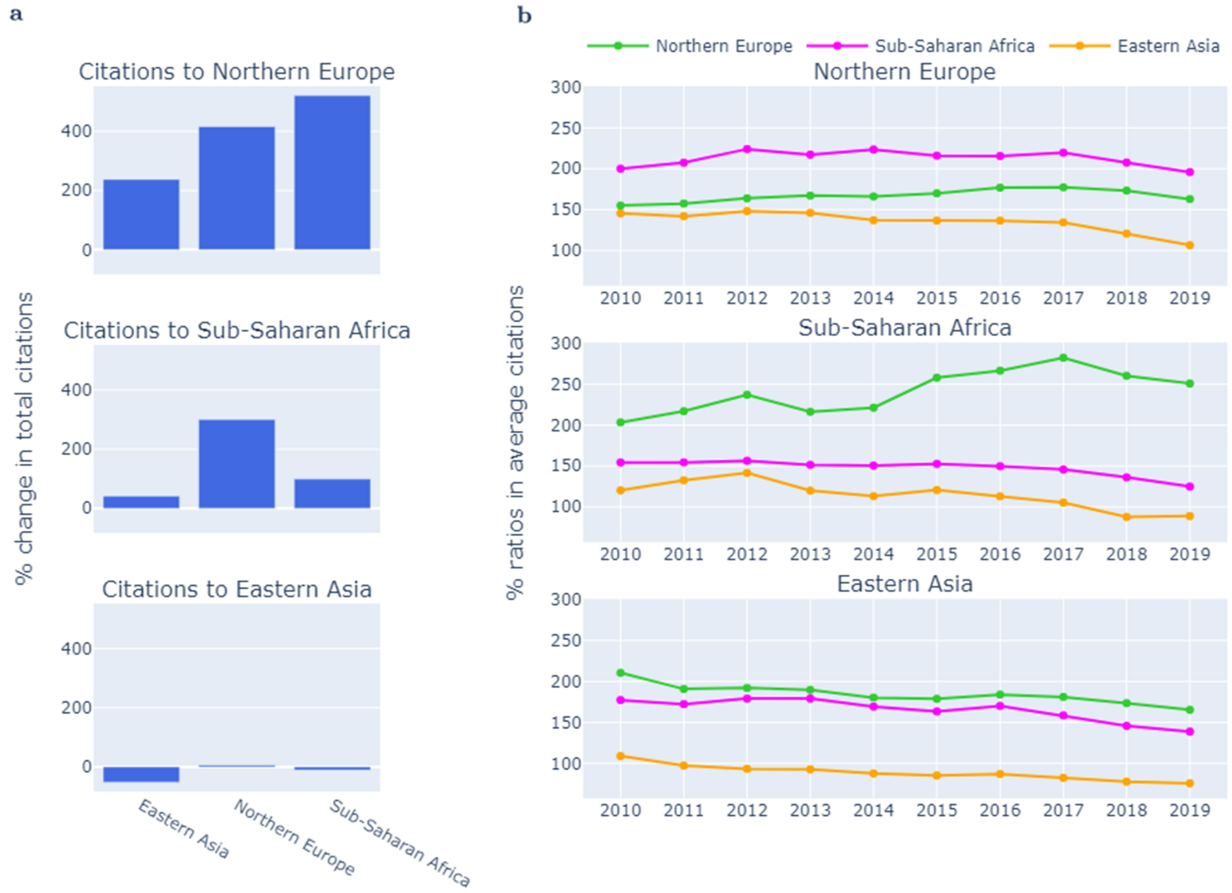
The debate over the citation count advantage is dominated by questions of confounding effects, specifically whether OA is more accessible to researchers from wealthier and more prestigious insitutitions and/or whether researchers selectively make their best work OA. To address this we showed that the citation diversity advantage is present, independent of citation counts (see Methods). The lack of overall correlation between citation counts and citation diversity provides evidence that citation counts and citation diversity track different aspects of usage and that there is limited common confounding at the global scale.

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.

# Comparing geographies

To further understand where increased citation diversity comes from and how it compares across subregions, we also examined the geographical distribution of citations. 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) .[19] 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. However, the citation diversity advantage also accrues preferentially to traditionally prestigious centres of research. For example, Fig. 3 shows differences between open and closed outputs with respect to citations to and from three selected subregions. Outputs within affiliations from Northern Europe benefit most from both increased citations to its open outputs (i.e., usage by other subregions), and for the increased citation of Northern European outputs to other subregions (i.e., 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 Sub-Saharan Africa (similarly for Latin America, not shown), but show an increase over time from low or negative levels (see Methods).

**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 Methods) 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 open 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.

# Discussion

The Budapest OA Initiative,[6] 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[6]

> "…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".

# Methods

## Input data

The COKI Academic Observatory data collection pipeline[18] 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), MAG (institutional affiliations, citation links, fields of research), ROR (institutional information) to generate the "DOI Table" - an enriched metadata source on research outputs.

These datasets are updated on a regular cycle with MAG updated fortnightly 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.

## 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.

We define R as the number of groups and p as the proportion of citing affiliation links assigned to a given group, or as the proportion of citing outputs assigned to a given field of research. The Shannon Entropy quantifies the level of uncertainty in predicting the group assignment of a randomly selected citing affiliation link or citing output to field as:

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

Whereas, the Gini-Simpson Index measures the probability that two randomly selected citing affiliation/field links belongs to the same group:

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

with ln as the natural logarithm.

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.[20]

# A Consistent Effect Across All Measures and Categories

We first 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 (Supplementary Figures A).

We then turn to the analysis of the citing affiliation links. We start by examining counts of unique citing affiliations grouped by institutions, countries, subregions, and regions, and also examining the numbers of unique fields of research of the citing outputs. 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 groups for each OA category are considered and show consistent advantage of OA outputs over closed outputs, i.e., OA outputs attract more unique citing groups, 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 affiliation links and large number of outputs with low counts.

To confirm this finding across the distributions of outputs, we also include the distribution summaries (in the form of box plots) 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 box plots 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 groups. Again, we note caveats around small numbers of groups and large numbers of outputs for certain cases in the study dataset. Green OA stands out as the best performing category in terms of the number of unique citing groups (institutions, countries, subregions, regions).

We then introduce citation diversity measures as the main part of our analysis. For both the Shannon and Gini-Simpson measures we see higher mean and median diversity scores for the OA outputs (vs. closed outputs) for every year of publication, with respect to citing institutions, countries, subregions and regions. With respect to citing fields there is a slight disadvantage for Gold OA 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 box plots (Supplementary Figures E). In addition to increased central tendency for the OA categories, there are also signs in these box plots 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 groups, and years of publications. The KDEs and histograms are compared between OA 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 OA) as well as for access via other repository platforms (i.e., Green OA) 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 OA (Supplementary Figures G). There is substantial variation for Gold OA 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.

# Dependence of Diversity Measures on Citation Counts

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 articles 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 groups 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 OA 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 OA vs closed outputs for each citation group for all years and all types of citing groups. We find that OA outputs perform no worse, and in fact better in most cases, than closed outputs in attracting unique numbers of citing groups.

Similarly, we construct the comparison of diversity scores across citation bins for 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 not be 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.

# 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 OA and closed outputs from all 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 OA 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.

# 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. Secondly, 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 reliable as a measure. Thirdly, 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

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).

# References


1. 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
2. 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
3. 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
4. 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
5. 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
6. Chan, L., et al. Read the Declaration - Budapest Open Access Initiative (2002). Retrieved September, 6, 2022 from https://www.budapestopenaccessinitiative.org/read/
7. Neylon, C., & Huang, C-K. The Global State of Open Access 2021. Zenodo (2022). https://doi.org/10.5281/zenodo.7059176
8. 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
9. 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.
10. 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
11. 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
12. 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

13. 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
14. 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
15. 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
16. 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
17. 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
18. 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
19. 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
20. 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

# 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).

# Acknowledgements

This work was funded by the Research Office of Curtin through a strategic grant, the Curtin University Faculty of Humanities, and the School of Media, Creative Arts and Social Inquiry. The Curtin Open Knowledge Initiative is also recipient of a grant from Arcadia a Charitable Fund of Lisbet Rausing & Peter Baldwin.

# Author Information

Chun-Kai (Karl) Huang (Centre for Culture and Technology, Curtin University, Bentley, WA 6102, Australia)

Cameron Neylon (Centre for Culture and Technology, Curtin University, Bentley, WA 6102, Australia)

Lucy Montgomery (Centre for Culture and Technology, Curtin University, Bentley, WA 6102, Australia)

Richard Hosking (Centre for Culture and Technology, Curtin University, Bentley, WA 6102, Australia)

James P. Diprose (Centre for Culture and Technology, Curtin University, Bentley, WA 6102, Australia)

Rebecca N. Handcock (Curtin Institute for Computation, Curtin University, Bentley, WA 6102, Australia)

Katie Wilson (Faculty of Education, Te Herenga Waka - Victoria University of Wellington, Wellington 6012, New Zealand)

### Contributions

C.K.H., C.N., and L.M. were involved with the conceptualization and project administration. C.K.H., C.N., and R.H. were responsible for data curation. C.K.H. and C.N. were responsible for formal analysis, investigation, methodology, validation, visualisation and writing the original draft of the article. C.N. and L.M. were responsible for funding acquisition and supervision. C.K.H. and C.N. were responsible for the software that produced the processed data, analysis, figures, and the final text documents. R.H. and J.P.D. were responsible for the software that collected and created data tables used as input for the project. C.K.H., C.N., L.M., R.H., J.P.D., R.N.H. and K.W. contributed to the final review, editing and approval of the manuscript.

### Corresponding author

Correspondence to Chun-Kai (Karl) Huang at karl.huang@curtin.edu.au.

## Ethics declaration

The authors declare that they have no competing interests.