ISSI 2023 Proceedings, v2, pp 515–519
DOI: 10.5281/zenodo.8432306 https://cns-iu.github.io/workshops/2023-07-02_issi/

Open access science in Wikipedia

Puyu Yang[1] and Giovanni Colavizza[2]

[1] *p.yang2@uva.nl,* [2] *g.colavizza@uva.nl*
University of Amsterdam, Institute for Logic, Language and Computation (ILLC), LAB42, 1098XH, Amsterdam, (The Netherlands)

Abstract
As one of the largest online encyclopedias, Wikipedia is essential in conveying scientific information. To ensure the reliability of its information, Wikipedia relies on a variety of sources, including scientific articles. In this study, we assess the Open Access (OA) status of scientific articles cited in Wikipedia. To reach this goal, we mainly use the Wikipedia Citations dataset and equip it with Open Access status data from OpenAlex and journal information from Scimago. We find that 42% of all citations to scientific articles in Wikipedia are to Open Access articles, and the distribution of different Open Access types cited by Wikipedia is similar to that observed in academic literature: Bronze, Green, Gold, Hybrid. The scientific disciplines of biology, physics and mathematics have a high proportion of OA articles cited in Wikipedia, while history is the lowest. Our results provide a preliminary picture of how scientific articles shaped Wikipedia from the perspective of Open Access.

Introduction
As one of the most popular knowledge acquisition platforms, Wikipedia is relied on by millions of users every day to meet a wide range of information needs (Singer et al., 2017), and its influence on academic research is also increasing year by year (Park, 2011). One of Wikipedia's core principles is neutrality in writing verifiable articles and trusted sources[1], which also makes citations one of the cornerstones of Wikipedia. Among all the citations, scientific articles are considered the ideal source. Previous research has shown that scientific resources play a critical and specialized role in Wikipedia, especially supporting scientific articles such as biology and medicine (Yang & Colavizza, 2022). Besides, journal articles cited from Wikipedia are more likely published in high-impact journals (e.g., by impact factor) and in open access (Nielsen, 2007; Teplitskiy et al., 2017).

However, it cannot be ignored that high-impact journals often require expensive subscription fees (Björk & Solomon, 2012). For this reason, many researchers advocate journals that provide free access to research—"Open Access" (OA) journals (Van Norden, 2013). As the OA movement has grown, OA journals have significantly impacted the dissemination of scientific knowledge outside the literature (Piwowar et al., 2018). As a platform for free content, OA research is a core pillar of Wikipedia. It is designed to be accessible to the public and its large community of volunteer editors. However, few studies reveal at article-level granularity the role of OA articles in Wikipedia. A better understanding of this issue helps advance our understanding of how the open science agenda contributes to public commons such as Wikipedia.

Data and methods
Our workflow proceeds as follows: firstly, we use the Wikipedia Citations open dataset to get all citations from Wikipedia to any source. Furthermore, we rely on Wikipedia Citation's own classification and DOI to find journal articles. Next, we equip these journal articles with OA information via OpenAlex API. Lastly, we use data from Scimago to equip each journal with related information. The main datasets we use are further described in what follows.

[1] https://en.wikipedia.org/wiki/Wikipedia:Core_content_policies

Wikipedia Citations
Wikipedia Citations is the main dataset in this research (Singh et al., 2021). It consists of more than 29M citations extracted from the over 6M articles composing the English Wikipedia as of May 2020. In Wikipedia Citations, approximately 2.5M citations are classified as journal articles, of which 1,705,085 are equipped with a DOI.

OpenAlex and Scimago
In order to study the effect of OA articles, we choose OpenAlex to equip each article with OA status, concepts, publisher, times cited and other related information. To get the journal information, we download the data from Scimago and equip each journal with SJR score, H index and so on.

Results
With the additional information from OpenAlex and Scimago, we equip 1,696,108 (99.4%) out of 1,705,085 valid citations with available metadata, of which 1,152,141 (99.5%) are unique DOIs with OA status. Among this, 716,278 (42.3%) citations are OA out of the 1,696,108 total citations (450,277 (39.1%) is OA out of the 1,152,141 unique DOIs), which is about 1.5 times higher than the distribution among all the scientific articles (Piwowar et al., 2018).

We present our first findings on the distribution of Open Access (OA) policies in Wikipedia citations in figure 1. Our results demonstrate that the most commonly observed OA policy in Wikipedia citations is the bronze policy, which is again consistent with trends in scholarly literature (Piwowar et al., 2018). The second most popular OA policy observed in Wikipedia citations is Green, which is significantly more prevalent than the Gold policy. This trend may be attributed to differences in the reference acquisition methods used by Wikipedia editors compared to researchers. Additionally, our comparison of OA policies in scientific articles and Wikipedia reveals a similar trend (Piwowar et al., 2018). Interestingly, the percentage of Green policy in Wikipedia is much higher than in scientific articles, suggesting that OA repositories are a valuable source of information for Wikipedia editors.

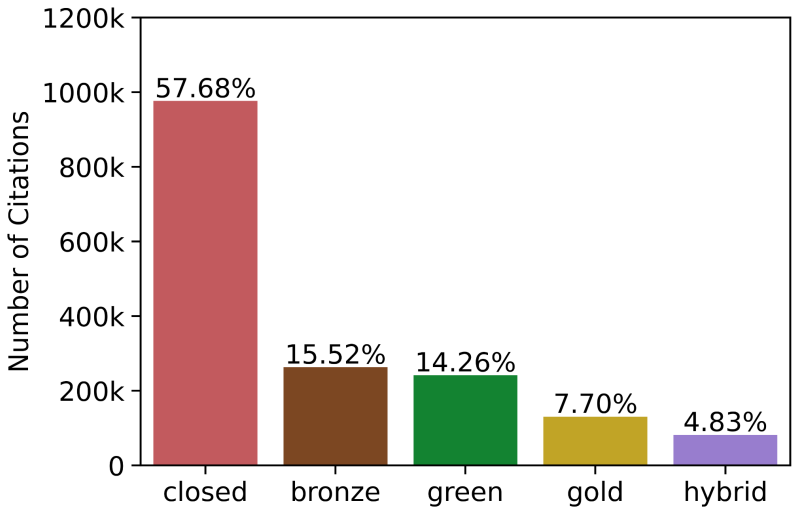

Figure 1. OA citations by policy

Then, we continue to observe the breakdown of OA status and OA policy by journals. In our dataset, we finally have 40,191 journals. To visualize it easily, we calculate the number of citations in each journal and list only the top 20 journals here. As is found in previous studies,

some high-impact journals appear more frequently on Wikipedia (Nielsen, 2007), *Nature*, *PNAS* and *Science* take up 5.7% of all the citations. From figure 2, we can find that most of the journals provide OA articles, and within journals, there is a high variance in whether articles are Open Access or not such as in *Nature* or *Science*, so it is misleading to characterize each journal as "Open" or "Closed". From figure 3, we can see again the bronze policy appears a lot, and in some OA journals like the *Journal of Biological Chemistry* and *PLOS ONE*, even if both of them are OA, the policy is totally different, one is hybrid and the other is gold.

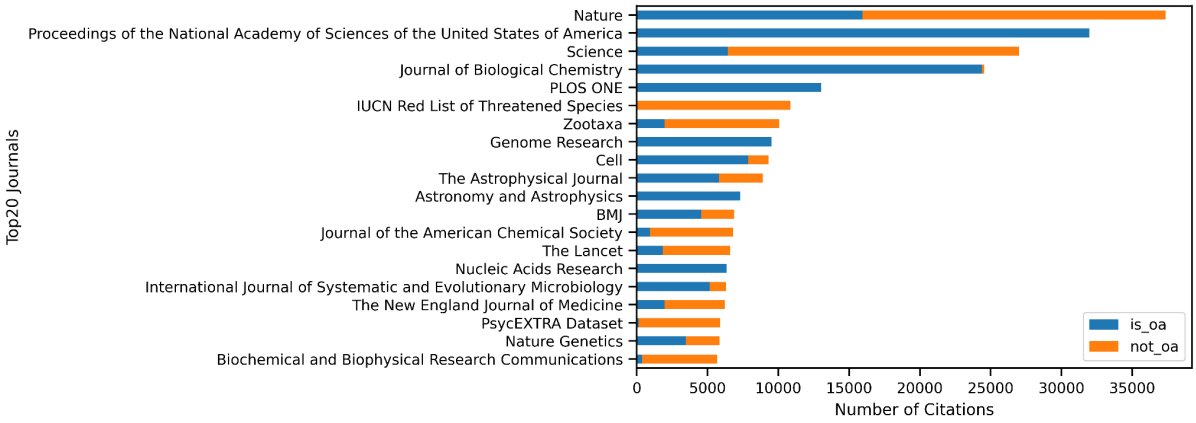

Figure 2. OA citations by top 20 journals

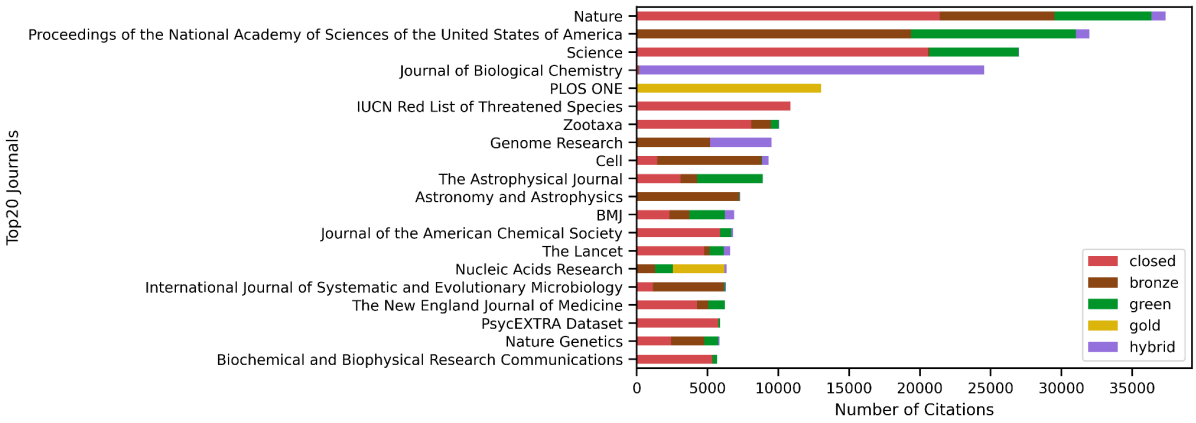

Figure 3. OA policies by top 20 journals

Similarly, we show the breakdown of OA status by concepts in figure 4 and 5. OpenAlex has 65k concepts[2] and 19 root-level concepts, here we only consider the root-level concepts. We use fractional counting to calculate the number of citations of each root-level concept, and as 42.3% of all citations on Wikipedia are OA, we treat this as a baseline. In figure 4, we plot the percentage of citations OA status on the left and the black dot line means the baseline (42.3%), on the right part we plot the total number of citations of concepts and order these concepts from large to small. Obviously, only four concepts have a higher proportion than the baseline: Biology, Medicine, Physics and Mathematics. On the contrary, History, Art, Psychology and Political Science have the lowest proportion of OA status. In general, if a field of research has more OA articles cited from Wikipedia, then it may mean that the associated Wikipedia articles may be accessible to a wider audience of editors and users. This effect would give some motivation to further study whether the Wikipedia articles which cite a higher number of OA

[2]https://docs.openalex.org/about-the-data/concept

articles are indeed more popular to users, and so pushing authors and publishers to release more studies under Open Access. For the OA articles in each concept, they are mainly based on bronze and green.

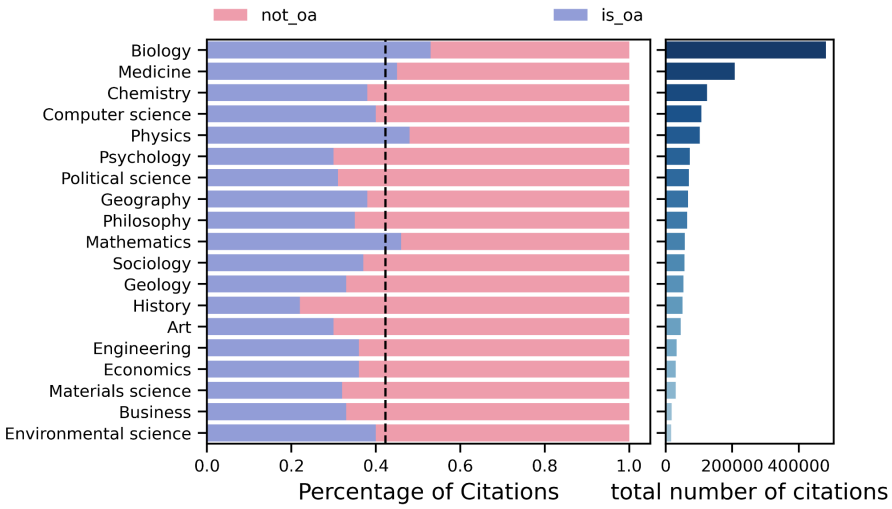

Figure 4. OA status by concept

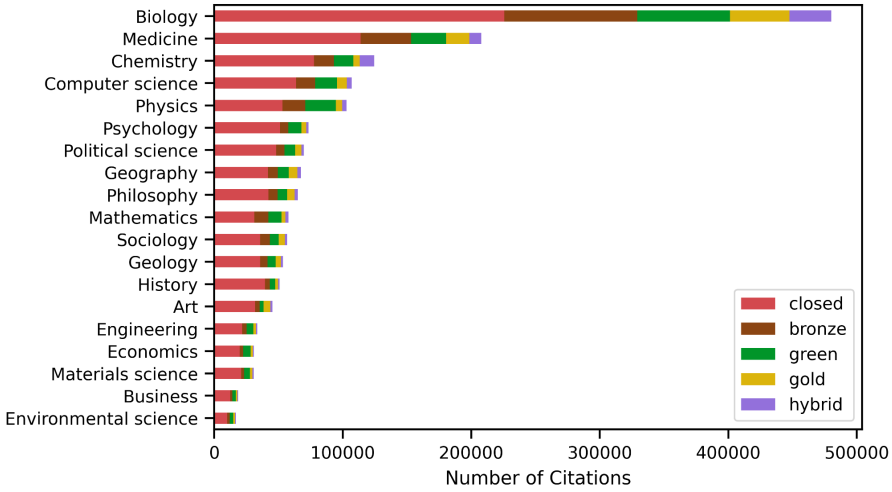

Figure 5. OA policies by concept

To observe the change in OA status by years, we plot figure 6. In the figure, the blue line means the proportion of OA status of each publication year and it uses the left y-axis, the black line represents the proportion of citations published in each year and it uses the right y-axis. It can be found that during the past 40 years, the increase in citing new OA articles has been steadily rising, i.e., new articles that are cited are increasingly OA rather than closed access, so OA articles will be paving the way for how science on Wikipedia is shaped in the future. Besides, the number of citations peaked in 2004 at 4.2%.

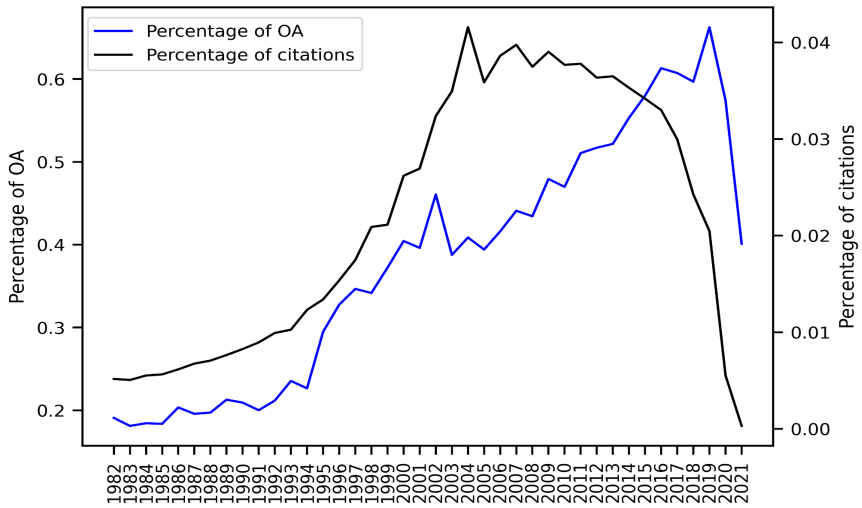

Figure 6. OA status by publication year

Conclusion

We have studied the relationship between Open Access (OA) articles and Wikipedia by journal, concepts and publication year. Taken together, Wikipedia extensively uses scientific articles from high-impact journals and the trend of citing OA articles is increasing over time. Moreover, the OA articles distribution has a noticeable difference within concepts. Specifically, STEM-related concepts receive more citations and have a higher OA share.

Analyzing citations from the perspective of Open Access also shows a more nuanced picture. On the one hand, the increasing trend of citing OA articles in Wikipedia is a confirmation of the positive impact of open science. On the other hand, some research fields, for example, in the humanities, show a gap in OA citations. Our study has a set of limitations, including those coming from our data sources and a preliminary focus on root-level concepts of analysis. We suggest that future work might profitably focus on lower, more granular levels of analysis, as well as on the effect of dissemination by these OA articles.

References

Björk, B. C., & Solomon, D. (2012). Open access versus subscription journals: a comparison of scientific impact. *BMC medicine*, 10(1), 1-10.

Nielsen, F. Å. (2007). Scientific citations in Wikipedia. *arXiv preprint arXiv:0705.2106*.

Park, T. K. (2011). The visibility of Wikipedia in scholarly publications. *First Monday*.

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., ... & Haustein, S. (2018). The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ*, 6, e4375.

Singer, P., Lemmerich, F., West, R., Zia, L., Wulczyn, E., Strohmaier, M., & Leskovec, J. (2017). Why We Read Wikipedia. *Proceedings of the 26th International Conference on World Wide Web*, 1591–1600.

Singh, H., West, R., & Colavizza, G. (2021). Wikipedia citations: A comprehensive data set of citations with identifiers extracted from English Wikipedia. *Quantitative Science Studies*, 2(1), 1–19.

Teplitskiy, M., Lu, G., & Duede, E. (2017). Amplifying the impact of open access: Wikipedia and the diffusion of science. *Journal of the Association for Information Science and Technology*, 68(9), 2116-2127.

Van Noorden, R. (2013). Open access: The true cost of science publishing. *Nature*, 495(7442), 426–429.

Yang, P., & Colavizza, G. (2022). A Map of Science in Wikipedia. *In Companion Proceedings of the Web Conference 2022* (pp. 1289-1300).