Scientometrics
https://doi.org/10.1007/s11192-024-05163-4

Open access improves the dissemination of science: insights from Wikipedia

Puyu Yang[1] · **Ahad Shoaib**[2] · **Robert West**[3] · **Giovanni Colavizza**[4]

Received: 8 May 2024 / Accepted: 18 September 2024

Abstract
Wikipedia is a well-known platform for disseminating knowledge, and scientific sources, such as journal articles, play a critical role in supporting its mission. The open access movement aims to make scientific knowledge openly available, and we might intuitively expect open access to help further Wikipedia's mission. However, the extent of this relationship remains largely unknown. To fill this gap, we analyse a large dataset of citations from the English Wikipedia and model the role of open access in Wikipedia's citation patterns. Our findings reveal that Wikipedia relies on open access articles at a higher overall rate (44.1%) compared to their availability in the Web of Science (23.6%) and OpenAlex (22.6%). Furthermore, both the accessibility (open access status) and academic impact (citation count) significantly increase the probability of an article being cited on Wikipedia. Specifically, open access articles are extensively and increasingly more cited in Wikipedia, as they show an approximately 64.7% higher likelihood of being cited in Wikipedia when compared to paywalled articles, after controlling for confounding factors. This open access citation effect is particularly strong for articles with high citation counts or published in recent years. Our findings highlight the pivotal role of open access in facilitating the dissemination of scientific knowledge, thereby increasing the likelihood of open access articles reaching a more diverse audience through platforms such as Wikipedia. Simultaneously, open access articles contribute to the reliability of Wikipedia as a source by affording editors timely access to novel results.

Keywords Wikipedia · Open access · Open science · Science communication

✉ Puyu Yang
p.yang2@uva.nl

[1] Institute for Logic, Language and Computation (ILLC), University of Amsterdam, Amsterdam 1098XH, The Netherlands

[2] University of Waterloo, Waterloo, Canada

[3] School of Computer and Communication Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

[4] Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy

Published online: 15 October 2024

Introduction

Open access (OA) publishing has emerged as a popular alternative to traditional subscription-based models, with the goal of making research more widely accessible to the public. This movement has gained momentum over the years, with many scholars recognizing the benefits of open access in promoting the dissemination of scientific knowledge and funding bodies adopting OA mandates (Piwowar et al., 2018; Holmberg et al., 2020).

Citations play a crucial role in supporting Wikipedia's mission to provide reliable and verifiable information.[1] Among various sources, academic and peer-reviewed publications are widely regarded as the most reliable.[2] The OA movement provides Wikipedia with a valuable opportunity to access a vast repository of reliable and verifiable scientific knowledge. By incorporating OA citations, Wikipedia can enhance its role in scientific communication. Tattersall et al. (2022). As a dynamic platform for sharing and disseminating knowledge across the globe, Wikipedia is relied upon by millions of users every day to satisfy a wide range of information needs (Singer et al., 2017). It has become a critical source of information for both the general public and academic researchers, and its impact is extending beyond the realm of general knowledge and into the academic sphere (Park, 2011; Kousha & Thelwall, 2017; Tohidinasab & Jamali, 2013).

Wikipedia's extensive use of citations makes it possible to analyze its reliance on academic publications, which is a central aspect of our investigation. Previous research utilized the Scopus database and an English Wikipedia database dump extracted from 2014, culminating in the identification of 32,361 unique articles for analysis. They found that articles from OA journals exhibit 47% higher odds of being cited in Wikipedia compared to those from paywalled journals (Teplitskiy et al., 2017). Notably, their adoption of journals as the unit of analysis, rather than individual articles, has the possible drawback of wrongly estimating the influence of OA on scientific knowledge dissemination through Wikipedia. This limitation also arises from overlooking articles accessible via green or hybrid routes (ElSabry, 2017). Moreover, their manual matching approach imposed constraints on the scale of their research. Therefore, an exploration conducted at the granularity of individual articles not only promises a more nuanced understanding of the relationship between OA and the dissemination of scientific knowledge through Wikipedia but also unveils the role of citation count in this process.

In light of this, our study seeks to fill this gap by examining how OA publications affect Wikipedia at the article-level granularity. Specifically, we aim to answer the following research questions:

1. RQ1: To what extent does Wikipedia rely on open access publications? How has this been changing over time?
2. RQ2: To what extent does the open access status of an article influence its likelihood of being cited in Wikipedia?

To address these questions, we will use descriptive statistics and regression analysis based on the *Wikipedia Citations* dataset (Singh et al., 2020). To identify the information in article-level granularity such as the OA status of publications, we will use the OpenAlex and

[1] https://en.wikipedia.org/wiki/Wikipedia:Core_content_policies.
[2] https://en.wikipedia.org/wiki/Wikipedia:Verifiability#Reliable_sources.

Scimago data. Our research contributes to understanding the role of OA in the dissemination of scientific knowledge and the impact of Wikipedia in this process, as well as informing policy and practice in the realm of open scholarly communication.

The remainder of the paper is structured as follows. "Previous work" Sect provides an overview of existing research in the field. "Data and methods" Sect. describes our dataset and methodology. "Results" Sect. presents descriptive statistics of OA publications in Wikipedia (RQ1), and then uses regression analysis to model the influence of OA status on the likelihood of a paper being cited in Wikipedia (RQ2). Finally, "Discussion" and "Conclusion" Sects. offer a discussion and conclusion of our findings.

Previous work

Open access in science

The key idea behind OA is to provide unrestricted and free access to scientific outcomes, thus enhancing their visibility and reach regardless of financial or geographical constraints (Tennant et al., 2016; Redalyc et al., 2003). The increasing popularity of OA in academic publications has generated extensive discussions among scholars in recent years. Empirical studies have shown that OA has had a significantly positive impact on the accessibility of scientific journal articles (Björk et al., 2010). However, the distribution of OA publications varies depending on the data source. A comprehensive analysis of OA publications based on Crossref data shows that at least 27.9% of the total 19 million scientific articles are OA (Piwowar et al., 2018). In contrast, studies report that around 55% of articles in the Google Scholar database from 2009 to 2014 are OA, and more than 50% of scientific papers published since 2007 can be accessed freely (Archambault et al., 2014; Martín-Martín, Costas, van Leeuwen, & Delgado López-Cózar, 2018). Among the various OA policies, Bronze OA is the most common type (Piwowar et al., 2018). Although the distribution of OA varies across different fields, General Science, Technology, and Biomedical research have relatively higher OA rates, while Engineering and Arts &Humanities have lower rates (Archambault et al., 2014; Martín-Martín et al., 2018).

An "open access citation advantage" (OACA) has also been a topic of ongoing debate. Some researchers have observed that a citation advantage linked to OA exists, although the effect magnitude varies based on the dataset and methods used. For example, OA articles have been found to receive 18% more citations than average based on Web of Science, while Scopus reports an even higher, positive 40% effect (Piwowar et al., 2018; Archambault et al., 2014). Kristin found that in four disciplines-philosophy, political science, electrical and electronic engineering, and mathematics, OA articles exhibit a greater research impact (Antelman, 2004). Distinct advantages are found for green OA articles hosted in institutional repositories, receiving 106% more citations than gold OA or non-OA articles, and OA articles receive up to 36% more diverse, interdisciplinary citations than non-OA articles (Young & Brandes, 2020). Despite these findings, a recent systematic review of OACA suggests that the debate continues, revealing diverse outcomes across different studies (Langham-Putrow et al., 2021). Out of 134 included studies, 47.8% confirm the existence of OACA, 27.6% deny it, 23.9% find OACA only in subsets, and 0.8% are inconclusive, with a notable association between the focus on multiple disciplines and the

identification of OACA in subsets (Langham-Putrow et al., 2021). Therefore, the effects of OA on citation patterns remain a topic of interest and active investigation.

Science and Wikipedia

With the rapid development of the internet, traditional peer review processes need to adapt to keep pace with the rapid knowledge creation in the 21st century (Black, 2008). As the largest encyclopedia worldwide, Wikipedia aims to effectively and globally distribute information based on scientific findings,[3] thereby making it a valuable altmetric source (Sugimoto et al., 2017; Mesgari et al., 2015). Evans and Krauthammer observed higher citation counts for articles linked in Wikipedia, suggesting its potential for impact assessment (Evans & Krauthammer, 2011). Altmetric.com integrated Wikipedia mentions into its tracking in 2015[4], but doubts have arisen about Wikipedia's reliability for impact assessment. Lin and Fenner found that only 4% of PLOS articles were cited in Wikipedia (Lin & Fenner, 2014), and Kousha and Thelwall concluded that Wikipedia citations are insufficient for impact assessment in most fields (Kousha & Thelwall, 2017).

Previous research indicates that Wikipedia's topical coverage is similar to that of scientific disciplines. With 13.44% of its citations coming from OA journals (Arroyo-Machado et al., 2020) and 31.2% of Wikipedia citations associated with OA sources, this percentage has exhibited an upward trend over the years (Pooladian & Borrego, 2017). Additionally, STEM fields, particularly biology and medicine, comprise the most prominently featured scientific topics on Wikipedia (Yang & Colavizza, 2022). Fields such as medicine and psychology have a comparatively high number of citations to research papers on Wikipedia and are sometimes used as a gateway to further academic research (Maggio et al., 2017; Schweitzer, 2008). Furthermore, journal articles cited in Wikipedia tend to be published in high-impact journals (e.g., with higher impact factors) and are more frequently OA than the average article (Nielsen, 2007; Teplitskiy et al., 2017).

Science significantly contributes to Wikipedia, but the influence is reciprocal. Previous studies have established that Wikipedia can enhance the citation impact of the articles it cites (Thompson & Hanley, 2018). Furthermore, Wikipedia has demonstrated its ability to rapidly and reliably incorporate novel scientific findings in response to ongoing public events or crises (Colavizza, 2020).

Citation analyses of Wikipedia

The open release of citation datasets from Wikipedia has led to a surge in studies examining citation analysis on Wikipedia (Singh et al., 2020; Zagorova et al., 2021). Among the articles on Wikipedia, 6.7% cite at least one journal article with an associated digital object identifier (DOI) (Singh et al., 2020), and the majority of these cited journal articles were published in the past two decades (Yang & Colavizza, 2022). Benjakob et al. (2022) conducted a study on the quality of citations in Wikipedia during COVID-19 and found that Wikipedia mostly cites reliable sources and prefers OA articles. Some researchers have focused on user behavior related to reference usage on Wikipedia. Piccardi et al. (2020)

[3] https://wikimediafoundation.org/about/mission/.

[4] https://www.altmetric.com/blog/new-source-alert-wikipedia/.

found that engagement with citations on Wikipedia is generally low, but references are more frequently looked up when the information is not included.

Despite the growing number of citation studies on Wikipedia, the relationship between OA and Wikipedia still requires further exploration. Previous research has examined the effect of OA on Wikipedia, and found that OA articles were 47% more likely to be cited than paywalled articles when controlling for journal and research fields (Teplitskiy et al., 2017). However, their focus on analyzing journals rather than individual articles leads to in an underestimation of OA impact on disseminating scientific knowledge through Wikipedia. This limitation stems from the oversight of articles accessible through green or hybrid routes (ElSabry, 2017). Additionally, their manual matching approach constrained the scope of the research. Consequently, this study aims to build on previous findings by employing a more rigorous and comprehensive methodology, examining individual articles, and accounting for additional confounding factors to better understand the relationship between OA and Wikipedia.

Data and methods

The data collection process adhered to the workflow outlined below. First, we obtained all citations from English Wikipedia to any source using the open dataset called *Wikipedia Citations* (Kokash & Colavizza, 2024). Next, to identify journal articles, we used the classification and DOI information provided by *Wikipedia Citations*. Then, to enrich the journal articles with article-level data such as citation counts, OA status, and OA policy, we used the OpenAlex API to retrieve relevant information through DOIs for each journal article. Finally, we used data from Scimago to obtain relevant information for each journal. The following sections provide a detailed description of the main datasets used in the study.

Wikipedia citations

The primary dataset used in this research is *Wikipedia Citations*, a comprehensive dataset of over 45 million citations extracted from the February 2024 dump of English Wikipedia (Kokash & Colavizza, 2024). This is an updated version of the 2020 dataset (Singh et al., 2020). Of these, approximately 2.2 million citations are classified as journal articles, with 2,197,461 of them containing a DOI.

OpenAlex and Scimago

To examine the impact of OA articles, we used OpenAlex, a free and open platform providing data on academic papers and researchers (Priem et al., 2022). OpenAlex draws data from sources such as Microsoft Academic Service (MAG) and Crossref, containing more than 240 million academic works. These works are useful for research in bibliometrics, science and technology studies, and science of science policy (Bredahl, 2022; Hao et al., 2022). To obtain the necessary data for journal articles in *Wikipedia Citations*, we utilized the OpenAlex API[5] to retrieve relevant article details such as OA status, OA policy,

[5] https://docs.openalex.org/how-to-use-the-api/get-single-entities.

publication date, publisher, and concepts, among others, for each DOI. After matching, we retrieved article information from OpenAlex for 2,154,524 journal articles.

In our paper, we used OA policy following the classification scheme proposed by OpenAlex, which includes the following categories:

1. Gold: Published in a fully OA journal.
2. Green: Toll-access on the publisher landing page, but there is a free copy in an OA repository.
3. Hybrid: Free under an open license in a toll-access journal.
4. Bronze: Free to read on the publisher landing page, but without any identifiable license.
5. Closed: All other articles.

OA status is treated as a binary variable in our analysis, defined as either True or False. According to OpenAlex, and as supported by previous literature (Piwowar et al., 2018), an article is considered OA if it has a URL where the full text can be read without payment or login.

Additionally, we collected journal information to conduct a regression analysis on the influence of OA. We obtained this information using data downloaded from Scimago.[6] Scimago is an OA resource that provides an internationally recognized journal rank indicator for analysis in the fields of scientometrics and informetrics (Falagas et al., 2008; Yuen, 2018; González-Pereira et al., 2010). We equipped each journal with the SCImago Journal Rank indicator (SJR) and other relevant information.

Model specification

Dependent variable To assess the potential advantage of OA articles in Wikipedia, we defined a binary dependent variable, ***is_wiki***, which indicates whether an article has been cited in Wikipedia or not. Since our primary dataset consists solely of articles cited in Wikipedia, we use OpenAlex to obtain negative samples of articles not cited in Wikipedia, via stratified sampling.

Independent variable To assess the impact of OA articles on their citation rates in Wikipedia, we analyze two types of variables: article-level and journal-level. At the article level, we consider the number of citations (***times_cited***), whether the article is OA (***is_oa***), the time of publication (***article_age***), and the field of research (***concept***). These features have been shown to influence citation impact in previous studies (Colavizza et al., 2020; Gargouri et al., 2010; Yegros-Yegros et al., 2015; Struck et al., 2018; Teplitskiy et al., 2017; Nielsen, 2007). At the journal level, we primarily consider the Scimago Journal Rank (***SJR***). To accurately represent the SJR for each article, we use the rank assigned to journals for the same year in which the article was published. Since a year-by-year breakdown of journal ranks is only available from 1999 to 2020, we assign the rank for 1999 to articles published before 1999, as it is the earliest available representation. This range (i.e., published before 1999) accounts for 29% of the citations in our curated set from Wikipedia.

[6] https://www.scimagojr.com/aboutus.php.

Although these variables have been widely used to model citation impact in previous studies, few analyses have directly linked these indicators to whether an article is cited in Wikipedia, particularly concerning different OA policies.

In this study, we use logistic regression to analyze the relationship between a binary dependent variable and one or more independent variables. The logistic regression coefficients represent the size of each predictor variable's contribution to the target variable. Figure 1 illustrates the assumed causal structure of Wikipedia's OA citation adoption effect, with a black line depicting an assumed causal relationship between two variables. Specifically, we assume that the likelihood of a journal article being cited in Wikipedia is directly influenced by its features, citation counts, and OA status. At the same time, OA status can also influence citation counts, leading to a mediated effect on the article's adoption in Wikipedia. Our models measure both the direct effect and the total effect of OA status on being cited in Wikipedia. The direct effect is shown as a thick black line in Fig. 1, while the total effect includes both direct and mediated (via citation counts) effects.

Dataset construction

We aim to create a balanced dataset of journal articles suitable for regression analysis. The initial dataset, sourced from Wikipedia, contained 1,499,021 unique scientific articles. To initiate our regression analysis, we constructed a dataset by adding Journal, Year of Publication, and Concept as stratifying variables. This dataset will be used to account for the influence of ***concept***. To restrict our focus to root-level concepts and avoid ambiguity, we filtered the citations to include only those with a single associated concept, resulting in a set of 410,573 articles. Subsequently, we assembled corresponding sets of articles for these two datasets from OpenAlex based on the stratifying variables, excluding those already cited in Wikipedia. To reduce noise in the sampling strategy, we removed journals with no corresponding name in Scimago and those with fewer than 20 citations. We also removed all articles published before 1900 to eliminate sparsely mentioned dates and accept a slight recency bias.

After pre-processing, we group the articles within each stratum and proceed as follows:

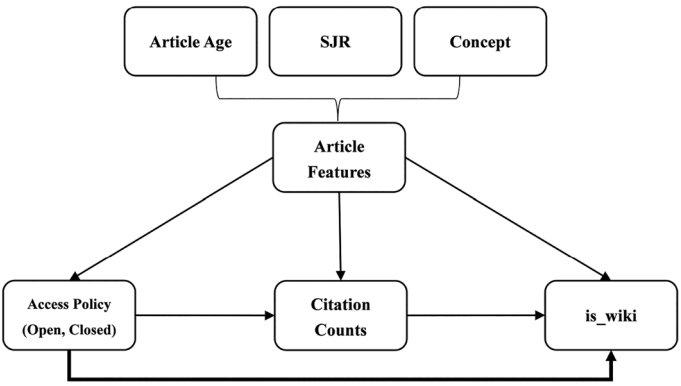


Fig. 1 Assumed causal structure of Wikipedia's OA citation adoption effect, with a black line representing an assumed causal relationship between two variables

1. Filter the entire set of OpenAlex articles to include only those matching the fields in the strata.
2. If the number of articles in this filtered set is fewer than in the curated Wikipedia dataset, discard the strata and remove the corresponding articles from the curated dataset.
3. Otherwise, randomly sample an equal number of articles from the filtered set and add them to the set of negative samples.

After iterating through all strata (90,019 in total), we derived a final negative set comprising 261,230 entries. When combined with the corresponding sets of Wikipedia-cited articles, this results in a comprehensive dataset totalling 522,460 entries.

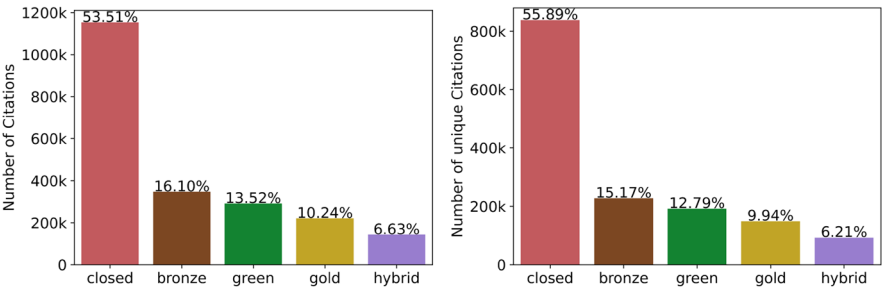
To ensure the robustness of our sampling methodology, we repeated the process five times, resulting in five different datasets that were used in the analyses. Although our method of matching strata to construct a set of negative samples approximates the more rigorous method of propensity score matching (PSM), the discrete nature of our strata and the large population size contribute to the robustness of our analysis. A descriptive overview of this curated dataset is provided in Tables 4, 5 and 6 in the appendix.

Results

We augmented our analysis by incorporating additional metadata from OpenAlex and Scimago, enabling us to obtain information for 98.0% (2,154,524) of the 2,197,461 citations with a valid DOI. From these, we extracted 1,499,021 unique publications (DOIs) and their associated OA status. Our findings show that 46.5% (1,021,820 out of 2,197,461) of the citations and 44.1% (661,068 out of 1,499,021) of the publications were OA.

Characterizing open access articles within Wikipedia

We present our findings on the distribution of OA policy in Wikipedia citations in Fig. 2. Our results show that the most commonly observed OA policy in Wikipedia citations is the bronze policy, which is aligns with trends in scholarly literature (Piwowar et al., 2018). The second most common OA policy observed in Wikipedia citations is green, which is significantly more prevalent than the gold policy.



(a) Over citations. (b) Over unique journal articles.

Fig. 2 Distribution of open access policy in Wikipedia

Figure 3 illustrates the distribution of OA articles based on the publication year. The grey bars represent the total number of articles cited by Wikipedia each year, encompassing both OA and non-OA articles. The green bars indicate the number of OA articles within each grey bar. The red dotted line indicates the annual proportion of OA articles from the Web of Science database, the blue dashed line represents the annual OA article ratio from the OpenAlex database, and the black line shows the proportion of OA articles cited in Wikipedia by publication year. All three lines are plotted against the right y-axis, representing the fraction of OA articles, while the left y-axis shows the total article count.

Overall, the data reveal a consistent increase in the proportion of OA articles cited by Wikipedia over the past four decades. Compared to the overall scientific literature, as measured by the OpenAlex and Web of Science databases, Wikipedia's proportion of OA articles is notably higher. This trend suggests a growing reliance on OA articles within Wikipedia, indicating their significant influence on scientific representation on the platform. Specifically, since 2015, the percentage of OA article citations in Wikipedia has consistently exceeded 50%.

We examined the breakdown of OA status and OA policies across the 40,806 journals in our dataset. To effectively visualize this information, we calculated the number of citations for each journal and selected the top 20 for further analysis. Figure 8 displays the total number of citations for the top 20 journals, where blue represents OA articles and orange represents non-OA articles. Consistent with previous studies, high-impact journals such as *Nature*, *PNAS*, and *Science* appear frequently on Wikipedia (Nielsen, 2007), accounting for 5.7% of all citations. However, inferring the OA status of articles based on whether journals are classified as "Open Access" or "Closed Access" can be misleading (Teplitskiy et al., 2017), due to the high variance in OA status among articles within the same journal. For instance, although some articles in *Nature* and *Science* are OA, others are non-OA.. Therefore, studying the relationship between OA and Wikipedia at the journal level is inappropriate.

To further explore the distribution of OA policies among the top 20 journals, we visualized the data in Fig. 9. Our analysis shows a growing trend towards bronze OA policies among journals. However, some journals that classify themselves as OA, such

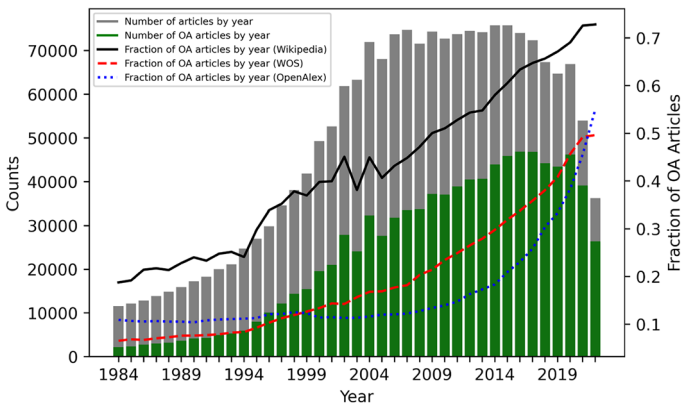


Fig. 3 Fraction of OA citations by publication date of citation

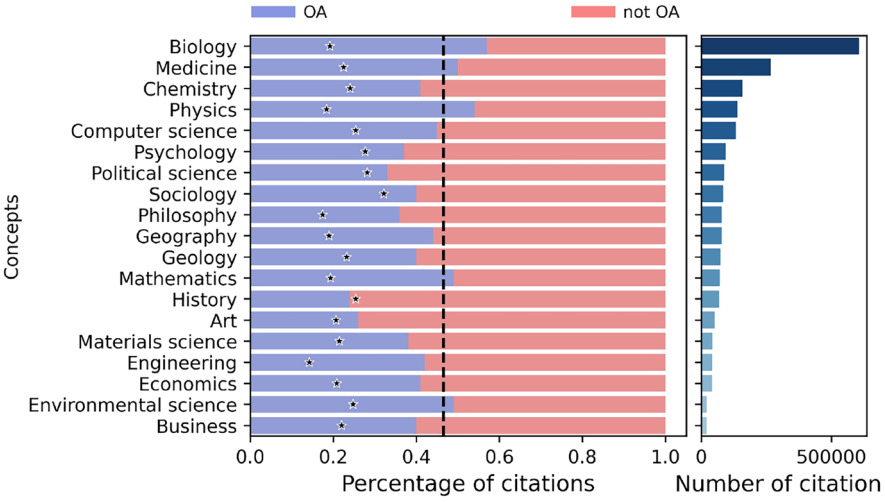


Fig. 4 Distribution of OA status and count of citations by OpenAlex concept

as "*Journal of Biological Chemistry*",[7] contain a significant proportion of articles classified as Hybrid or Gold OA. Despite potential limitations in OpenAlex's classification of OA articles, we adhered to their classifications in our study.

Additionally, we analyzed the distribution of OA status across OpenAlex concepts, as shown in Figs. 4 and 7. The analysis utilized the OpenAlex dataset, which comprises 65,000 concepts, including 19 root-level concepts. We employed fractional counting to assess the number of citations for each root-level concept. In Fig. 4, the left side displays the percentage of cited publications with OA status for each concept, while the right side shows the total number of citations per concept, ordered from the largest to the smallest. The blue bars represent the fraction of OA citations within each OpenAlex concept, while the red bars represent the fraction of paywalled articles. Given that 46.5% of citations on Wikipedia are OA, we used this percentage as a baseline for OA proportionality, represented by the black dotted line in Fig. 4. Additionally, the black star in the same figure denotes the percentage of OA articles for each concept across the entire OpenAlex dataset, serving as a benchmark for the broader scientific landscape. Our analysis revealed significant variance in OA proportions across different fields.

Notably, the OA proportions for all concepts in Wikipedia significantly exceed those observed in OpenAlex, underscoring the critical role of OA articles in shaping Wikipedia's information sources. Although the overall proportion of OA on Wikipedia is 46.5%, certain concepts have a relatively higher OA citation rate. Specifically, Biology (57%), Physics (53%), Medicine (50%), Environmental Science (49%), and Mathematics (49%) demonstrate a greater reliance on OA publications for shaping their scientific content. Conversely, Political Science (33%), Art (27%), and History (24%) show the lowest proportions of OA articles among Wikipedia's cited sources. Generally, Wikipedia exhibits a stronger dependence on OA articles in STEM-related fields than in the humanities, where citations of scientific articles are less prevalent.

[7] https://www.elsevier.com/journals/journal-of-biological-chemistry/0021-9258/open-access-journal.

Table 1 Regression results for the first sample with models 1 and 2

Regression model	Model 1 ($R^2 = 0.00034$)			Model 2 ($R^2 = 0.07182$)		
Feature	Coef	Odds ratios	P>z	Coef	Odds ratios	P>z
Intercept	− 0.363	0.695	0	− 0.979	0.376	0
ln1p_times_cited				0.442	1.557	0
ln(article_age)	0.064	1.066	0	− 0.067	0.935	0
ln(SJR)	− 0.002	0.998	0.487	− 0.246	0.782	0
is_oa	0.588	1.800	0	0.499	1.647	0
is_oa:ln_article_age	− 0.091	0.900	0	− 0.103	0.902	0
Art	− 0.001	0.999	0.980	0.669	1.952	0
Business	0.001	1.001	0.992	0.38	1.462	0
Chemistry	− 0.002	0.998	0.845	0.014	1.014	0.237
Computer science	− 0.004	0.996	0.846	0.251	1.285	0
Economics	− 0.006	0.994	0.885	0.042	1.043	0.387
Engineering	0.000	1.000	0.999	0.523	1.688	0
Environmental science	0.002	1.002	0.987	0.75	2.118	0
Geography	− 0.007	0.993	0.877	0.583	1.791	0
Geology	0.008	1.008	0.626	0.09	1.095	0
History	− 0.004	0.996	0.914	0.646	1.908	0
Materials science	0.009	1.009	0.822	− 0.069	0.934	0.109
Mathematics	0.002	1.002	0.931	0.415	1.514	0
Medicine	0.003	1.003	0.675	0.03	1.03	0
Philosophy	− 0.003	0.997	0.915	0.703	2.02	0
Physics	− 0.012	0.988	0.386	0.227	1.255	0
Political science	− 0.006	0.994	0.851	0.681	1.977	0
Psychology	− 0.002	0.998	0.905	− 0.069	0.934	0
Sociology	0.002	1.002	0.981	0.466	1.594	0

OA citation advantage

To comprehensively assess the impact of an article's OA status on its likelihood of being cited by Wikipedia, we developed a series of statistical models utilizing the datasets outlined in the data and methods section. The objective of this analysis is to elucidate the role of OA articles within the scientific discourse of Wikipedia, and identify any potential advantages associated with their citation patterns.

Model results

We use logistic regression for its interpretability and expressiveness, and we apply log transformations to continuous variables. To thoroughly evaluate the overall impact of OA status on citation adoption, our primary logistic regression model, designed to examine the influence of is_{oa}, is formulated as follows:

$$is\_wiki = is\_oa + ln(article\_age) + ln(SJR) + concept + is\_oa * ln\_article\_age \quad (1)$$

To assess the direct influence of OA status on the likelihood of being cited in Wikipedia, while accounting for the interplay between citation count and OA status, we introduce the second formula:

$$\begin{aligned} is\_wiki = {} & is\_oa + ln(article\_age) + ln(SJR) + concept + ln(times\_cited + 1) \\ & + is\_oa * ln\_article\_age \end{aligned} \tag{2}$$

To validate the robustness of our model, we evaluated the statistical significance of each coefficient across all five samples. A coefficient is considered statistically insignificant if it lacks significance in at least one of the samples. We then present the effects in terms of odds ratios, calculated from the mean odds ratios across all five samples. Additionally, we conducted a multicollinearity check on the variables in the model and found that all Variance Inflation Factors (VIF) were below 10, indicating the absence of significant multicollinearity issues.

Our analysis of the results is summarized in Table 1, which lists the regression results of model 1 and model 2. From Table 1, it is evident that OA articles exhibit substantially higher odds of being cited in Wikipedia compared to paywalled articles. Specifically, in model 1, OA articles have 80% higher odds of being cited, and this becomes 64.7% in model 2 when considering citation counts. These findings highlight that the OA status of an article plays a crucial role in its likelihood of being used as a reference on Wikipedia, suggesting that Wikipedia is more inclined to cite OA articles over paywalled ones.

Incorporating citation counts into the model enhances the interpretability and reveals additional insights. Similar to OA status, citation counts play a significantly positive role in the odds of scientific articles being cited in Wikipedia. This suggests that articles with higher citation counts are more likely to be referenced on Wikipedia, reflecting their impact and visibility within the scientific community. In terms of conceptual classification analysis, we use biology, which has the highest citation count on Wikipedia, as our reference point. By incorporating citation counts into our analysis, we identified 14 concepts that significantly influence the likelihood of articles categorized as OA being cited on Wikipedia. Notably, several concepts from the humanities and social sciences, such as Art, History, Philosophy, and Political Science, exhibited notably positive coefficients. This finding reflects the unique characteristics of these domains, known for being low-citation fields (Patience et al., 2017), despite their substantial importance within the Wikipedia ecosystem. Additionally, Environmental Science also demonstrated a high coefficient, likely due to its interdisciplinary nature, which incorporates knowledge from both natural and social sciences.

Furthermore, the age of the article demonstrates a modest yet significantly negative effect on the likelihood of OA articles being cited by Wikipedia. This finding implies that newer publications are more likely to be cited by Wikipedia compared to older articles, indicating a preference for recent and up-to-date scientific content on the platform.

Despite the negative effect of the SJR, insights can be gleaned from the distribution of SJR among Wikipedia citations, as shown in Fig. 10. In Fig. 10, the x-axis represents the SJR value obtained from Scimago, while the y-axis represents the proportion of Wikipedia citations. It is evident that nearly 90% of the cited journals on Wikipedia have an SJR value of less than 10. Additionally, the mean SJR in our dataset is 3.68, with a third quartile of 4.38. This finding further supports our regression results,

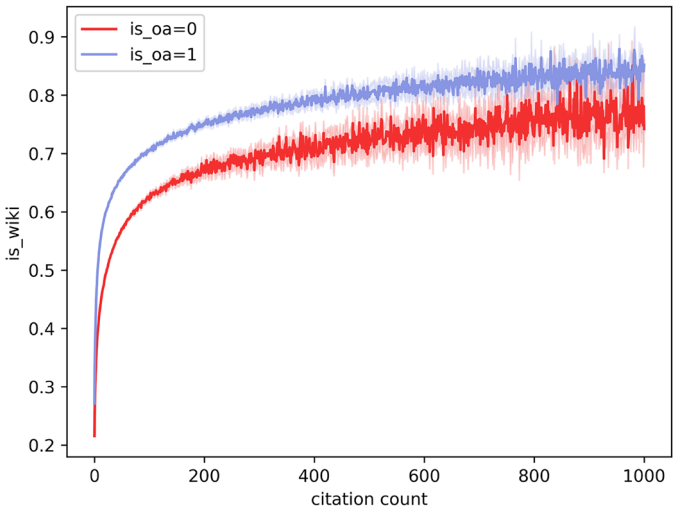


Fig. 5 OA adoption effect at varying citation counts, based on model 2

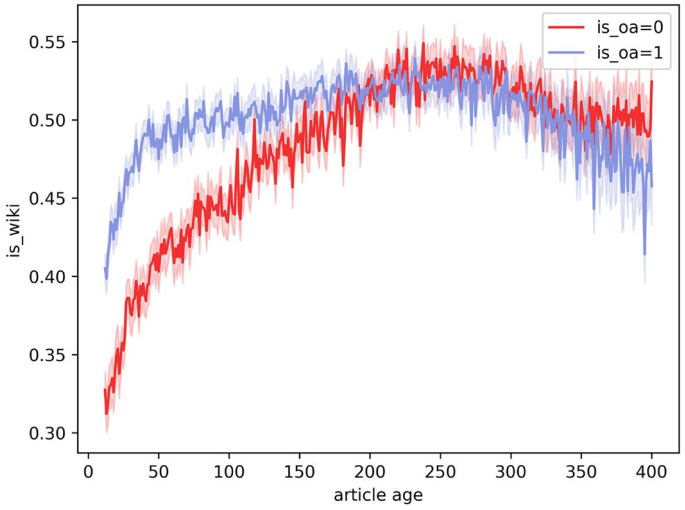


Fig. 6 OA adoption effect at varying article age, based on model 2

indicating that in Wikipedia, most citations come from journals with small SJR values. Thus, as the SJR decreases, the likelihood of OA articles being cited by Wikipedia increases.

To gain insight into the interaction between OA status and citation counts in Wikipedia, we use Formula 2 to create a graph that plots these two variables along with article features.

The graph, shown in Fig. 5, displays the dependent variable, ***is_wiki***, on the y-axis and the citation counts (variable ***times_cited***) on the x-axis. Articles are grouped according to their OA status. We plot the average model prediction for each group using the first data

Table 2 Coefficients for OA adoption by policy

Feature	Coef	Odds_ratios	P>z
Bronze	0.101	1.106	0.000
Gold	0.032	1.032	0.000
Green	0.162	1.176	0.000
Hybrid	0.190	1.210	0.000
ln_article_age	0.019	1.019	0.000
ln(SJR)	− 0.036	0.965	0.000
Bronze:ln_article_age	− 0.0398	0.961	0.000
Gold:ln_article_age	− 0.1392	0.870	0.000
Green:ln_article_age	− 0.1868	0.830	0.000
Hybrid:ln_article_age	− 0.2230	0.800	0.000

Results for the first sample, model 1, $R^2 = 0.0013$

Table 3 Coefficients for OA adoption by policy

Feature	Coef	Odds_ratios	P>z
Bronze	− 0.111	0.895	0.054
Gold	1.090	2.974	0.000
Green	0.834	2.302	0.000
Hybrid	1.156	3.177	0.000
ln_article_age	− 0.070	0.933	0.000
ln(SJR)	− 0.244	0.783	0.000
Bronze:ln(article_age)	0.029	1.030	0.007
Gold:ln_article_age	− 0.245	0.783	0.000
Green:ln_article_age	− 0.171	0.843	0.000
Hybrid:ln_article_age	− 0.274	0.760	0.000
$\ln(1 + \text{times}_{\text{cited}})$	0.447	1.564	0.000

Results for the first sample, model 2, $R^2 = 0.073$

sample and provide 95% bootstrapped confidence intervals for each group (faded color). The red line illustrates the trend of OA adoption by citation count under the condition that the OA status is closed, while the blue line shows the trend under the condition that the OA status is open. This graph reveals several insights. Firstly, when the citation counts are very low (near 0), there is a significant initial citation advantage for OA articles compared to the paywalled articles. As the citation count increases (up to 200), this advantage gradually expands. However, as the citation count continues to grow, this advantage becomes less distinguishable. Our previous work (Yang & Colavizza, 2022) shows that articles cited fewer than 100 times account for 70% of all cited articles, while only about 3% are cited 1,000 times or more. Therefore, most citations in Wikipedia benefit from this OA effect. We speculate that the OA adoption effect arises because Wikipedia editors may find it easier to discover and access open research results earlier in the publication timeline, before these articles accumulate citations and gain broader peer recognition.

In addition, we examined the interaction between OA status and article age on Wikipedia using model 2, as illustrated in Fig. 6. The figure reveals a significant advantage for younger articles, especially those aged less than 48 months (4 years). OA articles in this age group have a 10% higher likelihood of being cited by Wikipedia compared to a

paywalled article. However, as the age of the articles increases to around 240 months (20 years), the likelihood of adoption begins to decline. This trend highlights Wikipedia's preference for newer articles, particularly those published within the last four years, over older publications.

Moreover, we employed two regression models to investigate the impact of OA policy on citation adoption, using "closed" as the baseline. The results, presented in Table 2, demonstrate that all OA policies significantly enhance the overall adoption rate for OA articles. Additionally, the interaction between OA policies and article age shows a significant negative effect on OA adoption. Table 3 presents the results from the second model, which explores the indirect effect of OA policy and reveals a similar trend. However, the bronze policy exhibits a slightly significant negative impact. To validate the robustness of our findings, we conducted regressions across all five samples, with the results reported in Tables 9 and 10.

Discussion

The surge in popularity and growth of OA has significantly contributed to the dissemination of scientific knowledge. Our research highlights Wikipedia's increasing reliance on OA articles, constituting 46.5% of all scientific citations on the platform, a notable rise from the 31.2% reported in the prior study (Pooladian & Borrego, 2017). This trend has shown continuous growth, particularly evident in scientific articles cited by Wikipedia that were published after 2011, where at least 50% are OA. In comparison, only 30% of articles in the Web of Science database and 20% in the OpenAlex database were OA during the same period These findings align with the broader scientific community's trend, as evidenced by the percentage of OA articles steadily increasing to 28% in 2018, with OpenAlex reporting 47% (Piwowar et al., 2018). Despite high-impact journals remaining a preferred source for Wikipedia (Nielsen, 2007), variations in the distribution of OA articles within journals emphasize the necessity for a nuanced, article-level approach.

Our examination of OA policies in scientific articles and Wikipedia unveils a parallel trend (Piwowar et al., 2018). Bronze policy (16.10%) and green policy (13.52%) dominate as the most common OA policies in Wikipedia. The higher prevalence of green policy in Wikipedia compared to scientific articles suggests differences in reference acquisition methods between Wikipedia editors and researchers. This trend further reinforces the importance of not overlooking articles accessible through green routes, thereby avoiding underestimating the impact OA can have on disseminating scientific knowledge through Wikipedia (ElSabry, 2017).

Our study further reveals disparities in OA Wikipedia citations across disciplines, with biology, physics, and mathematics exhibiting higher OA citation rates, while social sciences and humanities show comparatively lower rates. Nevertheless, Wikipedia's robust reliance on OA articles persists across all OpenAlex root concepts.

In addition, our analysis reveals an "OA citation advantage" in Wikipedia, meaning that OA publications are more likely to be cited as references in Wikipedia compared to paywalled publications. Specifically, under similar conditions, OA articles have a 64.7% higher likelihood of being cited in Wikipedia than their paywalled counterparts. Despite the significantly negative effect of the SJR on citation likelihood, we found that over 90% of articles cited in Wikipedia have an SJR below 10, with nearly 80% below 5. This

distribution indicates that Wikipedia editors prioritize the accessibility of reliable sources over the prestige of the journals in which they are published. Furthermore, the likelihood of an OA article being cited increases with its citation count but decreases as the article ages. Wikipedia's editors demonstrate strong responsiveness to new scientific developments, frequently updating content referencing OA articles published within the past four years, reflecting a clear preference for recent and easily accessible scientific knowledge.

We acknowledge certain limitations in our study. First, by focusing exclusively on articles with DOIs, we excluded conference papers and earlier literature. Future research could benefit from including these additional sources. While our regression model accounted for significant factors, such as OA status, OA policy, and citation counts, other causal variables like article length may influence article citations on Wikipedia. Furthermore, our study did not consider time as an analytical dimension, prompting future research to delve into Wikipedia's edit history for specific data at the time of article citation, facilitating a deeper understanding of the causal mechanisms underpinning the interplay between OA and Wikipedia. Additionally, Wikipedia supports dual citations, allowing both paywalled and OA versions of a source to be cited together. This functionality, supported by tools like Wikipedia's OABOT,[8] facilitates the addition of OA links to paywalled citations, improving access without violating copyrights. As a result, some paywalled publications can even be accessed through these OA links. Future studies could more comprehensively explore the impact of OA on Wikipedia by considering this broader context.

Conclusion

This study assessed the impact of open access (OA) on Wikipedia by analyzing article-level features using a comprehensive dataset of Wikipedia citations, OA metrics from OpenAlex, and journal data from Scimago. Our findings reveal that OA articles are increasingly cited over time, with their proportion on Wikipedia significantly exceeding that in the broader scientific literature. Moreover, OA articles enjoy a citation advantage on Wikipedia, with a greater likelihood of being referenced compared to similar paywalled articles. This advantage is particularly pronounced for highly cited articles and those published within the past four years. These results underscore the importance of OA in broadening the dissemination of scientific knowledge, especially on influential platforms like Wikipedia, where newer and more impactful articles are more likely to reach a wider audience.

Our study lays the groundwork for further research on Wikipedia and open science. Future studies should consider a broader range of sources and variables to more fully understand the OA effect on Wikipedia. Additionally, exploring other aspects of open science, such as open research data and software, through similar methodologies could provide further insights. In conclusion, our study highlights the significance of OA in Wikipedia and its potential broader impact, offering a foundation for future research and contributing to the understanding of OA's role in the dissemination of scientific knowledge.

[8] https://en.wikipedia.org/wiki/Wikipedia:OABOT#:%7E:text=Wikipedia%20links%20to%20hundreds%20of,does%20not%20violate%20any%20copyrights.

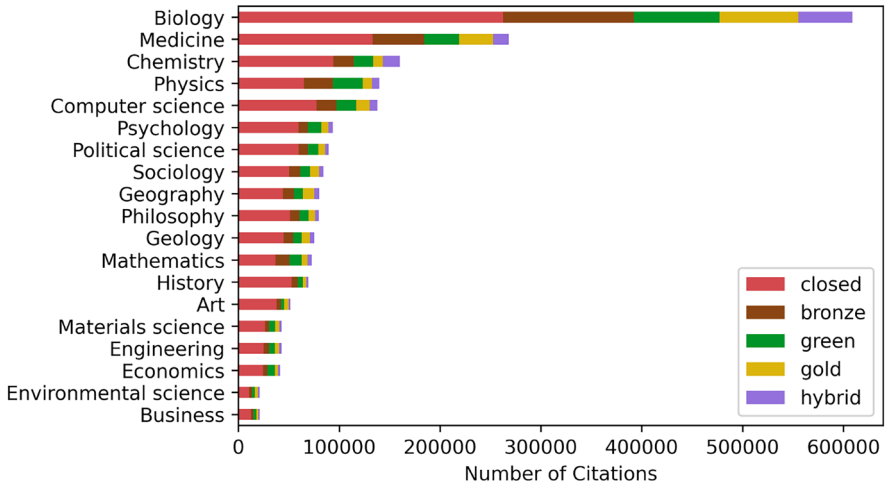


Fig. 7 Distribution of OA policies by OpenAlex concept

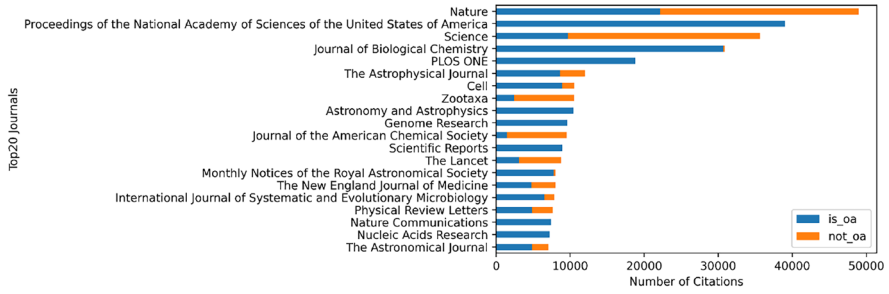


Fig. 8 Distribution of OA status by top 20 journals

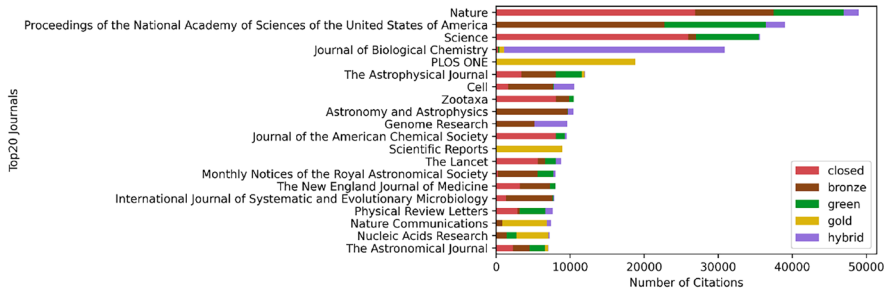


Fig. 9 Distribution of OA policies by top 20 journals

Appendix A: Figures

Presented below are two figures depicting the distribution of OA status and policies among the top 20 journals. We have discussed it in the results part. This observation

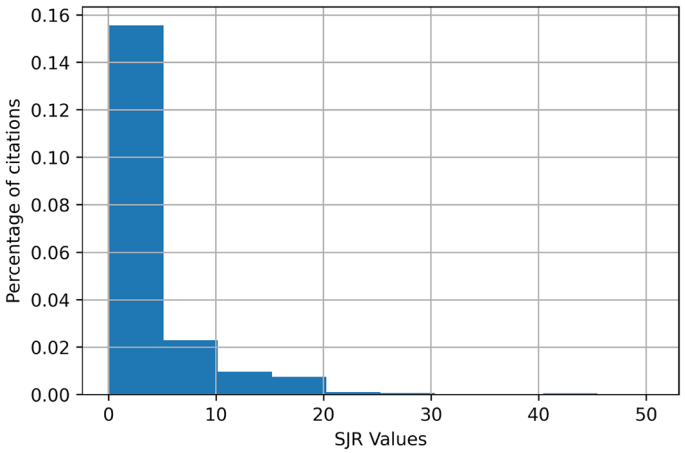


Fig. 10 Distribution of SJR among Wikipedia citations

Table 4 Descriptive statistics for the articles cited in Wikipedia

	Cited_by_count	Num_references	Article_age	H index	is_oa	ln(SJR)
Count	261,230	261,230	261,230	261,230	261,230	261,230
Mean	146.133	37.797	252.171	249.348	0.504	0.632
Std	834.450	49.357	201.939	259.576	0.500	1.166
Min	0	0	12	0	0	− 2.303
25%	14	10	122	87	0	− 0.140
50%	45	28	208	164	1	0.554
75%	124	48	310	305	1	1.460
Max	304,415	1976	1487	1331	1	3.922

Table 5 Descriptive statistics for the articles not cited in Wikipedia. Average over all samples

	Cited_by_count	Num_references	Article_age	H index	is_oa	ln(SJR)
Count	261,230	261,230	261,230	261,230	261,230	261,230
Mean	59.008	28.316	252.137	249.348	0.497	0.632
Std	228.084	38.377	202.060	259.576	0.500	1.166
Min	0	0	12	0	0	− 2.303
25%	3	2	122	87	0	− 0.140
50%	18	21	208	164	0	0.554
75%	54	41	310.2	305	1	1.460
Max	44,132.4	2891.2	1487	1331	1	3.922

Table 6 Count of articles by concepts in the final combined dataset

Num.	Concept	Counts	Num.	Concept	Counts
1	Biology	144,307	11	History	1618
2	Medicine	48,286	12	Art	1589
3	Chemistry	18,135	13	Materials science	1256
4	Physics	10,835	14	Economics	1005
5	Psychology	8491	15	Geography	911
6	Geology	8429	16	Business	480
7	Mathematics	5620	17	Sociology	466
8	Computer science	5274	18	Engineering	241
9	Philosophy	2145	19	Environmental science	211
10	Political science	1931			

Table 7 Coefficients for overall OA adoption

Index	Coef	Odds_ratios	P>z
ln_article_age	0.064	1.066	0.000
ln(SJR)	− 0.002	0.998	0.482
is_oa	0.583	1.791	0.000
is_oa:ln_article_age	− 0.104	0.901	0.000

Average results across all 5 samples, model 1, $R^2 = 0.00032$

Table 8 Coefficients for overall OA adoption

Index	Coef	Odds_ratios	P>z
ln1p_times_cited	0.442	1.556	0.000
ln_article_age	− 0.068	0.934	0.000
ln(SJR)	− 0.245	0.782	0.000
is_oa	0.494	1.639	0.000
is_oa:ln_article_age	− 0.103	0.902	0.000

Average results across all 5 samples, model 2, $R^2 = 0.072$

Table 9 Coefficients for OA adoption by the policy

	Coef	Odds_ratios	P>z
Bronze	0.194	1.215	0.002
Gold	0.672	1.959	0.000
Green	1.147	3.149	0.000
Hybrid	1.092	2.982	0.000
ln_article_age	0.064	1.066	0.000
Bronze:ln_article_age	− 0.031	0.970	0.010
Gold:ln_article_age	− 0.138	0.871	0.000
Green:ln_article_age	− 0.189	0.828	0.000
Hybrid:ln_article_age	− 0.232	0.793	0.000
ln(SJR)	− 0.001	0.999	0.654

Average results across all 5 samples, model 1, $R^2 = 0.0013$

Table 10 Coefficients for OA adoption by the policy

	Coef	Odds_ratios	P>z
Bronze	− 0.158	0.855	0.015
Gold	1.071	2.920	0.000
Green	0.834	2.303	0.000
Hybrid	1.212	3.361	0.000
ln1p_times_cited	0.447	1.564	0.000
ln_article_age	− 0.071	0.932	0.000
Bronze:ln_article_age	0.038	1.039	0.002
Gold:ln_article_age	− 0.241	0.786	0.000
Green:ln_article_age	− 0.173	0.841	0.000
Hybrid:ln_article_age	− 0.284	0.753	0.000
ln(SJR)	− 0.244	0.784	0.000

Average results across all 5 samples, model 2, $R^2 = 0.073$

Table 11 Coefficients for OA adoption by concept for all samples (*is_oa*)

Concept	Min OR	Max OR	OR mean	Highest P-value	Mean R^2
Biology	1.589	1.684	1.621	0.000	0.067
Computer science	1.314	1.737	1.471	0.264	0.052
Chemistry	3.828	4.068	3.918	0.000	0.060
Medicine	2.177	2.390	2.280	0.000	0.134
Psychology	2.524	3.554	3.150	0.001	0.090
Mathematics	1.536	1.748	1.672	0.227	0.073
Economics	2.277	3.635	3.250	0.276	0.104
Geology	1.347	1.531	1.428	0.136	0.053
Sociology	0.937	4.174	2.095	0.967	0.014
History	1.407	2.168	1.768	0.483	0.018
Geography	1.490	2.689	2.066	0.505	0.009
Philosophy	0.400	0.649	0.487	0.301	0.005
Materials science	1.511	3.379	2.253	0.497	0.098
Art	0.783	1.175	0.961	0.902	0.008
Environmental science	0.132	0.551	0.418	0.647	0.008
Physics	2.163	2.559	2.318	0.000	0.064
Engineering	0.863	1.965	1.313	0.911	0.006
Business	2.241	3.556	3.101	0.380	0.032
Political science	0.602	0.805	0.732	0.617	0.017

underscores the significance of conducting an article-level analysis for a more comprehensive understanding of the subject matter.

In Fig. 7, we illustrate the distribution of OA policies across various concepts. Our analysis reveals that bronze and green policies predominantly characterize most concepts in OA articles, except for Art, where the gold policy assumes significance (Figs. 8, 9, and 10).

Appendix B: Tables

The quality of our stratified samples is demonstrated through the descriptive statistics provided in Tables 4 and 5. Additionally, Table 6 presents a count of articles by concepts within our dataset. The regression results for formulas 1 and 2 for the entire sample are displayed in Tables 7, 8, 9, and 10.

Appendix C: Supplementary regression results

This section provides an in-depth analysis of OA citation advantage for each OpenAlex concept. To achieve this, we developed 19 distinct regression models, each dedicated to analyzing the adoption of OA citation for a single concept. We use the second formulation for each model, with data pertaining solely to the corresponding concept being considered in each case.

To gain insight into the effect of OA adoption on each concept, we present the coefficients for the is_{oa} variable in Table 11 and the coefficients for the $\ln(1 + times_{cited})$ variable in Table 12.

Table 11 indicates that OA articles across most concepts exhibit a positive OA Wikipedia citation advantage, with five concepts showing statistically significant advantages. The top five concepts with the highest OA adoption advantage are Chemistry, Economics, Psychology, Business, and Physics, suggesting that STEM-related subjects attract more attention on Wikipedia.

Table 12 Coefficients for OA adoption by concept for all samples ($(1 + times_{cited})$)

Concept	Min OR	Max OR	OR mean	Highest P-value	Mean R^2
Biology	1.594	1.602	1.599	0.000	0.067
Computer science	1.303	1.314	1.308	0.000	0.052
Chemistry	1.545	1.566	1.554	0.000	0.060
Medicine	1.820	1.835	1.824	0.000	0.134
Psychology	1.473	1.490	1.482	0.000	0.090
Mathematics	1.431	1.454	1.442	0.000	0.073
Economics	1.461	1.519	1.490	0.000	0.104
Geology	1.476	1.487	1.481	0.000	0.053
Sociology	1.113	1.211	1.153	0.003	0.014
History	1.252	1.297	1.274	0.000	0.018
Geography	1.124	1.149	1.137	0.000	0.009
Philosophy	1.109	1.126	1.116	0.000	0.005
Materials science	1.551	1.570	1.562	0.000	0.098
Art	1.191	1.267	1.213	0.000	0.008
Environmental science	1.011	1.149	1.069	0.824	0.008
Physics	1.409	1.418	1.413	0.000	0.064
Engineering	1.015	1.144	1.062	0.779	0.006
Business	1.205	1.240	1.221	0.000	0.032
Political science	1.203	1.240	1.227	0.000	0.017

Regarding $\ln(1 + \text{times}_{\text{cited}})$ in each concept, citation counts demonstrate a significantly positive effect in nearly all concepts, although Environment Science and Engineering do not show significance. OA articles in several OpenAlex concepts, including Biology, Computer Science, Chemistry, Medicine, Psychology, Mathematics, Economics, Geology, Materials Science, and Physics, exhibit, on average, over a 30% higher likelihood of being cited in Wikipedia compared to paywalled articles. Citation counts remain important factors in these concepts.

Acknowledgements Puyu Yang acknowledges the China Scholarship Council (CSC) grant. This study has been published as a pre-print before submission (Yang et al., 2024). We also declare that this manuscript is in part based on a conference paper that we presented at ISSI 2023 (Yang & Colavizza, 2023), on which the authors retain all copyrights.

Declarations

Data availability The code to replicate our work is made available online: https://github.com/alsowbdxa/Open_access_and_wikipedia. The Wikipedia Citations dataset is openly available (Kokash & Colavizza, 2024), while access to OpenAlex can be requested through their portal. All other supporting datasets we used are openly available and referenced from the Data and Methods section.

References

Antelman, K. (2004). Do open-access articles have a greater research impact? *College & Research Libraries, 65*(5), 372–382.

Archambault, É., Amyot, D., Deschamps, P., Nicol, A., Provencher, F., Rebout, L., & Roberge, G. (2014). Proportion of Open access papers published in peer-reviewed journals at the European and world levels-1996–2013. Copyright, fair use, scholarly communication, etc. Retrieved from https://digitalcommons.unl.edu/scholcom/8

Arroyo-Machado, W., Torres-Salinas, D., Herrera-Viedma, E., & Romero-Frías, E. (2020). Science through Wikipedia: A novel representation of open knowledge through co-citation networks. *PloS One, 15*(2), e0228713.

Benjakob, O., Aviram, R., & Sobel, J.A. (2022). Citation needed? Wikipedia bibliometrics during the first wave of the COVID-19 pandemic. GigaScience11giab095. Retrieved 14 Feb 2023, from https://doi.org/10.1093/gigascience/giab095

Björk, B.-C., Welling, P., Laakso, M., Majlender, P., Hedlund, T., & Guðnason, G. (2010). Open access to the scientific journal literature: Situation 2009. *PLOS ONE, 5*(6), e11273. https://doi.org/10.1371/journal.pone.0011273

Black, E. W. (2008). Wikipedia and academic peer review: Wikipedia as a recognised medium for scholarly publication? *Online Information Review, 32*(1), 73–88. https://doi.org/10.1108/14684520810865994

Bredahl, L. (2022). Chapter 1: Introduction to bibliometrics and current data sources. *Library Technology Reports, 58*(8), 5–11.

Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K., & McGillivray, B. (2020). The citation advantage of linking publications to research data. *PLOS ONE, 15*(4), e0230416. https://doi.org/10.1371/journal.pone.0230416

Colavizza, G. (2020). COVID-19 research in Wikipedia. *Quantitative Science Studies, 1*(4), 1349–1380. https://doi.org/10.1162/qss_a_00080

ElSabry, E. (2017). Who needs access to research? Exploring the societal impact of open access. Revue française des sciences de l'information et de la communication (11).
Evans, P., & Krauthammer, M. (2011). Exploring the use of social media to measure journal article impact. *AMIA Annual Symposium Proceedings, 2011*, 374.
Falagas, M. E., Kouranos, V. D., Arencibia-Jorge, R., & Karageorgopoulos, D. E. (2008). Comparison of SCImago journal rank indicator with journal impact factor. *The FASEB Journal, 22*(8), 2623–2628. https://doi.org/10.1096/fj.08-107938
Gargouri, Y., Hajjem, C., Larivière, V., Gingras, Y., Carr, L., Brody, T., & Harnad, S. (2010). Self-selected or mandated, open access increases citation impact for higher quality research. *PLOS ONE, 5*(10), e13636. https://doi.org/10.1371/journal.pone.0013636
González-Pereira, B., Guerrero-Bote, V. P., & Moya-Anegón, F. (2010). A new approach to the metric of journals' scientific prestige: The SJR indicator. *Journal of Informetrics, 4*(30), 379–391. https://doi.org/10.1016/j.joi.2010.03.002
Hao, H., Cui, Y., Wang, Z., & Kim, Y.-S. (2022). Thirty-two years of IEEE VIS: Authors, fields of study and citations. Retrieved 20 Feb 2023, from http://arxiv.org/abs/2208.03772
Holmberg, K., Hedman, J., Bowman, T. D., Didegah, F., & Laakso, M. (2020). Do articles in open access journals have more frequent altimetric activity than articles in subscription-based journals? An investigation of the research output of Finnish universities. *Scientometrics, 122*(1), 645–659. https://doi.org/10.1007/s11192-019-03301-x
Kokash, N., & Colavizza, G. (2024). *A comprehensive dataset of classified citations with identifiers from English Wikipedia*. Zenodo.
Kousha, K., & Thelwall, M. (2017). Are Wikipedia citations important evidence of the impact of scholarly articles and books? *Journal of the Association for Information Science and Technology*. https://doi.org/10.1002/asi.23694
Langham-Putrow, A., Bakker, C., & Riegelman, A. (2021). Is the open access citation advantage real? A systematic review of the citation of open access and subscription-based articles. *PloS One, 16*(6), e0253129.
Lin, J., & Fenner, M. (2014). An analysis of Wikipedia references across plos publications. Altmetrics14: Expanding impacts and metrics an ACM web science conference 2014 workshop (pp. 23–26).
Maggio, L. A., Willinsky, J. M., Steinberg, R. M., Mietchen, D., Wass, J. L., & Dong, T. (2017). Wikipedia as a gateway to biomedical research: The relative distribution and use of citations in the English Wikipedia. *PLOS ONE, 12*(12), e0190046. https://doi.org/10.1371/journal.pone.0190046
Martín-Martín, A., Costas, R., van Leeuwen, T. & Delgado López-Cózar, E. (2018). Evidence of open access of scientific publications in Google Scholar: A large-scale analysis. *Journal of Informetrics,12*(3), 819–841. https://doi.org/10.1016/j.joi.2018.06.012
Mesgari, M., Okoli, C., Mehdi, M., Nielsen, F. Å., & Lanamäki, A. (2015). "The sum of all human knowledge'': A systematic review of scholarly research on the content of Wikipedia. *Journal of the Association for Information Science and Technology, 66*(2), 219–245.
Nielsen, F. A. (2007). Scientific citations in Wikipedia. arXiv:0705.2106 [cs]
Park, T .K. (2011). The visibility of wikipedia in scholarly publications. First Monday. Retrieved from https://firstmonday.org/ojs/index.php/fm/article/download/3492/3031
Patience, G. S., Patience, C. A., Blais, B., & Bertrand, F. (2017). Citation analysis of scientific categories. *Heliyon, 3*(5), e00300. https://doi.org/10.1016/j.heliyon.2017.e00300
Piccardi, T., Redi, M., Colavizza, G., & West, R. (2020). Quantifying engagement with citations on Wikipedia. Proceedings of the web conference 2020 (pp. 2365–2376). New York: Association for Computing Machinery. Retrieved 20 Feb 2023, from https://doi.org/10.1145/3366423.3380300
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. https://doi.org/10.7717/peerj.4375
Pooladian, A., & Borrego, Á. (2017). Methodological issues in measuring citations in Wikipedia: A case study in library and information science. *Scientometrics, 113*(1), 455–464.
Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv:2205.01833 [cs]
Redalyc, L., Clase, R., & IN-COM UAB, S. (2003). Berlin declaration on open access to knowledge in the sciences and humanities. https://openaccess.mpg.de/Berlin-Declaration
Schweitzer, N. J. (2008). Wikipedia and psychology: Coverage of concepts and its use by undergraduate students. *Teaching of Psychology, 35*(2), 81–85. https://doi.org/10.1080/00986280802004594
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 (pp.

1591–1600). Republic and Canton of Geneva, CHE: International World Wide Web conferences steering committee. Retrieved 25 Nov 2022, from https://doi.org/10.1145/3038912.3052716

Singh, H., West, R., & Colavizza, G. (2020). Wikipedia citations: A comprehensive dataset of citations with identifiers extracted from english wikipedia. CoRRabs/2007.07022. arXiv: https://arxiv.org/abs/2007.07022

Struck, D. B., Durning, M., Roberge, G., & Campbell, D. (2018). Modelling the effects of open access, gender and collaboration on citation outcomes: Replicating, expanding and drilling. STI 2018 conference proceedings (pp. 436–447). Centre for Science and Technology Studies (CWTS). Retrieved 23 Feb 2023, from https://hdl.handle.net/1887/65337

Sugimoto, C. R., Work, S., Larivière, V., & Haustein, S. (2017). Scholarly use of social media and altmetrics: A review of the literature. *Journal of the Association for Information Science and Technology, 68*(9), 2037–2062. https://doi.org/10.1002/asi.23833

Tattersall, A., Sheppard, N., Blake, T., O'Neill, K., & Carroll, C. (2022). Exploring open access coverage of Wikipedia—Cited research across the white rose universities. Insights (35).

Tennant, J. P., Waldner, F., Jacques, D. C., Masuzzo, P., Collister, L. B., & Hartgerink, C. H. J. (2016). The academic, economic and societal impacts of Open Access: An evidence-based review (Tech. Rep. No. 5:632). F1000Research. Retrieved 20 Feb 2023, from https://f1000research.com/articles/5-632

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. https://doi.org/10.1002/asi.23687

Thompson, N., & Hanley, D. (2018). Science is shaped by Wikipedia: Evidence from a randomized control trial [SSRN Scholarly Paper]. Rochester. Retrieved 20 Feb 2023, from https://papers.ssrn.com/abstract=3039505

Tohidinasab, F., & Jamali, H. R. (2013). Why and where Wikipedia is cited in journal articles? *Journal of Scientmetric Research, 2*, 231–238. https://doi.org/10.4103/2320-0057.135415

Yang, P., & Colavizza, G. (2022). A map of science in Wikipedia. Proceedings of the web conference. arXiv: http://arxiv.org/abs/2110.13790

Yang, P., & Colavizza, G. (2023). Open access science in Wikipedia. Proceedings of ISSI 2023—The 19th international conference of the international society for scientometrics and informetrics (vol. 2, pp. 515–519). https://doi.org/10.5281/zenodo.8432306

Yang, P., Shoaib, A., West, R., & Colavizza, G. (2024). Open access improves the dissemination of science: Insights from Wikipedia. arXiv. arXiv:2305.13945 [cs] version: 2

Yegros-Yegros, A., Rafols, I., & D'Este, P. (2015). Does interdisciplinary research lead to higher citation impact? The different effect of proximal and distal interdisciplinarity. *PLOS ONE, 10*(8), e0135095. https://doi.org/10.1371/journal.pone.0135095

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

Yuen, J. (2018). Comparison of impact factor, Eigenfactor metrics, and SCImago journal rank indicator and h-index for neurosurgical and spinal surgical journals. *World Neurosurgery, 119*, e328–e337. https://doi.org/10.1016/j.wneu.2018.07.144

Zagorova, O., Ulloa, R., Weller, K., & Flöck, F. (2021) "I updated the": The evolution of references in the English Wikipedia and the implications for altmetrics. *Quantitative Science Studies,* 1–27. https://doi.org/10.1162/qss_a_00171